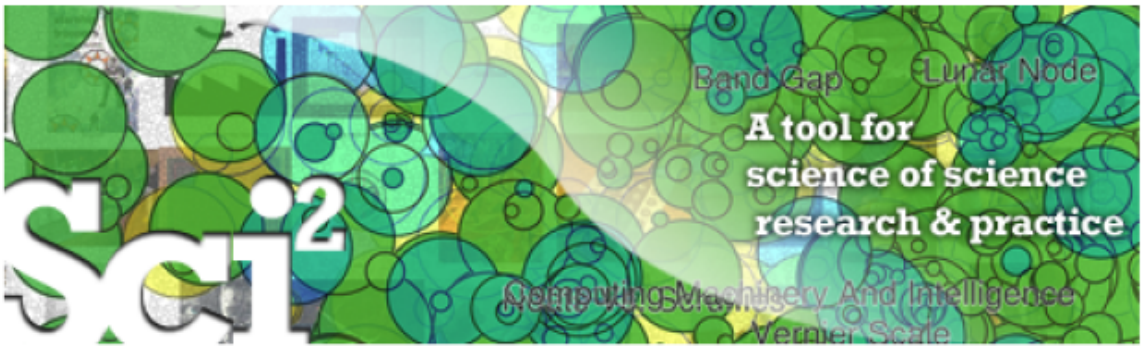
Science of Science (Sci2) Tool Manual v1.0 alpha

Download an offline copy of this manual (87.6 MB) (Updated May 31, 2011)

Download an offline copy of the CIShell Manual, which contains algorithm documentation for Sci2 (19.9 MB) (Updated May 31, 2011)



Watch the movie about CIShell-Powered tools on the SciVee: Making Science Visible website by clicking on the image above.

http://www.scivee.tv/node/27704

Science of Science (Sci^2) Tool
User Manual, v0.5 alpha
Updated 3.28.2011

Project Investigators: Katy Börner and Kevin W. Boyack (SciTech Strategies Inc.)
Programmers: Micah W. Linnemeier, Russell J. Duhon, Patrick A. Phillips, Chintan Tank, and Joseph Biberstine
Users, Testers & Tutorial Writers: Katy Börner, Hanning Guo, Scott Weingart

Cyberinfrastructure for Network Science Center
School of Library and Information Science
Indiana University, Bloomington, IN
http://cns.iu.edu

This work is funded by the School of Library and Information Science and the Cyberinfrastructure for Network Science Center at Indiana University, the James S. McDonnell Foundation, and the National Science Foundation under Grants No. IIS-0715303, IIS-0534909, and IIS-0513650. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Contents

Home

The page Science of Science (Sci2) Tool Manual v0.5 alpha does not exist.

1 Introduction

The Science of Science (Sci^2) Tool (http://cns.iu.indiana.edu) is a modular toolset specifically designed for the study of science. It supports the temporal, geospatial, topical, and network analysis and visualization of datasets at the micro (individual), meso (local), and macro (global) levels. Users of the tool can

- Access science datasets online or load their own.
- Perform different types of analysis with the most effective algorithms available.
- Use different visualizations to interactively explore and understand specific datasets.
- Share datasets and algorithms across scientific boundaries.

The Sci^2 Tool is built on the Cyberinfrastructure Shell (CIShell) (Cyberinfrastructure for Network Science Center, 2008), an open source software framework for the easy integration and utilization of datasets, algorithms, tools, and computing resources. CIShell is based on the OSGi R4 Specification and Equinox implementation (OSGi-Alliance, 2008).

2 Getting Started

2.1 Download, Install, Uninstall

The Sci^2 Tool is a stand-alone desktop application that installs and runs on all common operating systems. It requires Java SE 5 (version 1.5.0) or later to be pre-installed on your local machine.
You can check the version of your Java installation by running the command line:

```
java-version
```

If it is not already installed on your computer, download and install the latest version of Java from http://www.java.com/en/download/index.jsp.
To download the Sci^2 Tool, find the download link at http://sci.cns.iu.edu/registration/user/ and select your operating system from the pull down menu.

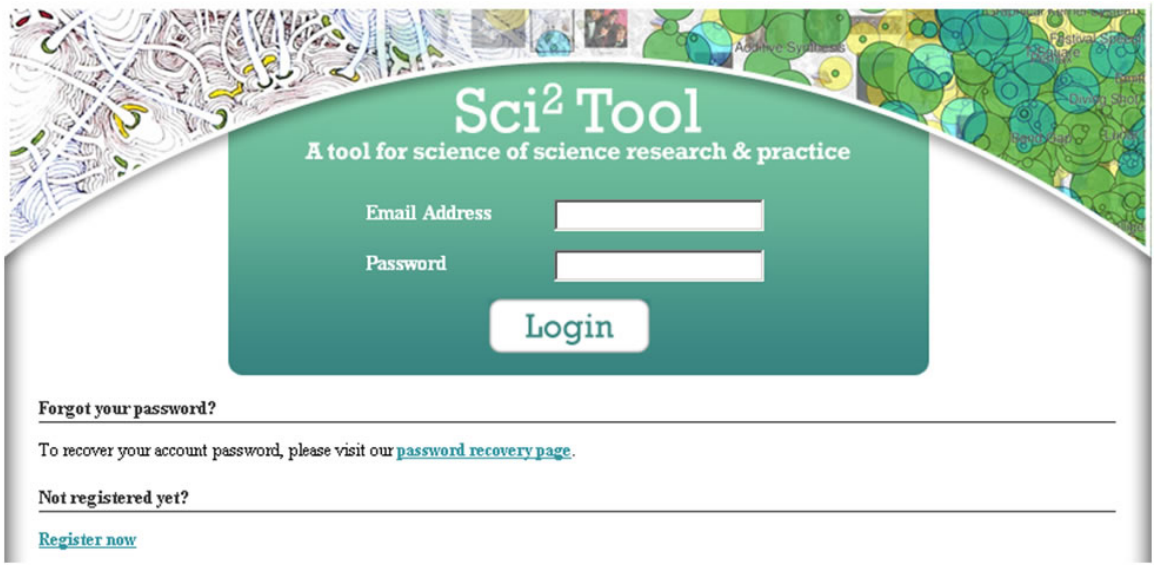


Figure 2.1: Downloading the Sci² Tool

Save the zip file in a new empty *yoursci2directory* and extract all files. After the files have been extracted, double click the S2 icon (sci2.exe) in *yoursci2directory* directory to run the program.

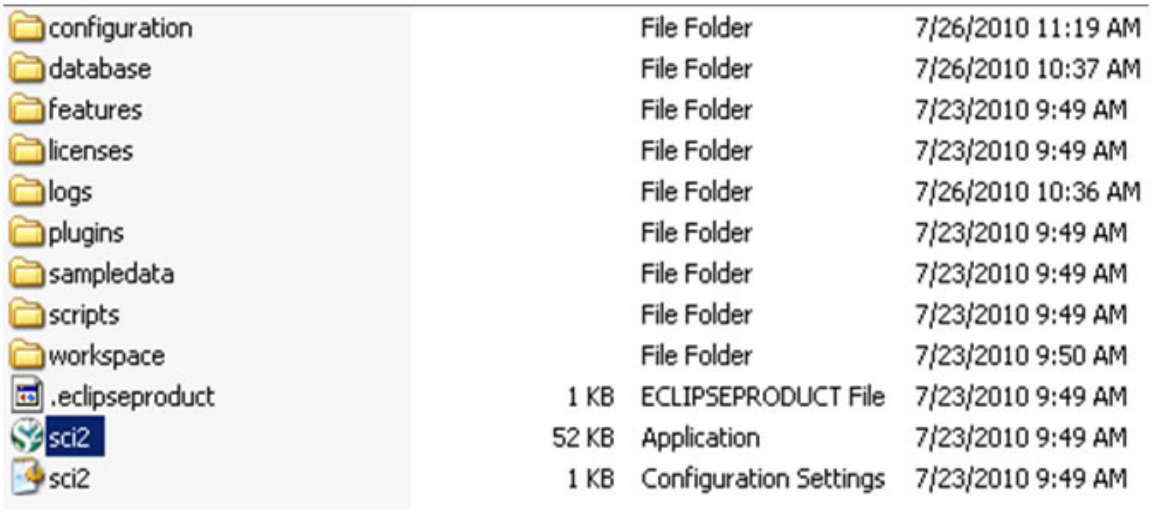


Figure 2.2: Click the 'sci2' icon to run the Sci² Tool

To uninstall the Sci² Tool, simply delete *yoursci2directory*. This will delete all sub-directories as well, so make sure to backup all files you want to save.

2.2 User Interface

- 2.2.2 Console
- 2.2.3 Data Manager
- 2.2.4 Scheduler

The general Sci^2 Tool user interface is shown in Figure 2.3.

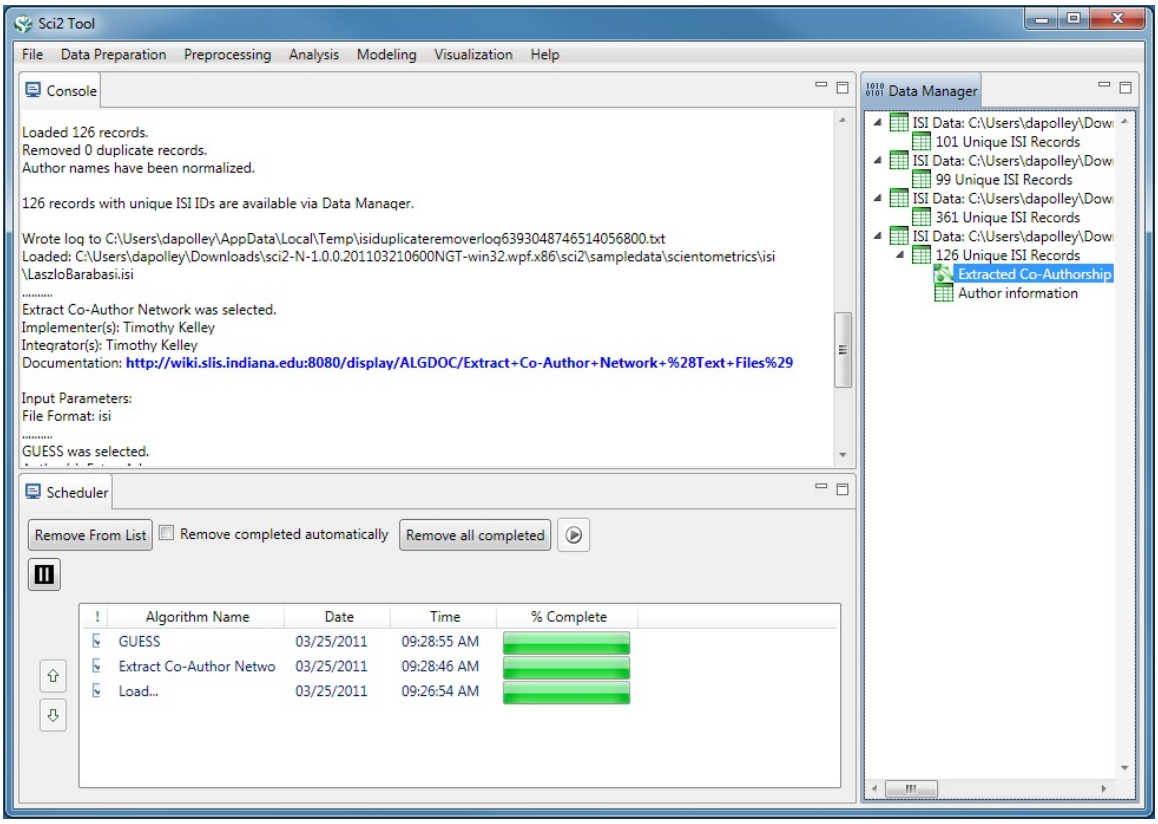


Figure 2.3: Sci^2 Tool Interface Components

2.2.1 Menus

The Sci^2 Tool menu structure is arranged such that a workflow runs from left to right. The '*File*' menu on the left allows a user to load data in a number of formats, which can then be prepared, preprocessed, analyzed, and finally visualized. Users also have the option of modeling new networks or finding help online.

2.2.1.1 File

In the 'File' menu, the first option after clicking on 'Load' is 'Select a File':

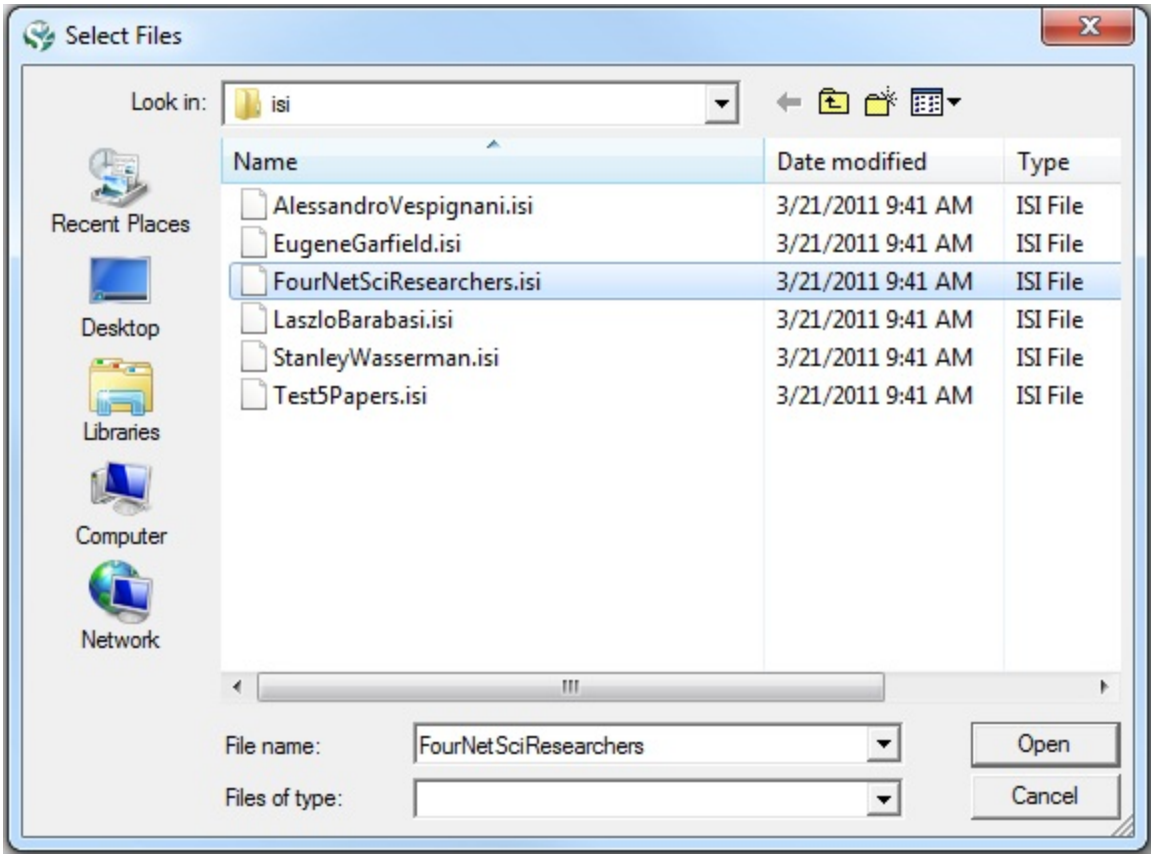


Figure 2.4: Using 'Load' to select a file in Sci²

The '*File*' menu functionality includes loading multiple data formats (see section 2.3 Data Formats for details), loading ISI and NSF data, saving and viewing results, and merging or splitting node and edge files.

The '*Console*' window documents all operations performed on the data.

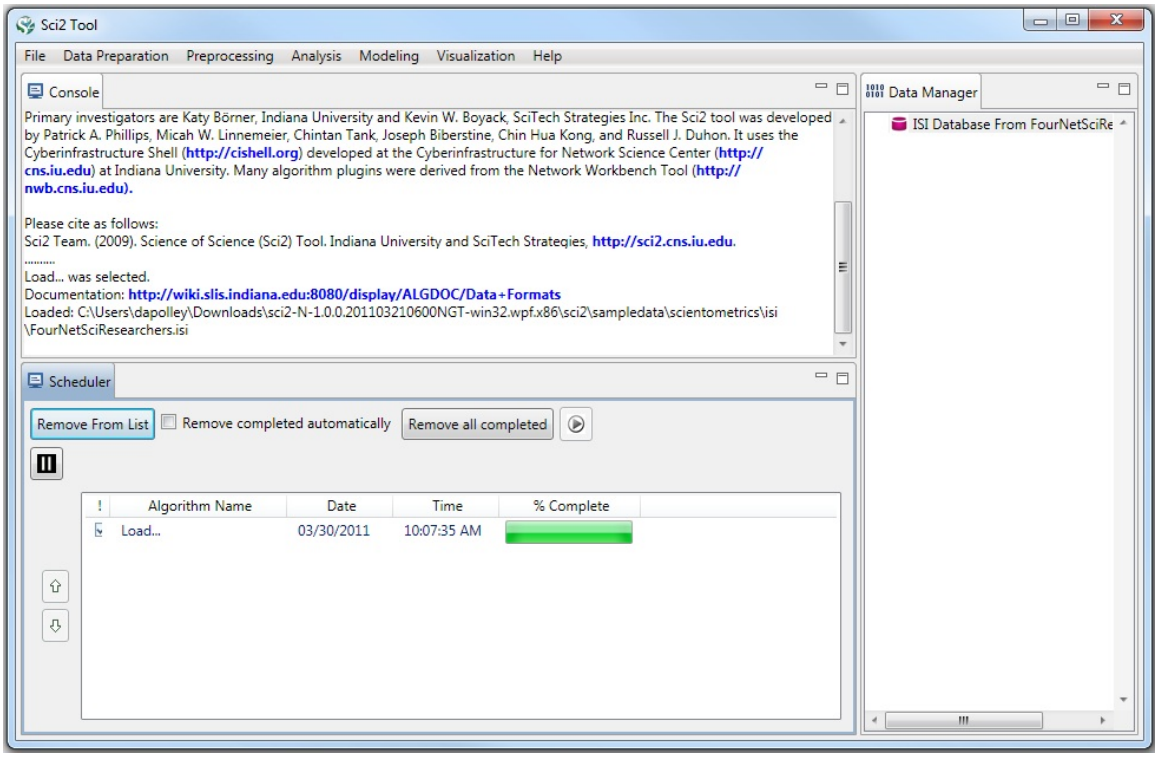


Figure 2.6: Sci^2's Console and Schedule after successfully loading a database file

When the scheduler indicates that the 'Load' operation is complete, the .isi file will appear in the data manager, preceded by a database icon. The converter graph and directory reader produces a sample graph based on filetypes supported by the Sci^2 Tool and a sample tree based on any directory structure on the hard drive, respectively.

2.2.1.2 Data Preparation

> **Extended Version**
>
> The screen shots in the rest of the section are from the extended version of the Sci2 Tool. To extend Sci2 see 3.2 Additional Plugins.

After loading a file, use options in the '*Data Preparation*' menu to clean the data and create networks or tables which can be used in the preprocessing, analysis, and visualization steps. The options in the top of the menu are for any table-based datasets (like csv files) and are used to extract networks. The '*Data Preparation* > *Database*' menu is specifically for ISI or NSF data previously loaded into a database. Find detailed information on each menu item in section 3.1 Sci2 Algorithms and Tools.

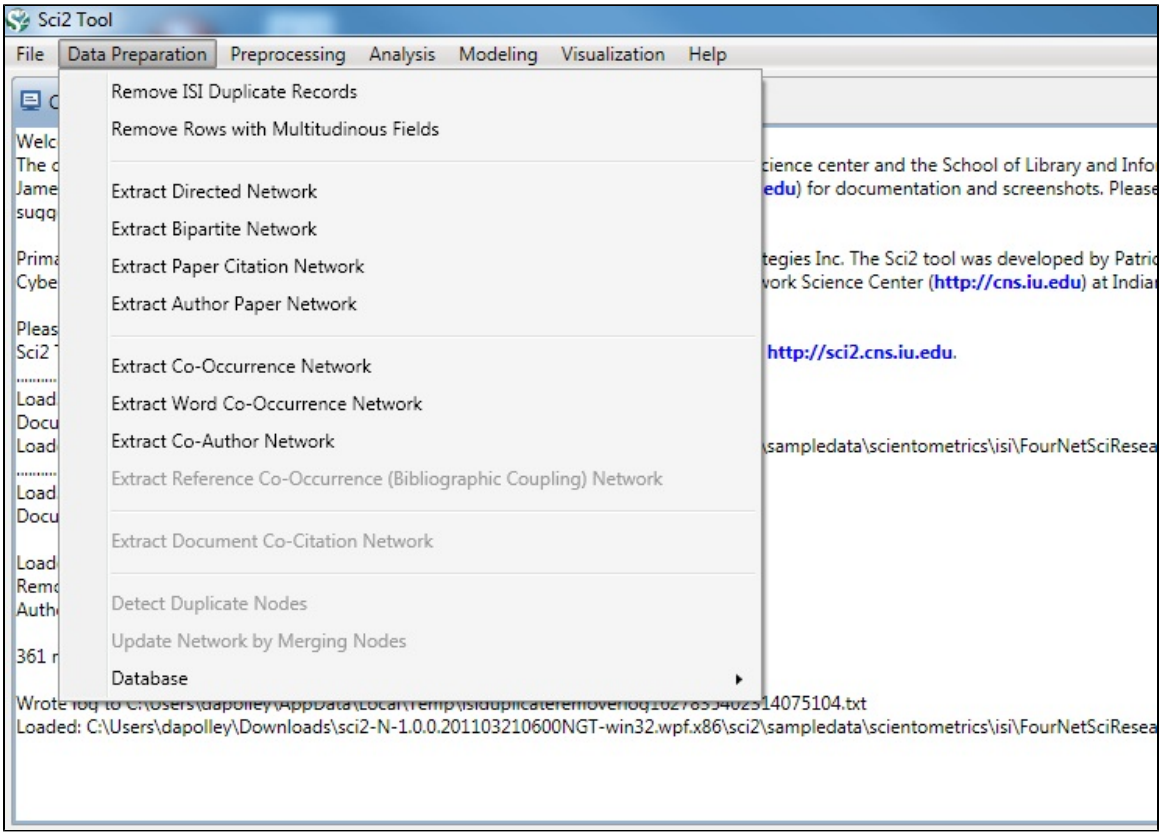


Figure 2.7: Data Preparation options

2.2.1.3 Preprocessing

Use preprocessing algorithms to prune or append networks or tables before analyzing and visualizing them. The menu is separated by domain, and most simple tasks require staying within the same domain. For example, to

visualize a co-authorship network, only use algorithms within the '*Networks*' domain under '*Preprocessing*', '*Analysis*' , and '*Visualization*'. Similarly, a geographic map requires only '*Geospatial*' algorithms. Find detailed information on each menu item in section 3.1 Sci2 Algorithms and Tools.

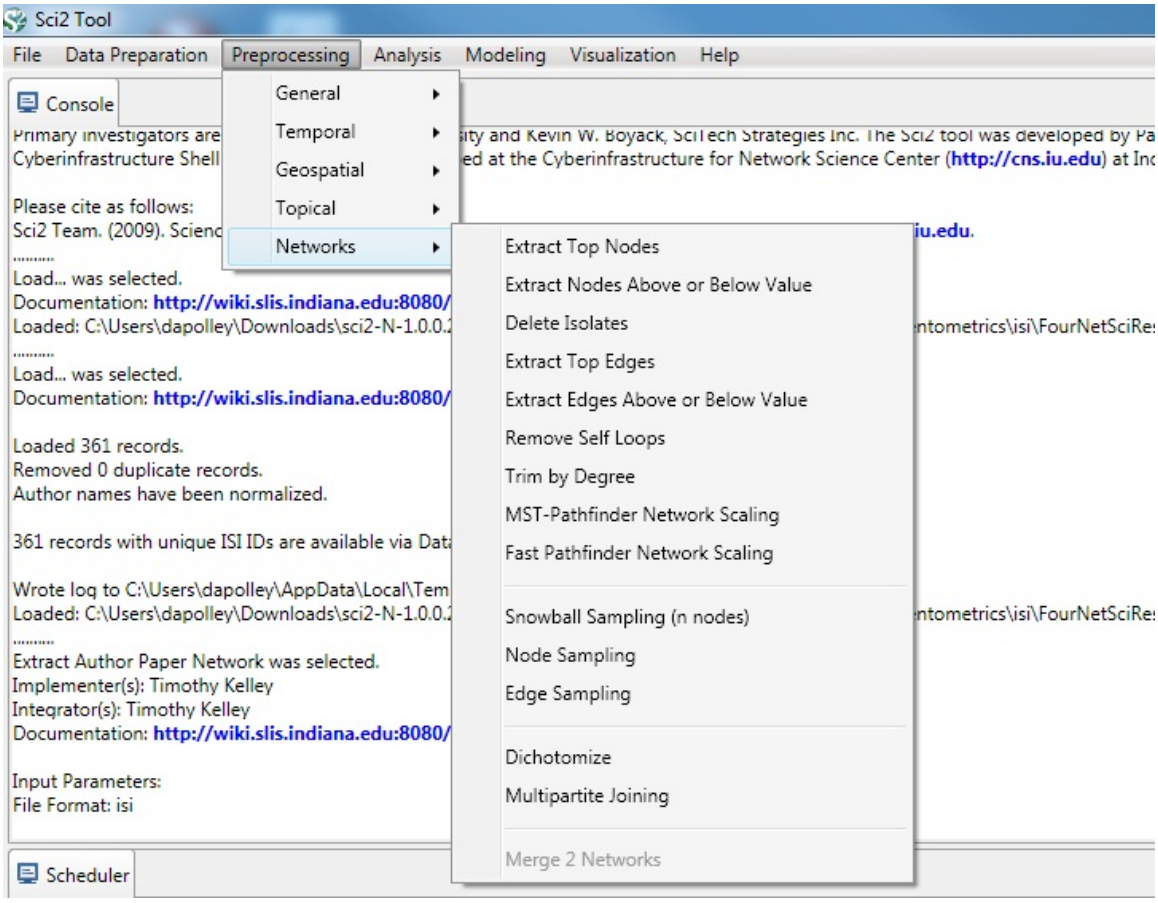


Figure 2.8: Preprocessing options

2.2.1.4 Analysis

Once data is loaded, prepared, and processed with whatever features needed, analysis is possible in each of the four domains: temporal, geospatial, topical, or network.

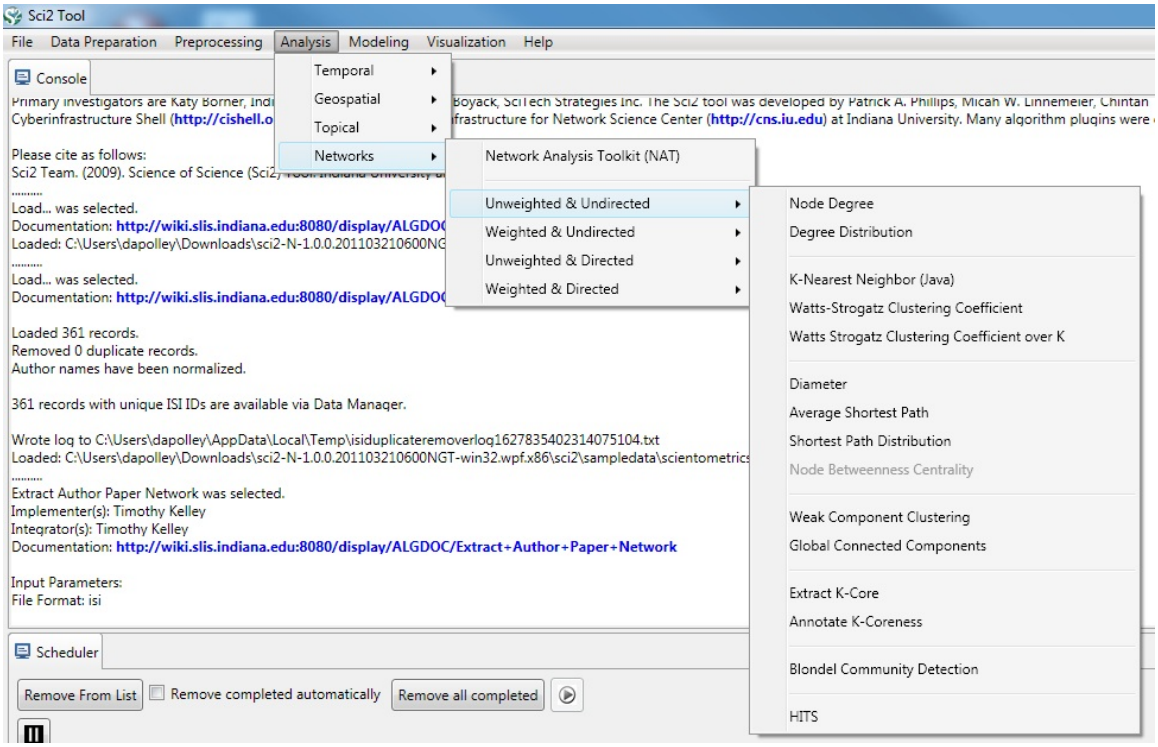


Figure 2.9: Analysis options

Analysis results can be used on their own or in conjunction with visualizations to gain insight into a dataset. The Sci^2 Tool features predominantly network analysis algorithms, however the tool also supports geocoding of table data and burst analysis for topical or temporal studies. Find detailed information on each menu item in section 3.1 Sci2 Algorithms and Tools

2.2.1.5 Modeling

The Sci^2 Tool supports the creation of new networks via pre-defined models. Learn more about modeling in section 4.10 Modeling (Why?).

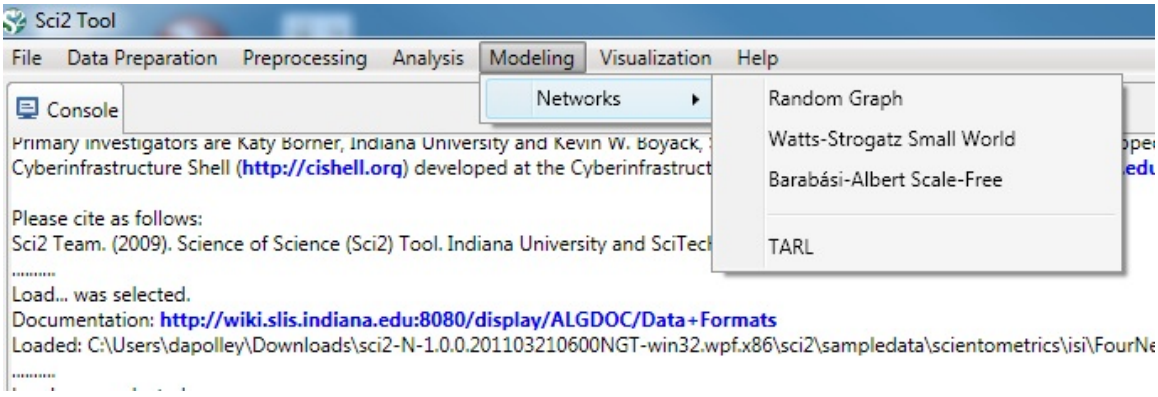


Figure 2.10: Modeling options

2.2.1.6 Visualization

Once all previous data steps are complete, the Sci^2 Tool can visualize the results. The most popular choice for visualizing networks is the GUESS toolkit, or DrL for much larger scale networks. Geocoded data can be represented on a map of the world or the United States, and temporal or topical data can be viewed using the horizontal bar graph. Find detailed information on each menu item in section3.1 Sci2 Algorithms and Tools.

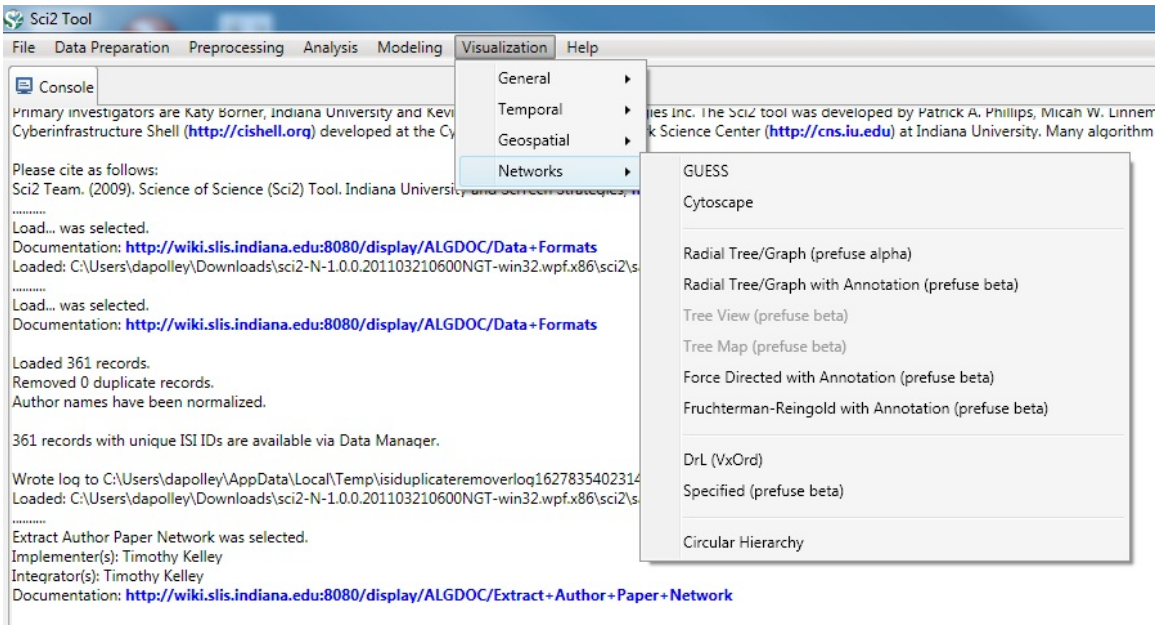


Figure 2.11: Visualization options

Cytoscape should be explained here. Coming soon.

2.2.1.7 Help

The '*Help*' menu leads to online documentation, advanced tool configuration, and detailed development information.

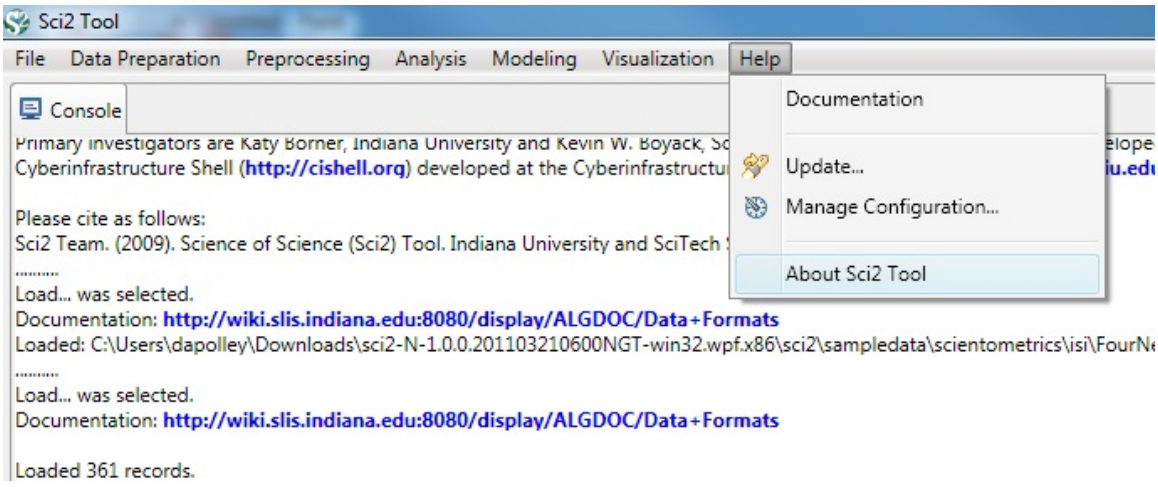


Figure 2.12: Help options

2.2.2 Console

All operations such as loading, viewing, or saving datasets, running various algorithms, and algorithm parameters, etc. are logged sequentially in the 'Console' window as well as in log files stored in the '***yoursci2directory*** */logs'* dire ctory. The Console window also displays the acknowledgement information about the original authors of the algorithm, the developers, the integrators, a reference paper, and the URL to the reference if available, together with an URL to the algorithm description in the NWB/Sci2 community wiki.

2.2.3 Data Manager

The 'Data Manager' window displays all currently loaded and available datasets. The type of a loaded file is indicated by its icon:

 Text – text file

 Table – tabular data (csv file)

 Matrix-data (Pajek .mat)

 Plot – plain text file that can be plotted using Gnuplot

 Network – Network data (in-memory graph/network object or network files saved as Graph/ML, XGMML, NWB, Pajek .net or Edge list format)

 Database – In-memory database

 Tree – Tree data (TreeML)

Derived datasets are indented under their parent datasets. That is, the children datasets are the results of applying certain algorithms to the parent dataset.

2.2.4 Scheduler

The 'Scheduler' lets users keep track of the progress of running algorithms.

2.3 Data Formats

In August 2010, the Sci^2 Tool supports loading the following input file formats:

- Network Formats
  - GraphML (*.xml or *.graphml)
  - XGMML (*.xml)
  - Pajek .NET (*.net)
  - NWB (*.nwb)
- Scientometric Formats
  - ISI (*.isi)
  - Bibtex (*.bib)
  - Endnote Export Format (*.enw)
  - Scopus csv (*.scopus)
  - NSF csv (*.nsf)
- Other Formats
  - Pajek Matrix (*.mat)
  - TreeML (*.xml)
  - Edgelist (*.edge)
  - CSV (*.csv)
- Special Load Algorithms
  - Load and Clean ISI File
- Databases
  - SDB
- Interfacing with other applications
  - UCINET

Internal database schema formats:

- ISI Database Tables
  - ADDRESS
  - AUTHORS
  - CITED_PATENTS
  - CITED_REFERENCES
  - DOCUMENT

Network file output formats:

- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)

Image output formats:

- JPEG (*.jpg)
- PDF (*.pdf)
- PostScript (*.ps) - Sample PS file

These format's relationships can be seen in Figure 2.13.

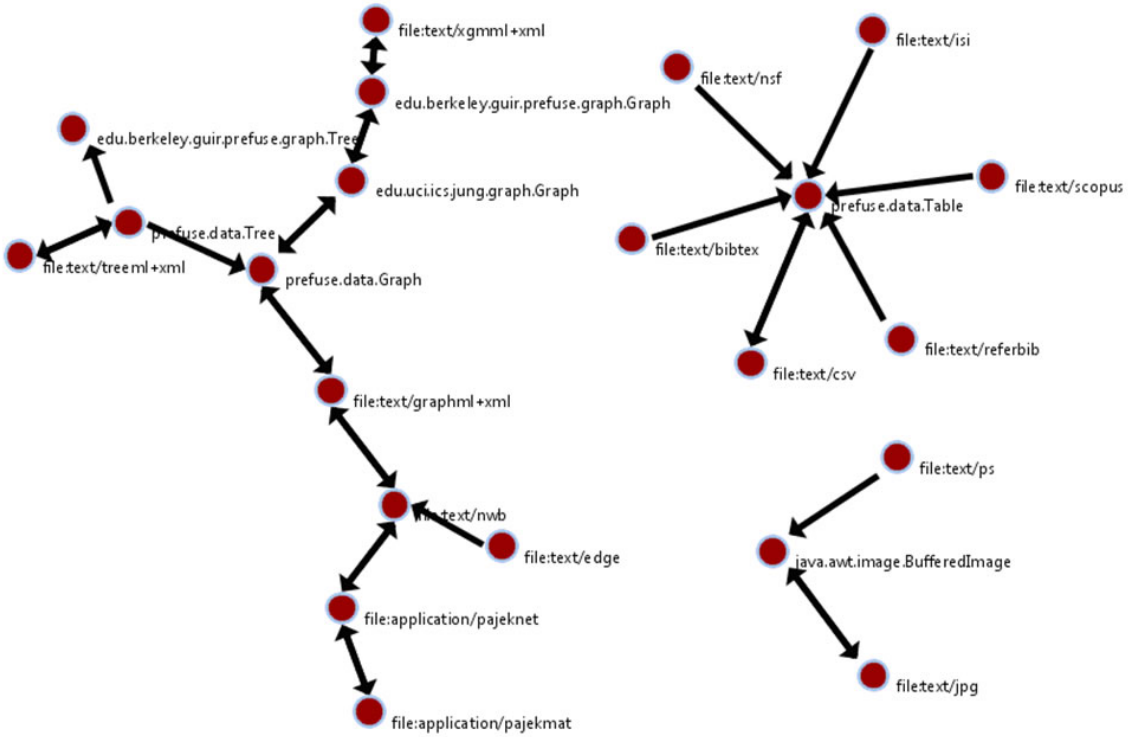


Figure 2.13: Visualization of compatible data formats

2.3.1 Database Schema

- ISI Database Tables
  - ADDRESS
  - AUTHORS
  - CITED_PATENTS
  - CITED_REFERENCES
  - DOCUMENT
  - DOCUMENT_KEYWORDS
  - DOCUMENT_OCCURRENCES
  - EDITORS
  - ISI_FILES
  - KEYWORD
  - PATENT
  - PERSON
  - PUBLISHER
  - PUBLISHER_ADDRESSES
  - REFERENCE
  - REPRINT_ADDRESSES

ISI Database Tables

> **Outdated**
>
> The ISI Database Tables have been replaced with Publication Database Tables.

ADDRESS

Addresses appear in standard ISI data in three forms: as Publisher Addresses, Reprint Addresses, and Research Addresses. Parsing addresses into their components is an inherently complex problem to solve algorithmically, so we simply do not do that. However, the database schema attempts to accommodate for the most common address components for flexibility with future loader versions and custom user queries.

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *ADDRESS_CITY*: Currently not used.
- *ADDRESS_COUNTRY*: Currently not used.
- *ADDRESS_POSTAL_CODE*: Currently not used.
- *RAW_ADDRESS_STRING*: The exact address string as found in the original ISI dataset. Also guarantees uniqueness.
- *ADDRESS_STATE_OR_PROVINCE*: Currently not used.
- *STREET_ADDRESS*: Currently not used.
- Addresses with the same *RAW_ADDRESS_STRING* are automatically merged to be the same addresses.

For further information on how the various types of addresses are handled, see PUBLISHER_ADDRESSES, REPRINT_ADDRESSES, and RESEARCH_ADDRESSES.

AUTHORS

This table does not contain actual author data besides e-mail addresses. Rather, it is a relationship table between **DOCUMENT** and **PERSON**. Each record in this table describes a "Person was an author of Document" relationship. That being said, both the **PERSON** and **AUTHORS** tables are being referred to when mentioning authors in this article.

Authors are parsed primarily out of the "AU" ISI field, which all valid ISI datasets should contain, but they can also be parsed out of References. The remainder of this article will speak as if parsing out of the "AU" field.

If provided, full author name data is parsed out of the "AF" field, which is expected to contain as many entries as the "AU" field. Entries in both fields are matched by the order they were listed; there is no error checking/correction if the orders mismatch.

If provided, the e-mail address data is parsed out of the "EM" field, which is expected to contain either as many entries as the "AU" field or one entry. If the number of entries in both fields match, the entries are matched by the order they were listed (similarly to full author names). Otherwise, the first (or only) e-mail address is used for all

authors.

The abbreviated name fields in the **PERSON** table (*FAMILY_NAME*, *FIRST_INITIAL*, *MIDDLE_INITIAL*, and *UNSPLIT_NAME*) are parsed from the "AU" field. This article describes how the "AU" field is parsed. Note: Even when the "AF" field is supplied, the *FAMILY_NAME* is parsed out of the "AU" field.

The full name fields in the **PERSON** table (*ADDITIONAL_NAME*, *FULL_NAME*, and *PERSONAL_NAME*) are parsed from the "AF" field. This article describes how the "AF" field is parsed.

This table contains the following fields:

- *AUTHORS_DOCUMENT_FK*: The foreign key into the **DOCUMENT** table.
- *AUTHORS_PERSON_FK*: The foreign key into the **PERSON** table.
- *EMAIL_ADDRESS*: Parsing/handling of e-mail addresses is described above.
- *ORDER_LISTED*: The order in which this author was listed in the "AU"/etc. fields. Currently is 0-based (so the first is 0, the second 1, etc.), though that is likely to change soon.

In the various built-in extractions, when Authors are specified in the "AU" field, they are considered to be inner authors. When they are specified as part of reference strings and are not merged, they are considered to be outer authors.

See DOCUMENT, PERSON, REFERENCE, How Abbreviated Names are Parsed, and How Full Names are Parsed for additional details.

CITED_PATENTS

This table does not contain actual patent data. Rather, it is a relationship table between **DOCUMENT** and **PATENT**. Each record in this table describes a "Document cites Patent" relationship.

This table contains the following fields:

- *CITED_PATENTS_DOCUMENT_FK*: The foreign key into the **DOCUMENT** table.
- *CITED_PATENTS_PATENT_FK* : The foreign key into **PATENT** table.

See DOCUMENT and PATENT for additional details.

CITED_REFERENCES

This table does not contain actual reference data. Rather, it is a relationship table between **DOCUMENT** and **REFERENCE**. Each record in this table describes a "Document cites Reference" relationship.

This table contains the following fields:

- *CITED_REFERENCES_DOCUMENT_FK*: The foreign key into the **DOCUMENT** table.
- *CITED_REFERENCES_REFERENCE_FK*: The foreign key into the **REFERENCE** table.

See DOCUMENT and REFERENCE for additional details.

DOCUMENT

This table contains data directly pertaining to documents. It includes all fields relevant for all possible (known) types of documents in ISI *Web of Science* data; that is, a Document is not necessarily always a(n) article, conference, paper, etc.

All "arbitrary" fields found in the original ISI dataset are also appended to this table, as this table is somewhat central to the database schema. Any fields that are not in the following list are considered to be arbitrary:

- AB (Abstract)

- AR (New Article Number)
- AU (Authors)
- AF (Author Full Names)
- BP (Beginning Page)
- SE (Book Series Title)
- BS (Book Series Subtitle)
- CP (Cited Patent)
- NR (Cited Reference Count)
- CR (Cited References)
- CY (Cited Year)
- HO (Conference Host)
- CL (Conference Location)
- SP (Conference Sponsors)
- CT (Conference Title)
- DT (Document Type)
- DI (DOI)
- ED (Editors)
- EM (E-mail Addresses)
- EF (End of File)
- EP (Ending Page)
- ER (End of Record)
- FN (File Type)
- SO (Full Journal Title)
- FU (Funding Agency and Grant Number)
- FX (Funding Text)
- BN (ISBN)
- SN (ISSN)
- GA (ISI Document Delivery Number)
- JI (Abbreviated ISO Journal Name)
- IS (Issue)
- LA (Language)
- ID (New ISI Keywords)
- PG (Number of Pages)
- DE (Original Keywords)
- PN (Part Number)
- PD (Publication Date)
- PT (Publication Type)
- PY (Publication Year)
- PA (Publisher Address)
- PI (City of Publisher)
- PU (Publisher)
- WP (Publisher Web Address)
- RP (Reprint Address)
- C1 (Research Addresses)
- SI (Special Issue)
- SC (Subject Category)
- SU (Supplement)
- TC (Times Cited)
- TI (Title)
- J9 (Abbreviated Journal Name)
- UT (Unique ID)
- VR (Version Number)
- VL (Volume)

Note that the above is merely a list of ISI fields handled by the loader, not fields the end up in this table. That being said, this table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *ABSTRACT_TEXT*: The exact string as found in the "AB" field.
- *ARTICLE_NUMBER*: The exact string as found in the "UT" field.
- *BEGINNING_PAGE*: The contents of the "BP" field parsed as an integer.
- *CITED_REFERENCE_COUNT*: The contents of the "NR" field parsed as an integer.
- *DIGITAL_OBJECT_IDENTIFIER*: The exact string as found in the "DI" field.
- *DOCUMENT_TYPE*: The exact string as found in the "DT" field.
- *DOCUMENT_VOLUME*: The contents of the "VL" field parsed as an integer.
- *ENDING_PAGE*: The contents of the "EP" field parsed as an integer.
- *FIRST_AUTHOR_FK*: If any authors were provided (in the "AU" field), this will be the foreign key into the **PERSON** table of the first author specified. See AUTHORS and PERSON.
- *FUNDING_AGENCY_AND_GRANT_NUMBER*: The exact string as found in the "FU" field.
- *FUNDING_TEXT*: The exact string as found in the "FX" field.
- *ISBN*: The exact string as found in the "BN" field.
- *ISI_DOCUMENT_DELIVERY_NUMBER*: The exact string as found in the "GA" field.
- *ISI_UNIQUE_ARTICLE_IDENTIFIER*: The exact string as found in the "UT" field.
- *ISSUE*: The exact string as found in the "IS" field.
- *LANGUAGE*: The exact string as found in the "LA" field.
- *PAGE_COUNT*: The contents of the "PG" field parsed as an integer.
- *PART_NUMBER*: The exact string as found in the "PN" field.
- *PUBLICATION_DATE*: The exact string as found in the "PD" field. Is not an actual date field in terms of data types because it may actually be a range or list of dates, or it may contain arbitrary text.
- *PUBLICATION_YEAR*: The contents of the "PT" field parsed as an integer.
- *DOCUMENT_SOURCE_FK*: If a source was provided (see SOURCE for an explanation of which ISI fields end up in the **SOURCE** table), this will be the foreign key into the **SOURCE** table.
- *SPECIAL_ISSUE*: The exact string as found in the "SI" field.
- *SUBJECT_CATEGORY*: The exact string as found in the "SC" field.
- *SUPPLEMENT*: The exact string as found in the "SU" field.
- *TIMES_CITED*: The contents of the "TC" field parsed as an integer.
- *TITLE*: The exact string as found in the "TI" field.

It is possible for References to be matched to Documents (as an "is a" relationship) during loading based on either Digital Object Identifiers and/or Article Numbers matching. See REFERENCE for more details on this. (There is also an explicit Match References To Papers step, which uses different metrics for matching Documents to References.).

In the various built-in extractions, when Documents are specified in the original ISI dataset, they are considered to be inner documents. When they are only specified as References and are not matched to actual Documents, they are considered to be outer documents.

DOCUMENT_KEYWORDS

his table does not contain actual keyword data. Rather, it is a relationship table between **DOCUMENT** and **KEYWORD**. Each record in this table describes a "Document was tagged with Keyword" relationship.

This table contains the following fields:

- *DOCUMENT_KEYWORDS_DOCUMENT_FK*: The foreign key into the **DOCUMENT** table.
- *DOCUMENT_KEYWORDS_KEYWORD_FK*: The foreign key into the **KEYWORD** table.
- *ORDER_LISTED*: The order in which this author was listed in the "ID"/"DE"/etc. fields. Currently is 0-based (so the first is 0, the second 1, etc.), though that is likely to change soon.

See DOCUMENT_KEYWORDS and KEYWORD for additional details.

DOCUMENT_OCCURRENCES

This table contains neither actual document data nor ISI file metadata. Rather, it is a relationship table between **DOCUMENT** and **ISI_FILES**. Each record in this table describes a "Document occurred in ISI File" relationship.

This table contains the following fields:

- *DOCUMENT_OCCURRENCES_DOCUMENT_FK*: The foreign key into the **DOCUMENT** table.
- *DOCUMENT_OCCURRENCES_ISI_FILE_FK*: The foreign key into the **ISI_FILES** table.

See DOCUMENT and ISI_FILES for additional details.

EDITORS

This table does not contain actual editor data. Rather, it is a relationship table between **DOCUMENT** and **PERSON**. Each record in this table describes a "Person was an editor of Document" relationship. That being said, both the **PERSON** and **EDITORS** tables are being referred to when mentioning editors in this article.

Editors are only parsed out of the "ED" ISI field. It is always assumed that they are specified in the abbreviated name form, the parsing of which is described in this article. As such, the abbreviated name fields in the **PERSON** table (*FAMILY_NAME*, *FIRST_INITIAL*, *MIDDLE_INITIAL*, and *UNSPLIT_NAME*) are the only fields that will potentially contain data.

This table contains the following fields:

- *EDITORS_DOCUMENT_FK*: The foreign key into the **DOCUMENT** table.
- *EDITORS_PERSON_FK*: The foreign key into the **PERSON** table.
- *ORDER_LISTED*: The order in which this editor was listed in the "ED" field. Currently is 0-based (so the first is 0, the second 1, etc.), though that is likely to change soon.

See DOCUMENT, PERSON, and How Abbreviated Names are Parsed for additional details.

ISI_FILES

This table contains originating ISI file metadata, some of which is specified in the originating ISI file and some of which is not. This table may be useful for things like determining which ISI datasets documents occurred in if/when multiple separate ISI files being loaded/merged into one database is supported. Currently, however, multiple separate ISI files being loaded/merged into one database is not supported.

Note: The name **ISI_FILES** breaks the naming scheme established throughout the rest of the ISI schema. This is a mistake that will likely be fixed in the (hopefully) near future.

This table contains the following fields:

- *PK*: Automatically generated.
- *FILE_FORMAT_VERSION_NUMBER*: The exact string as found in the "VR" field.
- *FILE_NAME*: Generated based on the ISI dataset the user chose in the Data Manager from within the tool.
- *FILE_TYPE*: The exact string as found in the "FN" field.

Every combination of all the fields in this table guarantees uniqueness.

See DOCUMENT_OCCURRENCES for additional details.

KEYWORD

This table contains keywords and their types. Currently, there are two known and supported types of keywords:

- *Original Keywords*: Keywords assigned to documents by their authors. Found in the "DE" ISI field. Keywords of this type have the value authorKeywords in the *TYPE* field of this table.
- *New ISI Keywords*: Keywords assigned to documents by ISI. Found in the "ID" ISI field. Keywords of this type have the value keywordsPlus in the *TYPE* field of this table.

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *KEYWORD*: The exact (keyword) string as found in the "DE"/"ID"/etc.
- *TYPE*: Determined by which keyword field this keyword was found in. Look above for possible values.

The combination of *KEYWORD* with *TYPE* also guarantees uniqueness.

See DOCUMENT_KEYWORDS for additional details.

PATENT

This table contains patent numbers of patents that documents cite. Patent numbers are treated as "blobs"; that is, they are not parsed into components at all.

Note: As of March 18, 2010, as a minor bug in the ISI loader, the entire contents of the "CP" field are treated as one patent (number) in the resulting database, even if multiple patents are actually listed. This should be easy to fix, and chances are it will be before anyone actually looks at this page.

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *PATENT_NUMBER*: The exact string as found in the "CP" field in the original ISI dataset. Also guarantees uniqueness.

Patents with the same *PATENT_NUMBER* are automatically merged to be the same patents.

See CITED_PATENTS for additional details.

PERSON

This table contains the abbreviated names and full names of people, both in unsplit and split forms. People can be parsed out of:

- the "AU"/"AF" ISI fields as Authors.
- References as Authors.
- the "ED" ISI field as Editors.

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *ADDITIONAL_NAME*: Also known as "middle name." Is only parsed out of the "AF" field.
- *FAMILY_NAME*: Also known as "last name." Hypothetically, could be parsed of the "AF" field, but is only parsed and used out of the "AU" and "ED" fields.
- *FIRST_INITIAL*: Can be parsed out of the "AU" and "ED" fields.
- *FULL_NAME*: The exact (unsplit/unparsed) string as found in the "AF" field.
- *MIDDLE_INITIAL*: Can be parsed out of the "AU" and "ED" fields.
- *PERSONAL_NAME*: Also known as "first name." Is only parsed out of the "AF" field.
- *UNSPLIT_NAME*: Not totally the exact (unsplit/unparsed) string as found in the "AU"/"ED" fields. The values in this field will always contain the same information as the strings found in the original ISI dataset, but the versions in this field are converted to a canonical form.

People are never merged during loading. Automatic People merging can be done via the Merge Identical ISI People algorithm, and merge suggestions (i.e. a merge table) can be generated via the Suggest ISI People Merges algorithm.

The abbreviated name fields in the **PERSON** table (*FAMILY_NAME*, *FIRST_INITIAL*, *MIDDLE_INITIAL*, and *UNSPLIT_NAME*) are parsed from the "AU" field. This article describes how the "AU" field is parsed. Note: Even when the "AF" field is supplied, the *FAMILY_NAME* is parsed out of the "AU" field.

The full name fields in the **PERSON** table (*ADDITIONAL_NAME*, *FULL_NAME*, and *PERSONAL_NAME*) are parsed from the "AF" field. This article describes how the "AF" field is parsed.

Also see AUTHORS, DOCUMENT, EDITORS, REFERENCE, How Abbreviated Names are Parsed, and How Full Names are Parsed for additional information.

PUBLISHER

This table contains a limited set of data on document publishers. It is possible for publisher addresses to be specified in the "PA" ISI field. Those are contained in the **ADDRESS** table and specified as Publisher Addresses by their entries in the **PUBLISHER_ADDRESSES** table.

This table has the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *PUBLISHER_CITY*: The exact string as found in the "PI" field.
- *PUBLISHER_NAME*: The exact string as found in the "PU" field.
- *PUBLISHER_SOURCE_FK*: If a source was found in the "SO"/"J9"/etc. fields, this field contains the foreign key into the **SOURCE** table of that Source.
- *PUBLISHER_WEB_ADDRESS*: The exact string as found in the "WP" field.

See SOURCE, ADDRESS, and PUBLISHER_ADDRESSES.

PUBLISHER_ADDRESSES

This table contains neither actual publisher nor address data. Rather, it is a relationship table between **PUBLISHER** and **ADDRESS**. Each record in this table describes an "Address of Publisher" relationship.

Publisher Addresses are specified in the "PA" ISI field. Since there is at most one Publisher per Document (specified in the "PU" ISI field), Publisher Addresses specified for a given Document can accurately be linked to the correct Publishers.

From what we have found, only one Publisher Address is ever specified per Publisher. The behavior of the loader reflects this.

This table contains the following fields:

- *PUBLISHER_ADDRESS_PUBLISHER_FK*: The foreign key into the **PUBLISHER** table.
- *PUBLISHER_ADDRESS_ADDRESS_FK*: The foreign key into the **ADDRESS** table.

See PUBLISHER and ADDRESS for additional details.

REFERENCE

This table contains data directly pertaining to References, which are always parsed out of the "CR" ISI field. Each subsequent reference string contains a series of tokens separated by the string ", " (a comma, then a single space character). Further down below, this article describes the limited set of reference string formats supported, which are defined by what tokens can be parsed out of them and in what positions they can be parsed.

There are various ways to determine what data a token is supposed to specify, depending on what position the token occurred in. For example, if a token that starts with the letter 'V', followed by numeric digits is in the third of four positions, those numeric digits in that token is treated as the reference volume (*REFERENCE_VOLUME*).

This table includes all fields that can possibly be parsed out of any support reference string format, which are:

- *PK*: Automatically generated. Guarantees uniqueness.
- *ANNOTATION*: It is possible for reference strings to contain Sources in various token positions (see below). These Sources are potentially prefixed with what we call annotations; similarly, it is possible for the entire Source token to just be an annotation. The set of annotations that the loader can separate from the actual Source title (if there is one) is:
    - *UNPUB*
    - *IN PRESS*
    - *PREPRINT*
    - *UNPUBLISHED*
    - *CITED INDIRECTLY*
    - *PRIVATE COMMUNICATIO*
    - *UNOPUB* (This was most likely a typo in a dataset we used for testing.)
- *REFERENCE_ARTICLE_NUMBER*: From what we have found, the standard reference string formats can contain an additional token at the end, which is the article number of the document being referenced. This can be used to match References to Documents (as a "is a" relationship).
- *REFERENCE_AUTHOR_FK*: It is possible for reference strings to contain an Author in various token positions (see below). When an Author is specified, it is always treated as the first author. The string specifying the Author is always considered to specify the unsplit abbreviated author name, and it is parsed as such.
- *DIGITAL_OBJECT_IDENTIFIER*: If you examine below, the only supported reference string format that contains 6 tokens has a Digital Object Identifier token in last/sixth position. A valid Digital Object Identifier begins with the string "DOI " and has the subsequent format "***/***", where *s are wildcards. The actual Java regular expression used to match that subsequent format is: "***/.***".
- *REFERENCE_OTHER_INFORMATION*: For lack of a better name, this is other information that may be specified at the end of any of the formats (similarly to Article Numbers). We determine if other information is specified if the last token starts with one of the following prefixes:
    - *PII*
    - *PMID*
    - *UNSP*
- *REFERENCE_PAGE_NUMBER*: When part of a support format, a page number token starts with "p" or "P", followed by a series of numeric digits. This field contains those numeric digits as an integer.
- *PAPER_FK*: After all references and documents have been parsed, it is possible to match references to documents if they have matching article numbers and/or digital object identifiers. When such matches have been made, a link is made from the Reference and Document that basically states that the Reference is the Document. This field contains the foreign key into the **DOCUMENT** table in such a case.
- *RAW_REFERENCE_STRING*: The exact string as found in the "CR" field.
- *REFERENCE_VOLUME*: When part of a support format, a volume token starts with "v" or "V", followed by a series of numeric digits. This contains those numeric digits as an integer.
- *REFERENCE_SOURCE_FK*: It is possible for reference strings to contain a Source in various token positions (see below). When a Source is specified, it is always assumed that the abbreviated name is provided (what appears in the "J9" field). This field contains the foreign key into the **SOURCE** table in such a case.
- *REFERENCE_WAS_STARRED*: This field is obsolete and will always be empty.

The following reference string formats are supported, organized by number of tokens present:

- 2 tokens:
    - year, source
    - author, source

- 3 tokens:
  - year, source, volume or page number
  - person, source, volume or page number
- 4 tokens:
  - year, source, volume, page number
  - person, year, source, volume or page number
  - person, source, volume, page number
- 5 tokens:
  - person, year, journal, volume, page number
- 6 tokens:
  - person, year, source, volume, page number, digital object identifier

In the various built-in extractions, when References are matched to Documents, they are considered as part of the set of inner documents. When they are not matched to Documents, they are considered as part of the set of outer documents.

See CITED_REFERENCES, DOCUMENT, PERSON, SOURCE, and How Abbreviated Names are Parsed.

REPRINT_ADDRESSES

This table contains neither actual document nor address data. Rather, it is a relationship table between **DOCUMENT** and **ADDRESS**. Each record in this table describes an "Address at which Document can be reprinted" relationship.

Reprint Addresses are specified in the "RP" ISI field.

From what we have found, only one Reprint Address is ever specified per Document. The behavior of the loader reflects this.

This table contains the following fields:

- *REPRINT_ADDRESS_DOCUMENT_FK*: The foreign key into the **DOCUMENT** table.
- *REPRINT_ADDRESS_ADDRESS_FK*: The foreign key into the **ADDRESS** table.

See DOCUMENT and ADDRESS for additional details.

RESEARCH_ADDRESSES

This table contains neither actual document nor address data. Rather, it is a relationship table between **DOCUMENT** and **ADDRESS**. Each record in this table describes an "Address at which research for Document was done" relationship.

Research Addresses are specified in the "C1" ISI field.

From what we have found, multiple Research Addresses can be specified per Document. The behavior of the loader reflects this.

This table contains the following fields:

- *RESEARCH_ADDRESS_DOCUMENT_FK*: The foreign key into the **DOCUMENT** table.
- *RESEARCH_ADDRESS_ADDRESS_FK*: The foreign key into the **ADDRESS** table.

See DOCUMENT and ADDRESS for additional details.

SOURCE

Sources are commonly journals, but they are not restricted to being journals. This table contains data directly pertaining to any and all types of Sources we know that ISI data describes. It has the following fields:

- *PK*: Automatically generated.
- *BOOK_SERIES_TITLE*: If this Source is a book in a series, this may be specified as the title of that series. It is the exact string as found in the "SE" field.
- *BOOK_SERIES_SUBTITLE*: If this Source is a book in a series, this may be specified as the subtitle of that series. It is the exact string as found in the "BS" field.
- *CONFERENCE_HOST*: If this Source is a conference that took place, this may be specified as the host of that conference. It is the exact string as found in the "HO" field.
- *CONFERENCE_LOCATION*: If this Source is a conference that took place, this may be specified as the location at which that conference took place. It is the exact string as found in the "CL" field.
- *CONFERENCE_SPONSORS*: If this Source is a conference that took place, this may be specified as one or more sponsors of that conference. It is the exact string as found in the "SP" field.
- *CONFERENCE_TITLE*: If this Source is a conference that took place, this may be specified as the title of that conference. It is the exact string as found in the "CT" field.
- *FULL_TITLE*: The exact string as found in the "SO" field.
- *ISO_TITLE_ABBREVIATION*: The exact string as found in the "JI" field.
- *ISSN*: The exact string as found in the "SN" field.
- *TWENTY_NINE_CHARACTER_SOURCE_TITLE_ABBREVIATION*: The exact string as found in the "J9" field.

Sources are guaranteed to be unique by the combination of all of their fields, not just their primary keys. So for example, even though Sources used to be merged by their matching "J9" fields, they will now not be matched unless all of their other fields are exactly the same (including empty).

In the various built-in extractions, when Sources are specified in the standard source fields ("J9", "SO", etc.), they are considered to be inner sources. When they are specified as part of reference strings and are not merged, they are considered to be outer sources.

See DOCUMENT and REFERENCE.

How Abbreviated Names are Parsed

Abbreviated names appear in two basic forms:

- <Last Name>, <Initials>
- <Last Name> <Initials>

(The only real difference is the comma.)

When parsing an abbreviated name, it is first split into one or two tokens, depending on what data was supplied. The first token is always considered to be the last name (**PERSON**.*FAMILY_NAME*). If a second token is supplied, it is considered to be the initials. The first character of this token is capitalized and treated as the first initial (**PERSON**.*FIRST_INITIAL*). Any subsequent characters of this token are capitalized and treated as the middle initial(s) (**PERSON**.*MIDDLE_INITIAL*).

All values that end up in **PERSON**.*UNSPLIT_NAME* as a result of this loader are converted to a canonical form, which is of the form:

- <Last Name>, <Initials>

Note: Titles such as Mr., Ms., Mrs., etc are not directly supported, in that there are no special parsing rules to handle them.

See AUTHORS, EDITORS, PERSON, REFERENCE, and How Full Names are Parsed.

How Full Names are Parsed

Full names appear in two basic forms:

- <Last Name>, <First and possibly Middle Names>
- <Last Name> <First and possibly Middle Names>

(The only real difference is the comma.)

When parsing a full name, it is first split into tokens. The first token is always considered to be the last name (which corresponds to the **PERSON**.*FAMILY_NAME* field, though practically speaking that field is always filled by the parsing results of the "AU" or "ED" ISI fields). If a second token is supplied, it is considered to be the first name (**PERSON**.*PERSONAL_NAME*). All tokens after the second are concatenated together, separated by a single space character, and treated as the middle name (**PERSON**.*ADDITIONAL_NAME*).

All values that end up in **PERSON**.*FULL_NAME* as a result of this loader are converted to a canonical form, which is of the form:

- <Last Name>, <First Name> <Middle Names>

where <Middle Names> is all tokens after the second concatenated as described above.

Note: Titles such as Mr., Ms., Mrs., etc are not directly supported, in that there are no special parsing rules to handle them.

See AUTHORS, EDITORS, PERSON, REFERENCE, and How Abbreviated Names are Parsed.

NSF Database Tables

- AWARD
- FIELD_OF_APPLICATION
- NSF_FILE
- ORGANIZATION
- PERSON (NSF)
- INVESTIGATORS
- PROGRAM_MANAGERS
- PROGRAM
- AWARD_OCCURRENCES
- AWARD_FIELD_OF_APPLICATIONS
- INVESTIGATOR_ORGANIZATIONS
- PROGRAM_NAME_AND_ELEMENT_CODES
- PROGRAM_REFERENCE_CODES

AWARD

This table contains data directly pertaining to awards. It includes all fields relevant for all possible (known) types of awards in NSF file.

All "arbitrary" fields found in the original NSF dataset are also appended to this table, as this table is somewhat central to the database schema.

In addition to "arbitrary" fields, this table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *ABSTRACT_TEXT*: The exact string as found in the "Abstract" field.
- *AWARD_NUMBER*: The exact string as found in the "Award Number" field.
- *RAW_START_DATE*: The exact string as found in the "Start Date" field.
- *AWARD_START_DATE*: The contents of the "Start Date" field parsed as a java date object.
- *RAW_EXPIRATION_DATE*: The exact string as found in the "Expiration Date" field.

- *EXPIRATION_DATE*: The contents of the "Expiration Date" field parsed as a java date object.
- *RAW_LAST_AMENDMENT_DATE*: The exact string as found in the "Last Amendment Date" field.
- *LAST_AMENDMENT_DATE*: The contents of the "Last Amendment Date" field parsed as a java date object.
- *RAW_AWARDED_AMOUNT_TO_DATE*: The exact string as found in the "Awarded Amount to Date" field.
- *AWARDED_AMOUNT_TO_DATE*: The contents of the "Awarded Amount to Date" field parsed as a double.
- *AWARD_INSTRUMENT*: The exact string as found in the "Award Instrument" field.
- *NSF_ORGANIZATION*: The exact string as found in the "NSF Organization" field.
- *AWARD_INSTRUMENT*: The exact string as found in the "Award Instrument" field.

The parsing of "Award" element is the first step when a new NSF row is encountered. It's primary key is used as a foreign key in other relationship tables like INVESTIGATORS, PROGRAM_MANAGERS, AWARD_FIELD_OF_APPLICATIONS, PROGRAM_NAME_AND_ELEMENT_CODES and PROGRAM_REFERENCE_CODES.

FIELD_OF_APPLICATION

This table contains fields of application that are associated with a particular award. The associated data is parsed from "Field Of Application(s)" field in the NSF data.

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *ORIGINAL_INPUT_FIELD*: The exact string as found in the "Field Of Application(s)" field.
- *EXTRACTED_NUMERIC_FIELD*: Field of application can be split up into Numeric & non-numeric part. The value of this field is the extracted numeric field.
- *EXTRACTED_TEXT_FIELD*: The value of this field is the extracted non-numeric field.

It's primary key is used as a foreign key in it's relationship table with Award; see AWARD_FIELD_OF_APPLICATIONS for additional information.

NSF_FILE

This table contains originating NSF file metadata. This table may be useful for things like determining which NSF datasets awards occurred in if/when multiple separate NSF files being loaded/merged into one database is supported. Currently, however, multiple separate NSF files being loaded/merged into one database is not supported.

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *FILE_MD5_CHECKSUM*: This is the value of the MD5 checksum of the entire NSF file selected for loading.
- *FILE_NAME*: Generated based on the NSF dataset the user chose in the Data Manager from within the tool.
- *FILE_TYPE*: Currently we have hard coded this value to be "NSF File". We had tried an automatic file type detection method earlier but it was a slow process and did not work properly on all the platforms.

Every combination of all the fields in this table guarantees uniqueness.

See AWARD_OCCURRENCES for additional details.

ORGANIZATION

Organization related information is stored in this table. Since the NSF format already provides separate fields having address information, it is easy to save this data in the table.

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *ORGANIZATION_NAME*: The exact string as found in the "Organization" field.
- *ORGANIZATION_PHONE*: The exact string as found in the "Organization Phone" field.

- *ORGANIZATION_STREET_ADDRESS*: The exact string as found in the "Organization Street Address" field.
- *ORGANIZATION_STATE*: The exact string as found in the "Organization State" field.
- *ORGANIZATION_ZIP*: The exact string as found in the "Organization Zip" field.
- *ORGANIZATION_CITY*: The exact string as found in the "Organization City" field.

It's primary key is used as a foreign key in it's relationship table with Investigators; see INVESTIGATOR_ORGANIZATIONS for additional information.

PERSON (NSF)

This table contains the full names of people, both in unsplit and split forms. People can be parsed out of following originating fields:

- Principal Investigator
- Co-PI Name(s)
- Program Manager

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *LAST_NAME*: Also known as "family name."
- *FIRST_NAME*: Also known as "personal name."
- "MIDDLE_INITIAL'
- *ORIGINAL_INPUT_NAME*: The exact (unsplit/unparsed) string as found in the originating field.
- *FORMATTED_FULL_NAME*: It is not always the exact (unsplit/unparsed) string as found in the originating field. The value in this field will always contain the information in the same syntax, which is - <Last Name>, <First Name> <Middle Initial>.

The patterns that can be properly parsed into respective respective fields are as follows:

- <LastName>, <FirstName>
- <LastName>, <FirstInitial>. <FirstName>
- <LastName>, <FirstName> <(AlternativeName)>
- <FirstName> <LastName>
- <FirstName> <MiddleInitial>. <LastName>
- <FirstName> <MiddleInitial> <LastName>

These patterns were derived by analyzing name generating fields from a large corpus of NSF files. We are working on making the Person name parser more complete in future versions of NSF Loader.

People are never merged during loading. Automatic People merging can be done via the Merge Identical NSF People algorithm.

Also see INVESTIGATORS and PROGRAM_MANAGERS for additional information.

INVESTIGATORS

This table contains specific information related to Principal Investigators (PIs) & Co-PIs. Additionally it is also contains relationship information that it shares with "PERSON" and "AWARD". Each record in this table also describes a "Person was the PI(or Co-PI) of Award" relationship.
As mentioned, information related to both PIs & Co-PIs is captured in this table. Although there is more information available for PIs like email address and state. A boolean field is used to differentiate between the current element being a PI or a Co-PI.
This table contains the following fields:

- _I_AWARD_FK_: The foreign key into the **AWARD** table.
- _I_PERSON_FK_: The foreign key into the **PERSON** table.

- *EMAIL_ADDRESS*: The exact string as found in the "PI Email Address" field. This is left blank for when the current element is a Co-PI.
- *STATE*: The exact string as found in the "State" field. This is left blank for when the current element is a Co-PI.
- *IS_MAIN_PI*: This is a boolean which is set to "TRUE" when the current element is a PI and "FALSE" otherwise. See AWARD and PERSON (NSF) for further reference.

PROGRAM_MANAGERS

This table does not contain actual program manager specific data. Rather, it is a relationship table between **AWARD** and **PERSON**. Each record in this table describes a "Person was the program manager of Award" relationship.

This table contains the following fields:

- *PM_AWARD_FK*: The foreign key into the **AWARD** table.
- *PM_PERSON_FK*: The foreign key into the **PERSON** table.

See AWARD and PERSON for further reference.

PROGRAM

This table is a mapping of Program element code to Program name. There is a direct co-relation between values in the "Program(s)" & "Program Element Code(s)" fields present in the NSF data. Also there is "Program Reference Code(s)" field in the NSF data that is used to provide program element codes.

This table contains the following fields:

- *PK*: Automatically generated. Guarantees uniqueness.
- *PROGRAM_NAME*: The exact string as found in the "Program(s)" field.
- *FUNDING_CODE*: The exact string as found in the "Program Element Code(s)" and/or "Program Reference Code(s)" field.

Programs with the same "name" and "funding code" are automatically merged to be the same program. Since "Program Element Code" and "Reference Code" are essentially part of the same NSF Funding Code Schema, while parsing the "Program Reference Code(s)" column we check if the code is already present in the "Program" funding code. If it is, then we will just get the reference to that object and save it in an appropriate relationship table. On another note if it is not present then we create a new Program record without the Name field.

Also see PROGRAM_NAME_AND_ELEMENT_CODES and PROGRAM_REFERENCE_CODES for further reference.

AWARD_OCCURRENCES

This table contains neither actual award data nor NSF file metadata. Rather, it is a relationship table between **AWARD** and **NSF_FILE**. Each record in this table describes a "Award occurred in NSF File" relationship.

This table contains the following fields:

- *AO_AWARD_FK*: The foreign key into the **AWARD** table.
- *AO_NSF_FILE_FK*: The foreign key into the **NSF_FILE** table.

See AWARD and NSF_FILE for further reference.

AWARD_FIELD_OF_APPLICATIONS

This table contains neither actual award data nor field of applications. It is a relationship table between **AWARD** and **FIELD_OF_APPLICATION**. Each record in this table describes a "Award has a Field of Application" relationship.

This table contains the following fields:

- *FOAS_AWARD_FK*: The foreign key into the **AWARD** table.
- *FOAS_FIELD_OF_APPLICATION_FK*: The foreign key into the **FIELD_OF_APPLICATION** table.

See AWARD and FIELD_OF_APPLICATION for further reference.

INVESTIGATOR_ORGANIZATIONS

This table contains neither actual investigator data nor organization. It is a relationship table between **INVESTIGATORS** and **ORGANIZATION**. Each record in this table describes a "Investigator belongs to an Organization" relationship.

This table contains the following fields:

- *IOS_INVESTIGATOR_FK*: The foreign key into the **INVESTIGATORS** table.
- *IOS_ORGANIZATION_FK*: The foreign key into the **ORGANIZATION** table.

See INVESTIGATORS and ORGANIZATION for further reference.

PROGRAM_NAME_AND_ELEMENT_CODES

This table does not contain "program" information. It is a relationship table between **PROGRAM** and **AWARD**. Each record in this table describes a "Program belongs to an Award" relationship. The "Program" information is captured from "Program Element Code(s)" field in the NSF Data.

This table contains the following fields:

- *PNEC_PROGRAM_FK*: The foreign key into the **PROGRAM** table.
- *PNEC_AWARD_FK*: The foreign key into the **AWARD** table.

See AWARD and PROGRAM for further reference.

PROGRAM_REFERENCE_CODES

This table does not contain "program" information. It is a relationship table between **PROGRAM** and **AWARD**. Each record in this table describes a "Program belongs to an Award" relationship.

This table contains the following fields:

- *PRC_PROGRAM_FK*: The foreign key into the **PROGRAM** table.
- *PRC_AWARD_FK*: The foreign key into the **AWARD** table.

The "Program" information is captured from "Program Reference Code(s)" field in the NSF Data. Since the originating field only has "funding code" information first we try to check if any of the existing "Program" have the same funding code. If yes, then we will just get the reference to that object and save it in an appropriate relationship table. On another note if it is not present then we create a new "program" record without the "name" field.

See AWARD and PROGRAM for further reference

Publication Database Tables

- Bridge Tables — Bridge tables are those that primarily hold relationship information about a publication, compared to Core Tables which primarily hold the data.
  - Authors Table — Links a document to its authors.
  - Cited Citations Table — Links a documents to its citations.
  - Cited Patents — Links a document to cited patents.
  - Document Keywords Table — Links a document to its keywords.

    - Document Occurrences Table — Links a document to the files that list it.
    - Editors Table — Links a document to its editors.
    - Publisher Addresses Table — Links a publisher to its address(es).
    - Reprint Addresses Table — Links a document to its reprint address(es).
    - Research Addresses — Links a document to its research address(es).
- Core Tables — Core tables are those that primarily hold the data about a publication, compared to Bridge Tables which primarily show the relationship between the data held in the core tables.
    - Addresses Table — The addresses table includes information about an address like city and country.
    - Citations Table — The citations table holds the information about a citation of a document including the volume, author, and year.
    - Documents Table — The documents table holds the information that describes a document like the title and abstract.
    - Files Table — The files table holds the information about the file that was used to create the database.
    - Keywords Table — The keywords table holds the information about a keywords of a document.
    - Patents Table — The patents table holds the information about a patent.
    - People Table — The people table includes information about authors, editors, and other people associated with a publication.
    - Publishers Table — The publishers table includes information about a publication's publisher like name and city.
    - Sources Table — The people table includes information about a publication's source like the full title, issn, and publication type.

Bridge Tables

Bridge Tables

Bridge tables are those that primarily hold relationship information about a publication, compared to Core Tables which primarily hold the data.

- Authors Table — Links a document to its authors.
- Cited Citations Table — Links a documents to its citations.
- Cited Patents — Links a document to cited patents.
- Document Keywords Table — Links a document to its keywords.
- Document Occurrences Table — Links a document to the files that list it.
- Editors Table — Links a document to its editors.
- Publisher Addresses Table — Links a publisher to its address(es).
- Reprint Addresses Table — Links a document to its reprint address(es).
- Research Addresses — Links a document to its research address(es).

Authors Table

Authors Table

Links a document to its authors.

Columns

Name	Type	Description	Notes
document_id	integer	PK of the document in the documents table	see Documents Table

person_id	integer	PK of the author in the people table	see People Table
email_address	integer	E-mail address given for the person in the document	
order_listed	integer	Author list position for the person in the document	

Source

For implementation details or more complete documentation, please see the source.

See Also

Cited Citations Table

Cited Citations Table

Links a documents to its citations.

Columns

Name	Type	Description	Notes
document_id	integer	PK of the document in the documents table	See Documents Table
citation_id	integer	PK of the reference in the references table	See Patents Table

Source

For implementation details or more complete documentation, please see the source.

See Also

Cited Patents

Cited Patents Table

Links a document to cited patents.

Columns

Name	Type	Description	Notes
document_id	integer	PK of the document in the documents table	See Documents Table

patent_id	integer	PK of the patent in the patents table	See Patents Table

Source

For implementation details or more complete documentation, please see the source.

See Also

Document Keywords Table

Document Keywords Table

Links a document to its keywords.

Columns

Name	Type	Description	Notes
document_id	integer	PK of the document in the documents table	See Documents Table
keyword_id	integer	PK of the keyword in the keywords table	See Keywords Table
ORDER_LISTED	integer	The order in which the keyword appeared in the document	

Source

For implementation details or more complete documentation, please see the source.

See Also

Document Occurrences Table

Document Occurrences Table

Links a document to the files that list it.

Columns

Name	Type	Description	Notes
document_id	integer	PK of the document in the documents table	See Documents Table
file_id	integer	PK of the file in the files table	See Files Table

Source

For implementation details or more complete documentation, please see the source.

See Also

Editors Table

Editors Table

Links a document to its editors.

Columns

Name	Type	Description	Notes
document_id	integer	PK of the document in the documents table	See Documents Table
person_id	integer	PK of the editor in the people table	See People Table
order_listed	integer	The order in which the editor was listed in the document	

Source

For implementation details or more complete documentation, please see the source.

See Also

Publisher Addresses Table

Publisher Addresses Table

Links a publisher to its address(es).

Columns

Name	Type	Description	Notes
publisher_id	integer	PK of the publisher in the publishers table	See Publishers Table
address_id	integer	PK of the address in the addresses table	See Addresses Table

Source

For implementation details or more complete documentation, please see the source.

See Also

Reprint Addresses Table

Reprint Addresses Table

Links a document to its reprint address(es).

Columns

Name	Type	Description	Notes
document_id	integer	PK of the document in the documents table	See Documents Table
address_id	integer	PK of the address in the addresses table	See Addresses Table

Source

For implementation details or more complete documentation, please see the source.

See Also

Research Addresses

Research Addresses Table

Links a document to its research address(es).

Columns

Name	Type	Description	Notes
document_id	integer	PK of the document in the documents table	See Documents Table
address_id	integer	PK of the address in the addresses table	See Addresses Table
order_listed	integer	The order in which the address was listed in the document	

Source

For implementation details or more complete documentation, please see the source.

See Also

Core Tables

Core Tables

Core tables are those that primarily hold the data about a

publication, compared to Bridge Tables which primarily show the relationship between the data held in the core tables.

- Addresses Table — The addresses table includes information about an address like city and country.
- Citations Table — The citations table holds the information about a citation of a document including the volume, author, and year.
- Documents Table — The documents table holds the information that describes a document like the title and abstract.
- Files Table — The files table holds the information about the file that was used to create the database.
- Keywords Table — The keywords table holds the information about a keywords of a document.
- Patents Table — The patents table holds the information about a patent.
- People Table — The people table includes information about authors, editors, and other people associated with a publication.
- Publishers Table — The publishers table includes information about a publication's publisher like name and city.
- Sources Table — The people table includes information about a publication's source like the full title, issn, and publication type.

Addresses Table

Addresses Table

The addresses table includes information about an address like city and country.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
city	varchar	City	Currently unparsed. Only the `raw_address` field is used.
country	varchar	Country	Currently unparsed. Only the `raw_address` field is used.
postal_code	varchar	Postal or ZIP code	Currently unparsed. Only the `raw_address` field is used.
raw_address	varchar	Raw address data	
state_or_province	varchar	State or province	Currently unparsed. Only the `raw_address` field is used.
street_address	varchar	Street address	Currently unparsed. Only the `raw_address` field is used.

Source

For implementation details or more complete documentation, please see the source.

See Also

- Publisher Addresses Table
- Reprint Addresses Table
- Research Addresses

Citations Table

Citations Table

The citations table holds the information about a citation of a document including the volume, author, and year.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
annotation	varchar	Annotation to the reference	
article_number	varchar	Reference article number	
author_id	varchar	The PK of the first author of the referenced document	
digital_object_identifier	varchar	DOI of the referenced document	
other_information	varchar	Other information about the reference	
page_number	integer	Page number referenced	
document_id	integer	The PK of the document being referenced in the documents table	
raw_citation	varchar	The raw reference string from the original document	
source_id	integer	The PK of the source of the referenced document	
year	integer	The year that the referenced document was published	
starred	varchar	Denotes whether the reference was starred	

volume	integer	Volume number	The volume is an `integer` field which causes problems with certain publication types, including Medline and Scopus, which include letters in their volume information.

Source

For implementation details or more complete documentation, please see the source.

See Also

Cited Citations Table

Documents Table

Documents Table

The documents table holds the information that describes a document like the title and abstract.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
abstract	varchar	Abstract of the document	
document_number	varchar	Article number for new APS journals	
beginning_page	integer	First page of the document	
cited_citation_count	integer	Number of references cited by the document	
digital_object_identifier	varchar	DOI of the document	
isi_type	varchar	Type of document	
volume	varchar	Volume number of the document	
ending_page	integer	Last page of the document	
first_author_id	integer	PK of the first author from the people table	
funding_agency_and_grant_number	varchar	Funding information	

funding_text	varchar	Raw text with funding information	
isbn	varchar	ISBN of the document	
isi_document_delivery_number	varchar	ISI document delivery number	
isi_unique_article_identifier	varchar	ISI unique article identifier	
issue	varchar	Issue	
language	varchar	Language	
page_count	integer	Number of pages	
part_number	varchar	Part number	
publication_date	varchar	Date of publication	
publication_year	integer	Year of publication	
source_id	integer	PK of the source	
special_issue	varchar	Denotes a special issue	
subject_category	varchar	List of related categories	
supplement	varchar	Name of the supplement where published	
times_cited	integer	Number of times cited	
title	varchar	Title	
cite_as	varchar	Preferred citation string	

Source

For implementation details or more complete documentation, please see the source.

See Also

- Authors Table
- Cited Citations Table
- Cited Patents
- Document Keywords Table
- Document Occurrences Table
- Editors Table
- Reprint Addresses Table
- Research Addresses

Files Table

Files Table

The files table holds the information about the file that was

used to create the database.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
format_version_number	varchar	What version of the standardized format is the file using	
name	varchar	Name of the file	
file_type	varchar	Type of file	

Source

For implementation details or more complete documentation, please see the source.

See Also

Document Occurrences Table

Keywords Table

Keywords Table

The keywords table holds the information about a keywords of a document.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
name	varchar	Keyword	
type	varchar	Type of keyword	E.g. `medline MeSH Terms`

Source

For implementation details or more complete documentation, please see the source.

See Also

Document Keywords Table

Patents Table

Patents Table

The patents table holds the information about a patent.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
number	varchar	Patent number	

Source

For implementation details or more complete documentation, please see the source.

See Also

Cited Citations Table
Cited Patents

People Table

People Table

The people table includes information about authors, editors, and other people associated with a publication.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
additional_name	varchar	Middle name or names	
family_name	varchar	Family name of the individual	
first_initial	varchar	First letter of the personal name	
full_name	varchar	Complete name	
middle_initial	varchar	First letter of the middle name	
first_name	varchar	Personal name of the individual	
raw_name	varchar	Raw name before processing	

Source

For implementation details or more complete documentation, please see the source.

See Also

Authors Table
Editors Table

Publishers Table

Publishers Table

The publishers table includes information about a publication's publisher like name and city.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
city	varchar	City	
name	varchar	Name	
source_id	integer	PK of the source	
web_address	varchar	Web address	

Source

For implementation details or more complete documentation, please see the source.

See Also

Publisher Addresses Table

Sources Table

Sources Table

The people table includes information about a publication's source like the full title, issn, and publication type.

Columns

Name	Type	Description	Notes
PK	integer	Primary key	
book_series_title	varchar	Title of the series that the source is part of	
book_series_subtitle	varchar	Subtitle of the series that the source is part of	
conference_host	varchar	Host of the conference	
conference_location	varchar	Location of the conference	
conference_sponsors	varchar	List of sponsors of the conference	

conference_title	varchar	Title of the conference	
conference_dates	integer	Conference dates	
full_title	varchar	Full title	
iso_title_abbreviation	varchar	ISO abbreviation of the title	
issn	varchar	ISSN	
publication_type	varchar	Type of publication	
twenty_nine_character_source_abbreviation	varchar	29 character source abbreviation	

Source

For implementation details or more complete documentation, please see the source.

See Also

2.4 Saving Visualizations for Publication

The Sci^2 Tool supports various image output formats. When in GUESS, use '*File> Export Image*' to export the current view or the complete network in diverse file formats such as jpg, png, raw, pdf, gif, etc.
To save image files created outside of GUESS, such as Circular Hierarchies or Geospatial Visualization, right-click on the PostScript file in the data manager and then click '*save*'. Select 'PostScript' and then save the file to your desired directory.

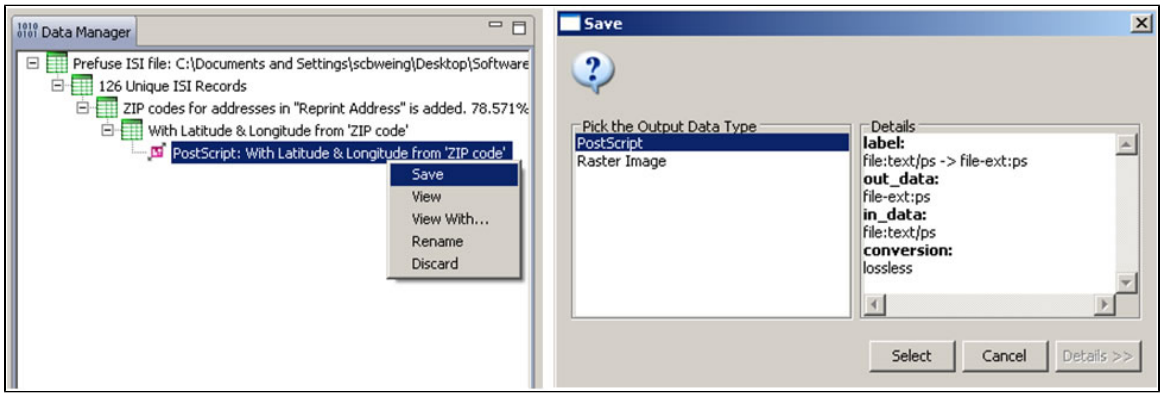


Figure 2.14: Saving a PostScript File

Adobe PostScript files can be converted to .pdf files using Adobe Distiller and viewed in Adobe Acrobat.

Alternately, PostScript files can be viewed in special interpreters, such as GSview, which requires Ghostscript software:

Ghostscript 9.01: http://pages.cs.wisc.edu/~ghost/doc/GPL/gpl901.htm
GSview 4.9: http://pages.cs.wisc.edu/~ghost/gsview/get49.htm

2.5 Sample Datasets

Meaningful analysis requires carefully collected data. The Sci^2 Tool can import many varieties of scientometric data, outlined in section 4.2 Data Acquisition and Preparation, but the tool also comes bundled with several sample

datasets. This sample data will be used throughout sections 4 Workflow Design and 5 Sample Workflows, and includes:

Datasets only available online:

- Scientometrics.isi - – All articles (2,126) published in the journal Scientometrics from 1978-2008.
- Medline_master_table.csv (most recent version is always available from the Scholarly Database site)
- Full Sample Dataset

Files available online and in the Sci2 tool (sci2/sampledata/):

- Geo
  - usptoInfluenza.csvusptoInfluenza.csv – Heavily pre-processed and geocoded data covering USPTO patents awarded containing the keyword 'Influenza'.
- Network
  - convertGraph_v0.8.0.graphml
  - kidscontest.net
  - netsci06-conference.net
  - seiyu.graphml
- Scientometrics
  - Bibtex
    - Bibsonomy.bib
    - LaszloBarabasi.bib
  - CSV
    - LaszloBarabasi.csv
  - Endnote
    - KatyBorner.enw – 146 publications authored or co-authored by Katy Börner from 1992-2010.
    - LaszloBarabasi.enw
  - ISI
    - AlessandroVespignani – 101 publications authored or co-authored by Alessandro Vespignani from 1990-2006.
    - EugeneGarfield
    - FourNetSciResearchers – 361 publications spanning 52 years and four network scientists: Albert-László Barabási, Eugene Garfield, Alessandro Vespignani, & Stanley Wasserman.This data was collected in 2007 from Web of Science.
    - LaszloBarabasi
    - StanleyWasserman
    - Test5Papers
    - ScientometricsNewFormat.isi - Papers published in *Scientometrics* from 2010.
  - Models
    - TARL
      - Agingfunction
      - iniscript.tarl
  - NIH
    - CTSA2005-2009 – 2,546 papers and 534 grants dealing with Clinical and Translational Science from 2005-2009.
  - NSF
    - BethPlale.nsf – 45 grants between three Indiana University researchers, totaling $39,031,960 from 1978-2008.
    - Cornell.nsf – Three universities' grant profiles, totaling $951,478,510 from 2000-200.
    - GeoffreyFox.nsf – 45 grants between three Indiana University researchers, totaling $39,031,960 from 1978-2008.
    - Indiana.nsf – Three universities' grant profiles, totaling $951,478,510 from 2000-200.
    - KatyBorner.nsf – 13 grants awarded from the NSF to Katy Börner as PI or Co-PI, from

2003-2008, totaling $3,527,728.
    - MedicalAndHealth.nsf – 288 grants awarded from the NSF containing the words 'Medical' or 'Health', from 2003-2010, totaling $152,015,288.
    - MichaelMcRobbie.nsf – 45 grants between three Indiana University researchers, totaling $39,031,960 from 1978-2008.
    - Michigan.nsf – Three universities' grant profiles, totaling $951,478,510 from 2000-200.
    - Stanford.nsf
- Properties
    - bibtexCoAuthorship.properties
    - endnoteCoAuthorship.properties
    - isiCoAuthorship.properties
    - isiCoCitation.properties
    - isiPaperCitation.properties
    - mergeBibtexAuthors.properties
    - mergeEndnoteAuthors.properties
    - mergeIsiAuthors.properties
    - mergeIsiPaperCitation.properties
    - mergeNsfPIs.properties
    - mergeScopusAuthors.properties
    - nsfCoPI.properties
    - nsfPIToProject.properties
    - scopusCoAuthorship.properties
- Scopus
    - BrainCancer.scopus
- SDB
    - RNAi
        - Medline_co-author_table_(nwb_format)
        - USPTO_citation_table_(nwb_format)

2.6 Visualization Color and Font Specification

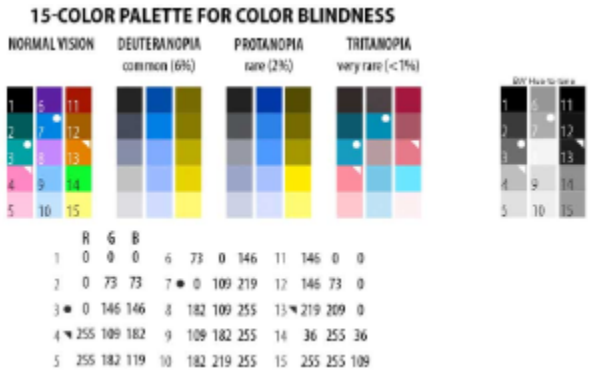
Color Specification

About The Color Selection Process

Color selection based on work by Martin Krzywinski (http://mkweb.bcgsc.ca**) and Cynthia Brewer (**http://colorbrewer2.org**).**

Temporal Bar Graph

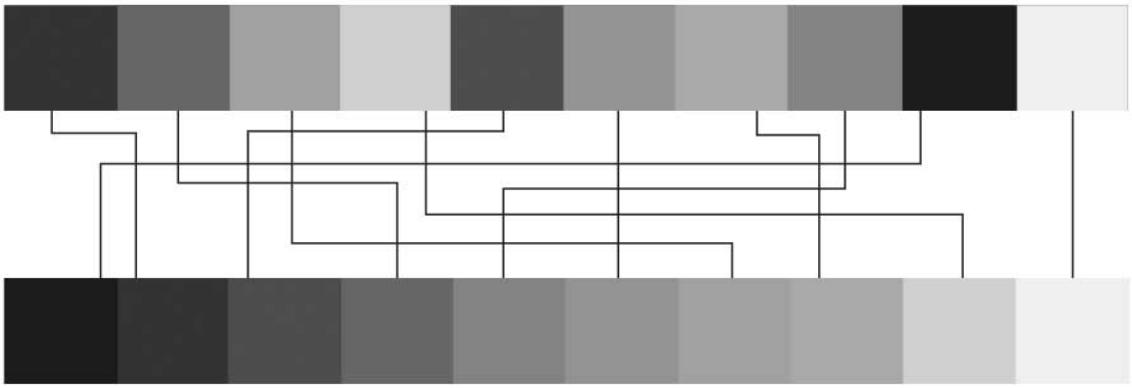
Starting off with an example of a 15-color palette that is safe for color blindness (below).



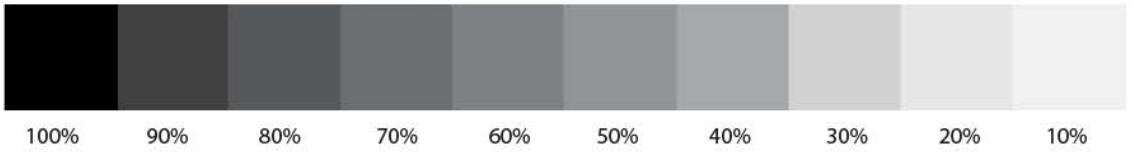
Eliminating 5 of the colors that are either too dark or too light, we get our 10-color palette for the Temporal Bar Graph.

We convert these colors to greyscale to check if there is enough contrast between shades for people that are visually challenged.

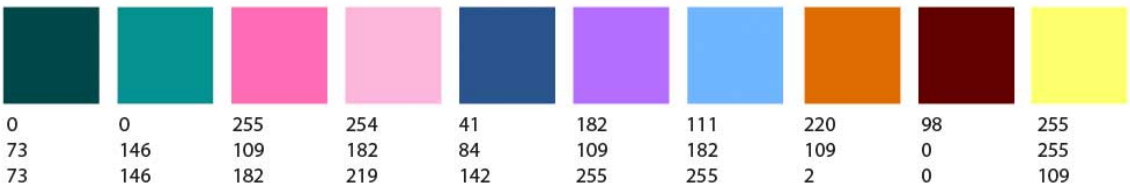
Next, we place them side by side and then arrange them from dark to light.



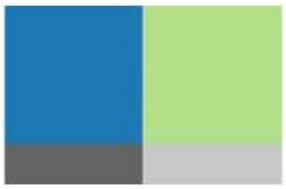
For comparison, here's a line of swatches using only variations of black and reduced by 10% in each successive swatch.



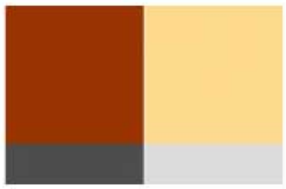
Based off of the image above, it looks like we're doing well with the original color values in the first image. Below are the swatches and RGB values for the swatches and for use in the Temporal Bar Graph.



Bipartite Network Colors



31 178
120 223
180 138

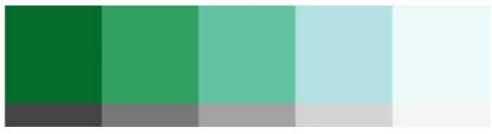


153 254
52 217
4 142

129 179
14 205
124 227

Geographical Map Colors

These are taken directly from Katy's ColorBlender examples from email

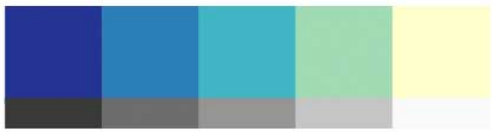


1 45 102 178 237
109 162 194 226 248
44 95 164 226 251

129 136 140 179 237
14 86 150 205 248
124 167 198 227 251

153 217 254 254 255
52 95 153 217 255
4 15 41 142 212

189 240 253 254 255
0 59 141 204 255
38 32 60 92 178



38 45 66 161 255
52 127 182 218 255
148 184 196 180 204

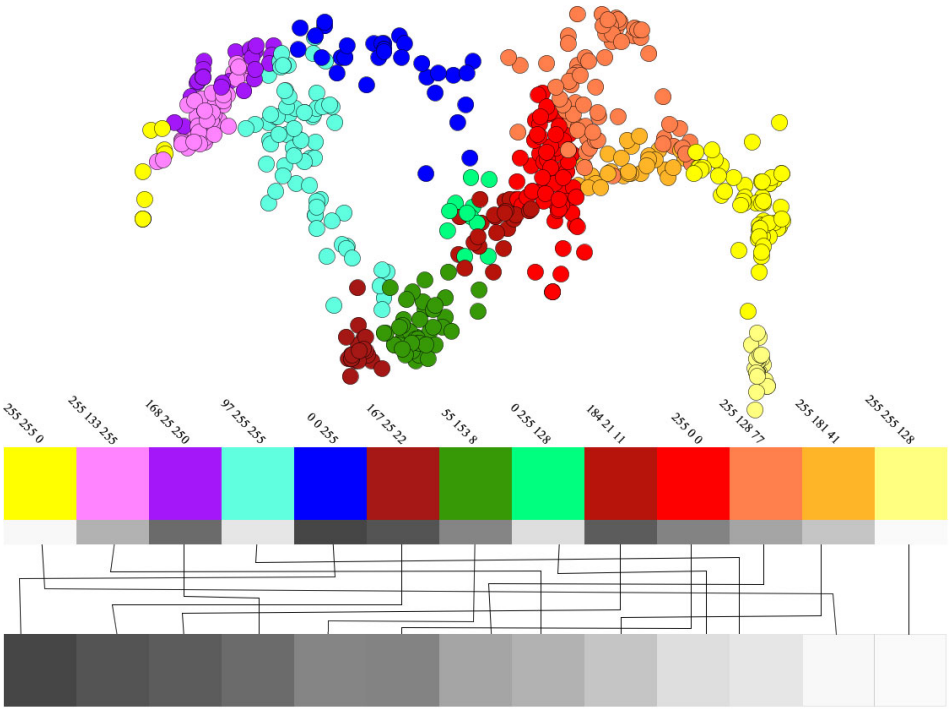
37 99 150 204 247
37 99 150 204 247
37 99 150 204 247

Map of Science Colors

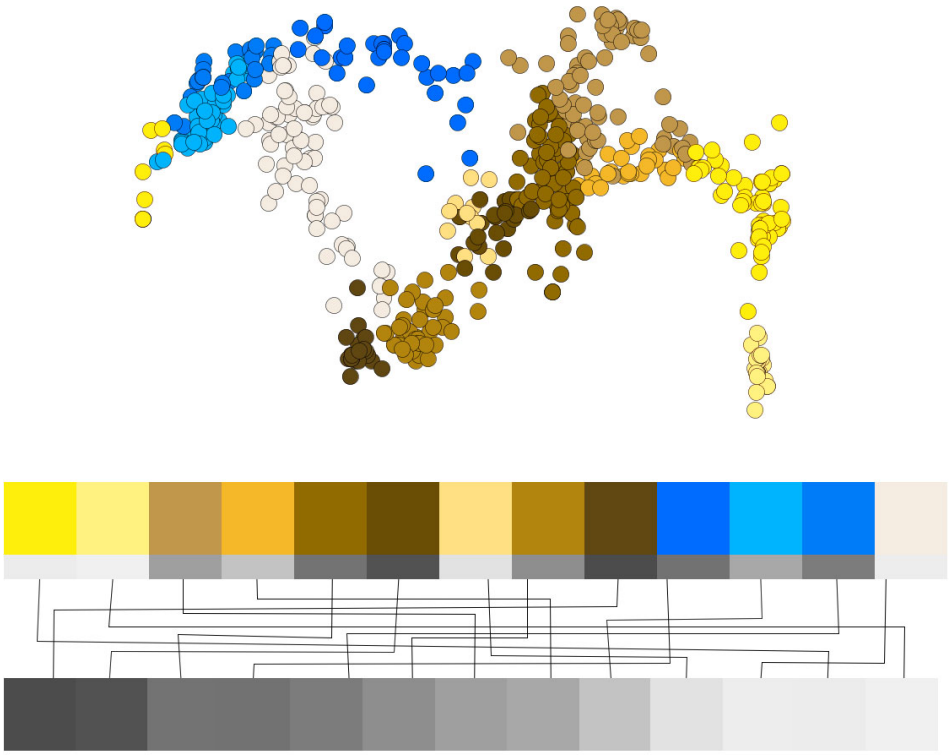
See: http://sci.cns.iu.edu/standards.html

Science Discipline	Pajek Color Name	Color	RGB
Biology	Olive Green		55, 153, 8
Biotechnology	Emerald		0, 255, 128
Brain Research	Dandelion		255, 181, 41
Chemical, Mechanical & Civil Engineering	SkyBlue		97, 255, 255
Chemistry	Blue		0, 0, 255
Earth Sciences	Mahogany		167, 25, 22
Electrical Engineering & Computer Science	Lavender		255, 133, 255
Health Professionals	Peach		255, 128, 77
Humanities	Canary		255, 255, 128
Infectious Diseases	Brick Red		184, 21, 11
Math & Physics	Mulberry		168, 25, 250
Medical Specialties	Red		255, 0, 0
Social Sciences	Yellow		255, 255, 0

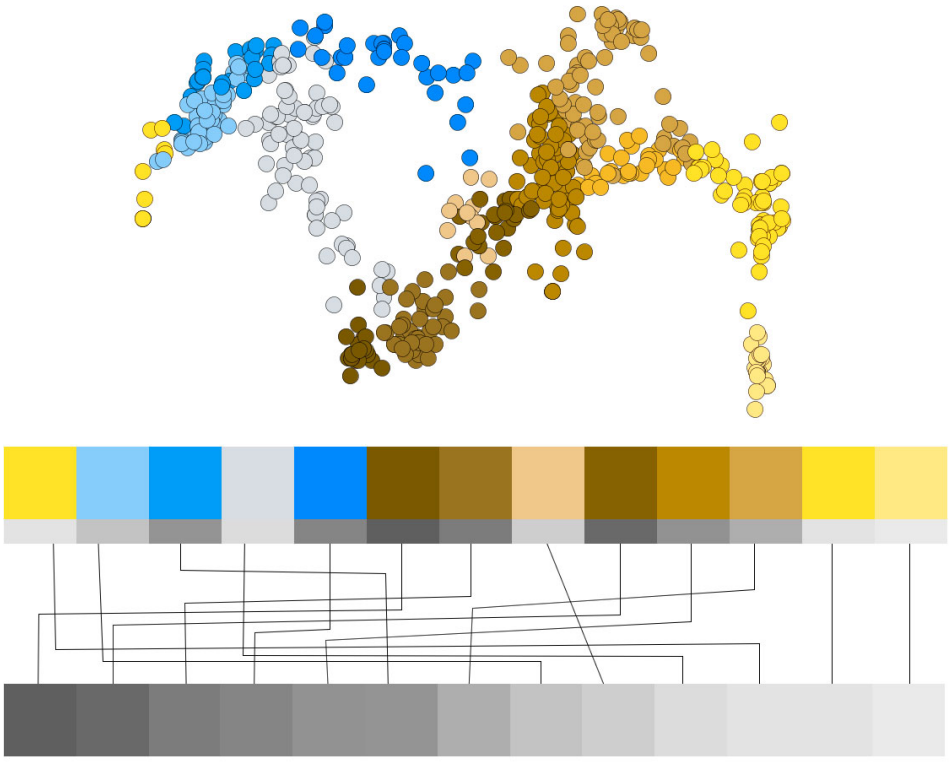
Original Map of Science Colors



P100 Comparisons

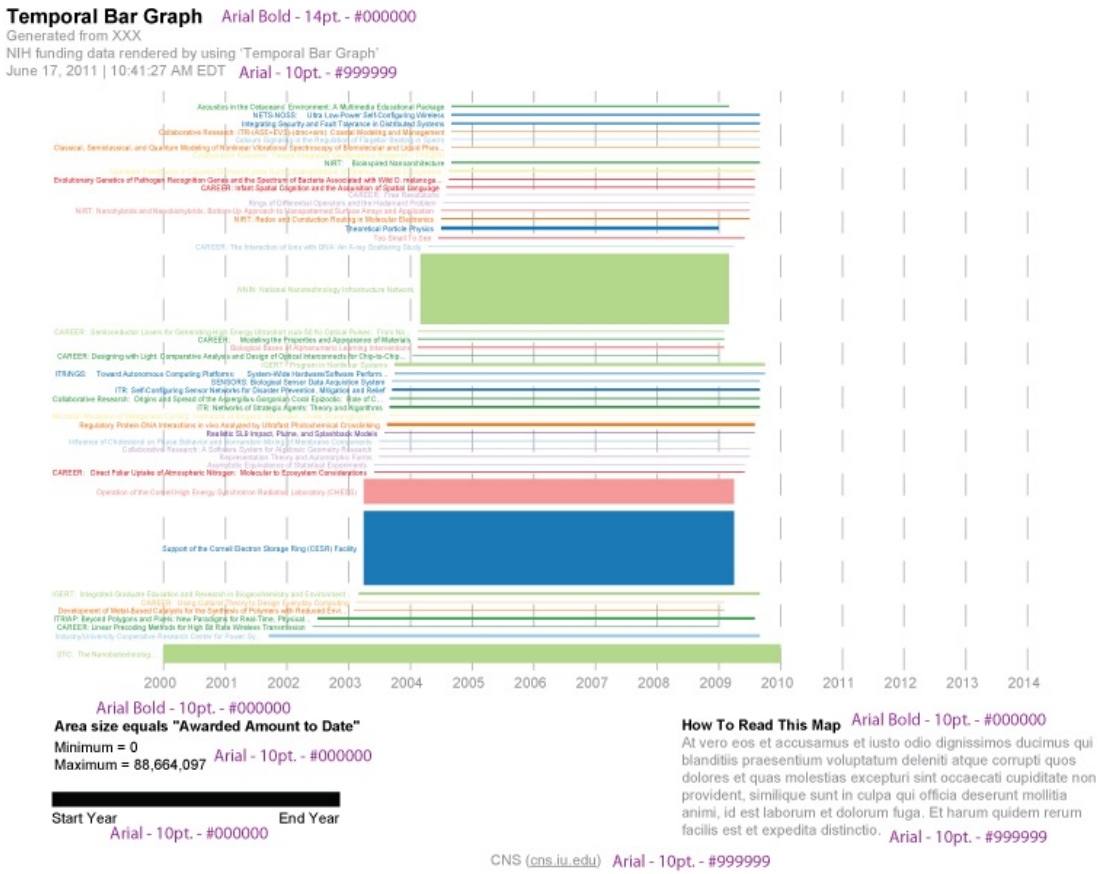


D100 Comparisons



Fonts Specification

Please see the following image for information regarding fonts and sizes used on visualizations.



3 Algorithms, Tools, and Plugins

The Sci^2 Tool menu provides easy access to diverse preprocessing, modeling, analysis, visualization, and scientometrics algorithms that are listed here. Note that the data analysis algorithms are grouped by data-type (weighted, unweighted, directed, undirected, and textual). Please see the online documentation (https://nwb.slis.indiana.edu/community/?n=Algorithms.HomePage) for more details.

3.1 Sci2 Algorithms and Tools

- **Loading Data**
  - ***Google Scholar***
- **Data Preparation**
  - *Text Files*
- **Preprocessing**
  - *General*
  - *Temporal*
  - *Geospatial*

- *Topical*
- *Networks*

- **Analysis**
  - *Temporal*
  - *Geospatial*
  - *Topical*
  - *Networks*
    - *Unweighted & Undirected*
    - *Weighted & Undirected*
    - *Unweighted & Directed*
    - *Weighted & Directed*

- **Modeling**
  - Networks

- **R**
- **Visualization**
  - General
  - Temporal
  - Geospatial
  - Networks

Loading Data

- ***Google Scholar***
  - Google Citation User ID Search Algorithm - The Google citation user ID is a web reader algorithm that retrieves the Google citation for the specified author.
  - Attach Citation BibTex File from Google Scholar - This algorithm downloads the citations listed in Google scholar for a given author in the form of a BibTex file.
  - Attach Citation Indices from Google Scholar - This algorithm reads the citation indices from the Google scholar profile for a given set of authors and creates a csv file with the corresponding citation indices.
  - Attach Citation Table from Google Scholar - This algorithm reads citations in the Google scholar profile for a given author and returns the citation information in the form of a series of tables for all the queried Google Citation user ID

Data Preparation

- ***Text Files***
  - Remove ISI Duplicate Records – Removes duplicate publications form ISI records based on ISI Unique ID attribute.
  - Remove Rows with Multitudinous Fields – Removes rows having at least N entries within a given field.

  - Extract Directed Network – Creates a directed network by placing a directed edge between the values in a given column to the values of a different column.
  - Extract Bipartite Network – Creates an unweighted bipartite network by placing a directed edge between the values in a given column to the values of a different column.
  - Extract Paper Citation Network – Extracts an unweighted directed network from papers to their citations.
  - Extract Author Paper Network – Extracts an unweighted directed network from authors to their papers.

  - Extract Co-Occurrence Network – Extracts a network from a delimited table.
  - Extract Word Co-Occurrence Network – Creates a weighted network where each node is a word and edges connect words to each other. The strength of an edge represents how often two words occur in the same body of text together.

- Extract Co-Author Network – Extracts a weighted network with authors as nodes and edge weights as the number of times those authors co-wrote a paper.
- Extract Reference Co-Occurrence (Bibliographic Coupling) Network – Extracts a weighted network from a Paper Citation network, with papers as nodes and edge weights as the number of citations two papers share.

- Extract Document Co-Citation Network – Extracts a weighted network from a Paper Citation network, with papers as nodes and edge weights as the number of times two papers are cited together.

- Detect Duplicate Nodes – Cleans graph data by detecting and preparing to merge nodes that are likely to represent the same entity.
- Update Network by Merging Nodes – Creates a new network by running the algorithm with both the Merge Table from "*Detect Duplicate Nodes*" and the original network selected.

Preprocessing

- ***General***
  - Extract Top N% Records – Returns the top N% rows of a table by selecting the percentage of rows to keep and column to sort by.
  - Extract Top N Records – Returns the top N rows of a table by selecting the number of rows to keep and column to sort by.
  - Aggregate Data – Summarizes the input table by column, allowing the aggregation of values such as "Cited Reference Count," "Number of Pages," "Publication Year," "Times Cited," as well as values represented by many other delimiters.
- ***Temporal***
  - Slice Table by Time – Slices a table into groups of rows by time.
- ***Geospatial***
  - Extract ZIP Code – Extracts a ZIP code from a given address.
- ***Topical***
  - Lowercase, Tokenize, Stem, and Stopword Text – Replaces spaces and punctuation from a field with a standard delimiter of the user's choosing.
- ***Networks***
  - Extract Top Nodes – Extracts the top N nodes from a graph, based on a given attribute.
  - Extract Nodes Above or Below Value – Extracts nodes with an attribute above or below a certain value.
  - Delete Isolates – Removes nodes which are not connected to any other in the graph.
  - Extract Top Edges – Extracts the top N edges from a graph, based on a given attribute.
  - Extract Edges Above or Below Value – Extracts all edges with an attribute above or below a certain number from a graph.
  - Remove Self Loops – Removes edges whose source and target nodes are equivalent from a graph.
  - Trim by Degree -- Deletes edges at random until each node has at most N edges.
  - MST-Pathfinder Network Scaling – Prunes a network using the MST-Pathfinder algorithm.
  - Fast Pathfinder Network Scaling – Prunes a network using the Fast Pathfinder algorithm.

  - Snowball Sampling (N nodes) – Picks a random node and traverses its edges iteratively until N nodes are extracted.
  - Node Sampling – Extracts N random nodes and their intervening edges, and then deletes isolates.
  - Edge Sampling – Extracts N random edges and their target and source nodes.

  - Symmetrize-- Turns a directed network into an undirected network.
  - Dichotomize – Trims edges above, equal to, or below a certain value.
  - Multipartite Joining – Joins a multipartite graph for one node type across another node type.

- Merge 2 Networks – Merges identical networks. Emphasis is put on edge attributes merging.

Analysis

- ***Temporal***
  - Burst Detection – Determines periods of increased activity in a table with dates/timestamps.
- ***Geospatial***
  - Geocoder – Converts place names to latitudes and longitudes.
  - Congressional District Geocoder - This algorithm converts the given **9-digits U.S. ZIP codes (ZIP+4 codes)** into its congressional districts and geographical coordinates (latitude and longitude).
  - Yahoo Geocoder - This algorithm converts place names and addresses into latitudes and longitudes (requires Yahoo! API Key)
- ***Topical***
  - Burst Detection – Determines periods of increased activity in a table with dates/timestamps.
- ***Networks***
  - Network Analysis Toolkit (NAT) – Calculates basic network statistics, such as number of nodes (and isolated nodes), node attributes, number of edges, presence of self loops and parallel edges, average degree ("total," "in," and "out"), strength of component connections, and overall density.
  - ***Unweighted & Undirected***
    - Node Degree – Calculates the amount of edges adjacent to a node, and then appends that value to each node.
    - Degree Distribution – Builds a histogram of the degree values of all nodes.

    - K-Nearest Neighbor – Calculates the correlation between the degree of a node and that of its neighbors, and then appends that value to each node.
    - Watts-Strogatz Clustering Coefficient – Calculates the degree to which nodes tend to cluster together, and then appends that value to each node.
    - Watts Strogatz Clustering Coefficient over K – Correlates the clustering coefficient and the degree of the nodes of a network.

    - Diameter – Calculates the length of the longest shortest path between pairs of nodes in a network.
    - Average Shortest Path – Calculates the average length of the shortest path between pairs of nodes in a network.
    - Shortest Path Distribution – Builds a histogram of the lengths of shortest paths between pairs of nodes in a network.
    - Node Betweenness Centrality – Appends a value to each node which correlates to the number of shortest paths that node resides on. The more shortest paths between node-pairs a certain node resides on, the higher its betweenness centrality.

    - Weak Component Clustering – Extracts the N largest weakly connected components of a network.
    - Global Connected Components – Calculates the number of connected components or subgraphs with a path between each pair of nodes.

    - Extract K-Core – Extracts the kth K-Core from a graph. The kth K-Core is what remains of the graph after every node with fewer than k edges connected to it is removed from the graph recursively.
    - Annotate K-Coreness-- Appends to each node the K-Core that node belongs to.

    - HITS – Computes authority and hub score for every node.
  - ***Weighted & Undirected***

- Clustering Coefficient – Calculates the degree to which nodes tend to cluster together, and then appends that value to each node.
- Nearest Neighbor Degree - Determines the average nearest neighbor degree.
- Strength vs Degree -
- Degree & Strength - Determines the degree and strength of each node.
- Average Weight vs End-point Degree - Determines the average weight as a function of end-point-degree.
- Strength Distribution - Determines the strength distribution.
- Weight Distribution - Determines the weight distribution.
- Randomize Weights - Redistributes the weights, while keeping the topology invariant.

- Blondel Community Detection – Extracts a hierarchical community structure for a large network.

- HITS – Computes authority and hub score for every node.

- ***Unweighted & Directed***
  - Node Indegree – Appends the number of incoming edges to each node.
  - Node Outdegree – Appends the number of outgoing edges to each node.
  - Indegree Distribution – Builds a histogram of the values of the indegree of all nodes.
  - Outdegree Distribution – Builds a histogram of the values of the outdegree of all nodes.

  - K-Nearest Neighbor – Calculates the correlation between the degree of a node and that of its neighbors, and then appends that value to each node.
  - Single Node In-Out Degree Correlations – Calculates the correlations between indegree and outdegree of a node.

  - Dyad Reciprocity – The ratio of dyads with a reciprocated tie to dyads with any tie.
  - Arc Reciprocity – The ratio of reciprocal edges to total edges.
  - Adjacency Transitivity – The ratio of transitive triads to intransitive triads (triads missing one edge).

  - Weak Component Clustering – Extracts the N largest weakly connected components of a network.
  - Strong Component Clustering – Extracts the N largest strongly connected components of a network.

  - Extract K-Core – Extracts the kth K-Core from a graph. The kth K-Core is what remains of the graph after every node with fewer than k edges connected to it is removed from the graph recursively.
  - Annotate K-Coreness – Appends to each node the K-Core that node belongs to.

  - HITS – Computes authority and hub score for every node.
  - PageRank – Ranks the importance of a node by how many other important nodes point to it.
- ***Weighted & Directed***
  - HITS – Computes authority and hub score for every node.
  - PageRank – Ranks the importance of a node by how many other important nodes point to it, taking into account edge weights.

Modeling

- **Networks**

  - Random Graph – Generates a graph with a fixed number of nodes connected randomly by undirected edges.

- Watts-Strogatz Small World – Generates a graph whose majority of nodes are not *directly* connected to one another, but are still connected to one another via relatively few edges.
- Barabási-Albert Scale-Free – Generates a scale-free network by incorporating growth and preferential attachment.

- TARL (Topics, Aging and Recursive Linking) – Incorporates "aging" to generate bipartite coevolving networks of authors and papers. Can also be applied to other datasets with different aging distribution.

R

- Create an R Instance
- Run Rgui
- Send a Table to R
- Get a Table from R

Visualization

- **General**
  - GnuPlot – Used to plot two-dimensional functions and data points in many different formats. For full documentation of this open-source software, please visit http://www.gnuplot.info/documentation.html
- **Temporal**
  - Temporal Bar Graph
  - Horizontal Bar Graph
  - Horizontal Bar Graph (not included version) – Uses csv (tabular) datasets (including NSF grant data and the output of burst detections) to visualize numeric data over time, generating labeled horizontal bars that correspond to records in the original dataset.
- **Geospatial**
  - Proportional Symbol Map – Maps geospatial coordinates as circles that can be size- and color-coded in proportion to associated numeric data.. Result is a PostScript file.
  - Choropleth Map – Color-codes named regions on a geographical map in proportion to associated numeric data. Result is a PostScript file.
  - Geospatial Network Layout with Base Map – Allows for the geospatial visualization of network data, by producing a network file and corresponding blank map.
- **Networks**
  - GUESS – Interactive data analysis and visualization tool.
  - Gephi - Interactive data analysis and visualization tool at http://gephi.org.

  - Radial Tree/Graph (prefuse alpha) – A single node is placed at the center and all others are laid around it in a tree structure.
  - Radial Tree/Graph with Annotation (prefuse beta) – A single node is placed at the center and all others are laid around it in a tree structure, with labels.
  - Tree View (prefuse beta) – Visualizes directory hierarchies in a tree structure. Warning: Does not work on Macs.
  - Tree Map (prefuse beta) – Visualizes hierarchies using the Treemap algorithm. Warning: Does not work on Macs.
  - Force Directed with Annotation (prefuse beta) – Sorts randomly placed nodes into a more aesthetically pleasing visual layout.
  - Fruchterman-Reingold with Annotation (prefuse beta) – Visualization which lays out nodes based on some force between them.

  - DrL (VxOrd) – A force-directed graph layout toolbox focused on real-world large-scale graphs.
  - Specified (prefuse beta) – Visualization tool for use with graphs having pre-specified node coordinates.

  - Circular Hierarchy – Generates a circular visualization of the output produced by a multi-level

aggregation method such as Blondel Community Detection. Result is a Postscript file.
- Bipartite Network Graph - Generates a bipartite network visualization of the output produced by Extract Bipartite Network algorithm.

3.2 Additional Plugins

The Sci^2 Tool is an 'empty shell' filled with plugins. Some plugins run on the core architecture, OSGi and CIShell (see Section 7 Extending the Sci2 Tool). Others convert loaded data into in-memory objects, formatted for different algorithms to read it. The algorithm plugins themselves can be divided into different menus, in this case data preparation, preprocessing, analysis, modeling and visualization. Users are not limited to using pre-packaged plugins; instead, they can create, download, share, and import their own.

There are two ways plugins can be added to the menu. First, if an algorithm is listed in the *default_menu.xml* file, which is found in the *configuration/* directory of a CIShell tool, it is added in the place specified by the *default_menu.xml* file. Second, if an algorithm is not listed in this file, but has specified a *menu_path* property in the *algorithm.properties* file, the menu manager CIShell service will add it to the appropriate menu, as described in the by the *menu_path* property's value. For example, Analysis/additions will place an algorithm on the bottom of the Analysis menu.

It is possible to extend Sci^2 Tool, by adding plugins related to database functionalities and Cytoscape tool (open source software platform for visualizing networks).

By adding the plugins related to database functionalities, the Sci^2 Tool can support the creation of databases from ISI and NSF files. Database loading improves the speed and functionality of data preparation and preprocessing. While the initial loading can take quite some time for larger datasets (see sections 3.4 Memory Allocation and 3.5 Memory Limits) it results in vastly faster and more powerful data processing and extraction.

Cytoscape (http://www.cytoscape.org) is an open source software platform for visualizing networks and integrating these with any type of attribute data. Although Cytoscape was originally designed for biological research, now it is a general platform for network analysis and visualization. A variety of layout algorithms are available, including cyclic, tree, force-directed, edge-weight, and yFiles Organic layouts. Cytoscape tool can also be added to the Sci^2 Tool as a plugin, permitting that the networks generated at Sci^2 be visualized in Cystoscape.

To add Database, Cytoscape, Congressional District Geocoder, and New ISI File Format plugins to Sci^2

1. Download these files that contain additional plugins to add Database, Balloon Graph, Cytoscape, Congressional District Geocoder, and New ISI File Format plugins to Sci^2:
   - DatabasePlugins.zip (unzip to obtain the jar files that will be inside a folder named DatabasePlugins)
   - BalloonGraph.zip (unzip to obtain the jar files that will be inside a folder named BalloonGraph)
   - WOS-plugins.zip provides the support for the newest version of the **ISI Web of Science file format**. See ISI (*.isi) for more information
   - org.textrend.visualization.cytoscape_0.0.3.jar
   - edu.iu.sci2.preprocessing.zip2district_0.0.1.jar
2. Copy these jar files to the directory */plugins* of Sci^{2}directory
3. Download the file default_menu.xml and copy it to the directory *configuration/* under Sci^2 directory, replacing the older version of this file. This file specifies where each plugin added will be located at the menu.
4. The next time the Sci^2 Tool is loaded after this modifications, this additional plugins will appear as new menu items.

Algorithms available at extended version

- ***Database***
  - ***ISI***

- Merge Identical ISI People – Merges all people in an ISI database who have the same name.
- Suggest ISI People Merges – Generates a pre-annotated merging table based on a user-selected threshold and string similarity metric.
- Merge Document Sources – Merge document sources of documents and references based upon known name variants and abbreviations.
- Create Document Source Merging Table – Essentially the same as Create Merging Table, but including two analyses of your dataset that may help in determining good candidates for merging.
- Match References to Papers – Matches references to papers if they have the same first author, source (journal), start page, volume, and year.

- Extract Authors – Outputs a table containing one row per author in the database.
- Extract Documents – Outputs a table containing one row per document in the database.
- Extract Keywords – Outputs a table containing one row per keyword in the database.
- Extract Document Sources – Outputs a table containing one row per document source in the database.

- Extract Authors by Year – Outputs a table containing the amount of publications per author per year.
- Extract References by Year – Outputs a table containing the amount of references to a publication per year.
- Extract Original Author Keywords by Year – Outputs a table containing one row per original author keyword per year.
- Extract New ISI Keywords by Year – Outputs a table containing one row per new ISI keyword per year.

- Extract Authors by Year for Burst Detection – Used for author burst detection.
- Extract Documents by Year for Burst Detection – Used for word occurrence based burst detection.
- Extract Original Author Keywords by Year for Burst Detection – Used for author keyword burst detection.
- Extract New ISI Keywords by Year for Burst Detection – Used for new ISI keyword burst detection.
- Extract References by Year for Burst Detection – Used to detecting bursting references to publications.

- Extract Longitudinal Summary – Outputs a table with the total number of documents published, references published, references made, distinct authors, distinct sources, distinct author keywords, distinct ISI keywords, and distinct other keywords by year.

- Extract Co-Author Network – Extracts a weighted, undirected network with authors as nodes and edges between authors who co-wrote papers. The extraction appends to nodes the number of authored documents, ISI's times-cited count, the publication of the earliest document, and the publication year of the most recent document. The extraction appends to edges weights for the number of co-written papers, and the publication years of the earliest and most recent collaboration.

- Extract Author Citation Network – Extracts a weighted, directed network with authors as nodes and edges from a citing author to a cited author. Nodes include all data from the PERSON tabl e and the number of documents authored in the current dataset.
- Extract Document Citation Network (Core Only) – Extracts an unweighted, directed network with documents as nodes and edges from a citing paper to a cited paper. Only those documents with full entries in the dataset are included in the network. Nodes include all data from the DOC

UMENT table.

- Extract Document Citation Network (Core and References) – Extracts an unweighted, directed network with documents as nodes and edges from a citing paper to a cited paper. Nodes include all data from the DOCUMENT table.
- Extract Document Source Citation Network (Core Only) – Extracts a weighted, directed network with sources (journals) as nodes and edges from a citing source to a cited source. Citations are via documents within sources, and only those sources represented by documents within the dataset are included in the network. Nodes include all data from the SOURCE table, and edges are weighted by the number of citations between sources.
- Extract Document Source Citation Network (Core and References) – -- Extracts a weighted, directed network with sources (journals) as nodes and edges from a citing source to a cited source. Citations are via documents within sources. Nodes include all data from the SOURCE table, and edges are weighted by the number of citations between sources.

- Extract Document Co-Citation Network (Core Only) – Extracts a weighted, undirected network with documents as nodes and edges between documents which have been cited together. Only those documents with entries in the dataset are included in the network. Edge weight is determined by the number of times two articles are cited together, and edges are appended with the publication years of their earliest and most recent co-citations.
- Extract Document Co-Citation Network (Core and References) – Extracts a weighted, undirected network with documents as nodes and edges between documents which have been cited together. Edge weight is determined by the number of times two articles are cited together, and edges are appended with the publication years of their earliest and most recent co-citations.
- Extract Document Source Co-Citation Network (Core Only) – Extracts a weighted, undirected network with journals as nodes and edges between journals which have been cited together by a common document. Only those journals containing documents with entries in the dataset are included in the network. Edge weight is determined by the number of times two journals are cited together, and edges are appended with the publication years of their earliest and most recent co-citations.
- Extract Document Source Co-Citation Network (Core and References) – Extracts a weighted, undirected network with journals as nodes and edges between journals which have been cited together by a common document. Edge weight is determined by the number of times two journals are cited together, and edges are appended with the publication years of their earliest and most recent co-citations.
- Extract Author Co-Citation Network – Extracts a weighted, undirected network with authors as nodes and edges between authors who have been cited together by a common document. Edge weight is determined by the number of times two authors are cited together, and edges are appended with the publication years of their earliest and most recent co-citations.

- Extract Author Bibliographic Coupling Network – Extracts a weighted, undirected network with authors as nodes and edges between authors who cite a common reference. Edge weight is determined by the number of common references between authors.
- Extract Document Bibliographic Coupling Network – Extracts a weighted, undirected network with documents as nodes and edges between documents which cite a common reference. Edge weight is determined by the number of common references between documents.
- Extract Document Source Bibliographic Coupling Network – Extracts a weighted, undirected network with journals as nodes and edges between journals whose documents cite a common reference. Edge weight is determined by the number of common references between documents.

- ***NSF***
  - Merge Identical NSF People – Merges all people in an NSF database who have the same name.

- Extract Investigators – Extracts a table containing one row per investigator from an NSF database.
- Extract Awards – Extracts a table containing one row per award from an NSF database.
- Extract Organizations – Extracts a table containing one row per organization from an NSF database.

- Extract Co-PI Network – Extracts a weighted, undirected network with principle investigators as nodes and edges between them if they co-investigated an award in the database. Nodes are appended with the number of awards investigated, total amount across each investigated award, start date of earliest award, and expiration date of most recent award. Edges are appended with the number of awards co-investigated by the two investigators and the joint award total between the investigators.

- ***General***
  - Create Merging Table – Given a database, this generates a spreadsheet that can be annotated to indicate how entities should be merged. The user selects the type of entity.
  - Merge Entities – Given a database and a merging table, this generates a new database where the merges in the merging table have been performed.
  - Custom Table Query – This algorithm executes the given SQL query on the selected database, returning a table with the results.
  - Custom Graph Query – This algorithm executes the two given SQL queries on the selected database, returning a graph with the results.
  - Extract Raw Tables from Database – Lets you explore the raw contents of a database, extracting every table in the database into the Data Manager.
- ***Generic-CSV***
  - Extract Co-Occurrence Network – Allows the user to create a co-occurrence network based on columns in a table.
  - Extract Bipartite Network – Extract a bipartite network from tabular data.
  - Extract Co-Entity Occurrence Network
  - Extract Table

- ***Analysis***
  - ***Geospatial***
    - Congressional District Geocoder - This algorithm converts the given **9-digits U.S. ZIP codes (ZIP+4 codes)** into its congressional districts and geographical coordinates (latitude and longitude).

- ***Visualization***
  - ***Networks***
    - Cytoscape - Open source software platform for visualizing networks and integrating these with any type of attribute data.
    - Bipartite Network Graph

Menus at Extended Version

As stated before, this additional plugins will appear as new menu items in the Sci^2 Tool.

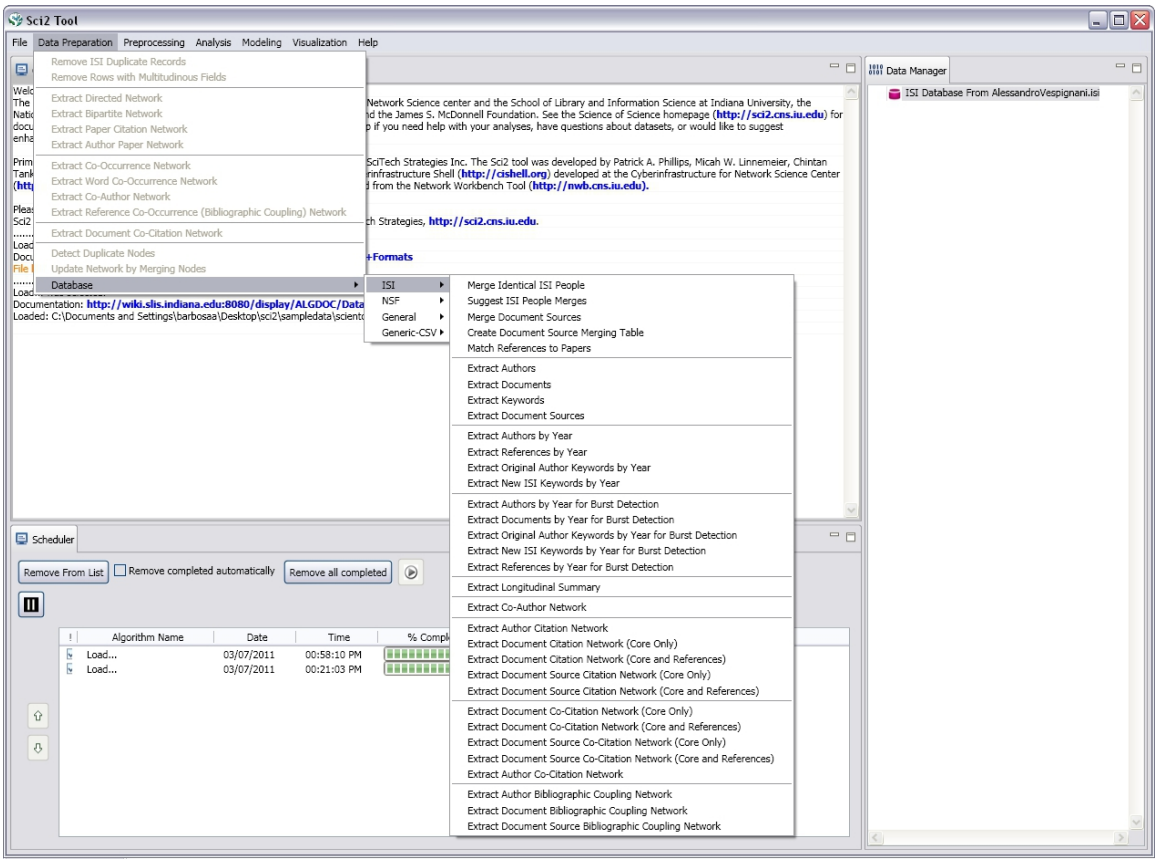


Figure 2.1: Menus in the extended version of Sci^2.

How Algorithms are Added to the Menu

There are two ways algorithms are added to the menu. First, if an algorithm is listed in the *default_menu.xml* file, which is found in the *configuration/* directory of the Sci^2 Tool, it is added in the place specified by the default_menu.xml file. Secondly, if an algorithm is not listed in this file, but has specified a *menu_path* property in the *algorithm.properties* file, the menu manager CIShell service will add it to the appropriate menu, as described in the by the *menu_path* property's value (for example, Analysis/additions will place an algorithm on the bottom of the Analysis menu).

Bipartite Network Graph

Description

The Bipartite Network Graph algorithm plots a bipartite network, that is, a network with exactly two distinct types of nodes. For instance, in the network below, one type of nodes represents a person, and the other type represents an action. The Extract Bipartite Network algorithm is an easy way to generate networks that are graphable by this plugin.

Bipartite Network Graph Example (web format) (pdf)
Bipartite Network Graph Example (print format) (pdf)

The nodes and edges can each be independently weighted. The algorithm chooses which column to put a node in based on a node attribute called "bipartitetype". For details on the input format, see Usage Hints below.

Links

- Source code in SVN

    - http://in.cns.iu.edu/svn/nwb/branches/tycho/sci2/plugins/normal_plugins/edu.iu.sci2.visualization.bipartitenet/ AT REVISION 5989
- Also requires:
    - http://in.cns.iu.edu/svn/nwb/branches/tycho/sci2/plugins/libs/javaGeom/
    - http://in.cns.iu.edu/svn/nwb/branches/geo-nete/edu.iu.sci2.visualization.geomaps and its dependencies (see Geospatial Visualization)
        - All these weird branch-y things will be fixed in the next week or so - Thomas, 3/21/2012
- Test code:
    - http://in.cns.iu.edu/svn/nwb/branches/tycho/sci2/plugins/unittests/edu.iu.sci2.visualization.bipartitenet.tests/

Pros & Cons

It's easy to see the network structure in bipartite networks, because of the obvious separation between the two different node types. However, medium to large networks (50+ nodes) may not look very good.

Applications

Implementation Details

The graph is rendered to a PNG file and a PostScript (PS) file.

Usage Hints

This is the network file (NWB (.nwb) format) that was used to generate the example graphs above:

```
*Nodes
id*int label*string  totaldesirability*int bipartitetype*string
1 "Applicant's Proposal" 1 "Who"
2 "Kiss My Red Ruby Lips" 5 "What"
3 "Shoe My Pretty Little Feet" 10 "What"
4 "Glove My Hand" 9 "What"
5 "Be My Man" 4 "What"
6 "Papa" 9 "Who"
7 "Mama" 8 "Who"
8 "Sister" 4 "Who"
9 "No Man" 3 "Who"
*DirectedEdges
source*int target*int linkdesirability*int
1 4  1
1 2  1
1 3  1
1 5  1
6 3  9
7 4  8
8 2  4
9 5  3
```

The nodes are required to have an attribute called "bipartitetype". Note that in the above network, the two values of "bipartitetype" are "Who and "What". These are also the titles of the columns in the graph. This attribute is used to determine which column the nodes should go in. You can generate your own network with this attribute, or you can use the Extract Bipartite Network algorithm.

The weights ("totaldesirability" and "linkdesirability") were generated using an aggregation function file with the

Extract Bipartite Network algorithm. For more on this, see Extract Bipartite Network and the Extract Co-Occurrence Network page (for info on aggregation functions).

To generate the graph file above, I imported no-man.csv into Sci2, then ran the Extract Bipartite Network algorithm. I chose "Who" and "What" as the First and Second Column parameters, and used aggfunc-man.txt as the Aggregate Function File. I saved the resulting file as a NWB file.

Parameters:

- **Left side node type**: which of the two values of "bipartitetype" should put a node in the left-hand column? The other type of nodes will go on the right.
- **Node size**: which node attribute should be used to scale the size of the nodes? If none is selected, the nodes will all be the same size.
- **Edge weight**: which edge attribute should be used to set the darkness of the edges between nodes? If none is selected, the edges will all be the same color.
- **Title for left column**: the title to print above the left column of nodes. If left blank, the value of "bipartitetype" for this column will be used.
- **Title for right column**: just like the left column.

Acknowledgements

This algorithm makes use of the javaGeom library.

References

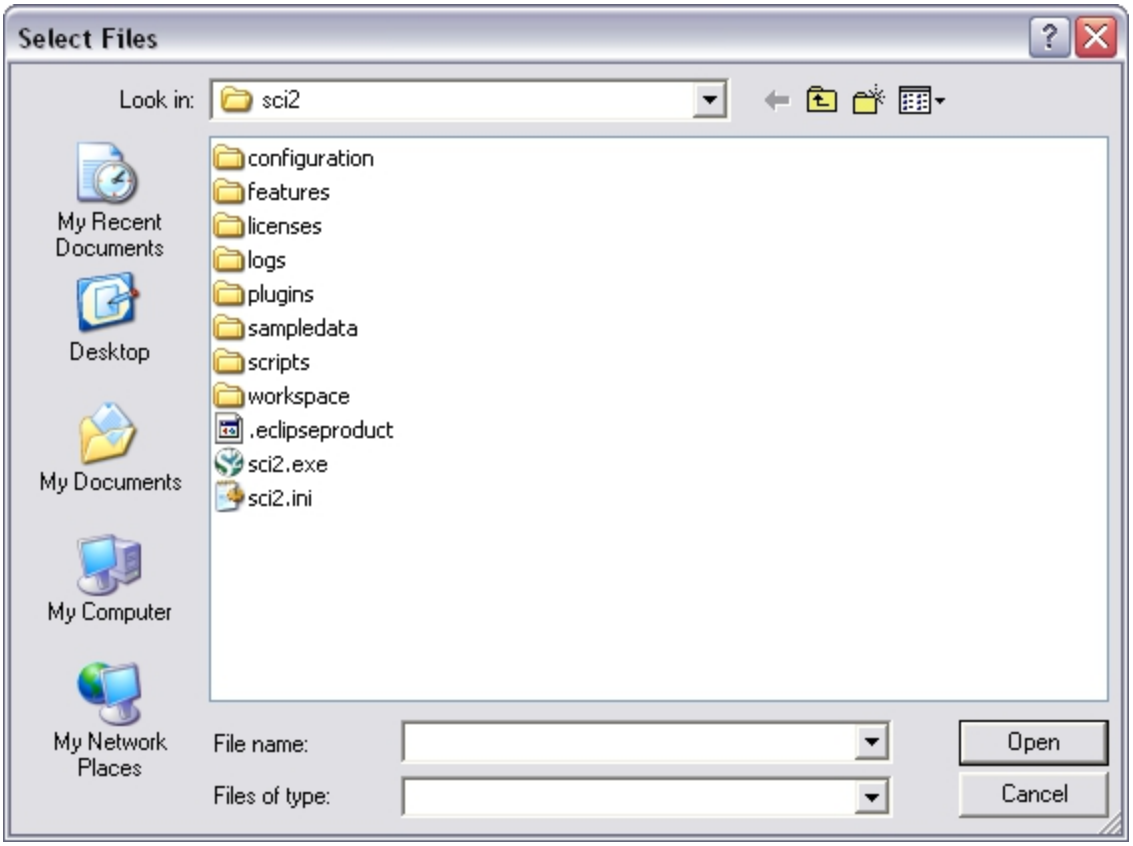
See Also

 3.2 Additional Plugins
 Shared Binaries
 Thomas's Log

3.3 Load, View, and Save Data

Standard

In the Sci^2 Tool, use '*File > Load ...*' to load one of the sample datasets provided in '***yoursci2directory****/sampledata*', or load any dataset of your own choosing.



It is possible to select more than one dataset at a time. Hold the *Ctrl* key and click several files to make multiple selections.

The '*File > Load ...*' command can be used not only to load datasets, but also to load any file that is one of the supported formats by Sci^2 . Once a dataset has been loaded it will show up in the data manager in the Sci^2 Tool.

Any file listed in the 'Data Manager' can be saved, viewed, renamed, or discarded by right clicking it and selecting the appropriate menu options. If '*File > View With ...*' was selected, the user can select among different application viewers. Choosing "Microsoft Office Excel" for a tabular type file will open MS Excel with the table loaded.

The Sci^2 Tool can save a network using '*File > Save...*' which brings up the "Save" window. Note that some data conversions are lossy, i.e., not all data is preserved. For example, when converting a network to a matrix format, attributes are lost because of the limitations of the format.

Drag-and-Drop

Sometimes it's simpler to drag-and-drop the desired file to be loaded into Sci^2 Tool than selecting it the standard way.

Press and hold down the mouse pointer to "grab" the file or group of files you want to import to Sci^2 Tool. Drag this file to the Data Manager window and drop it there.

Extended Version

If the tool was extended using the instructions contained at 3.2 Additional Plugins, the process to load datasets will be a little different. Using the extended version, it is also possible to load the datasets in the database format.

Choosing the "Load" option from the "File" dropdown menu, provides the following dialog box after a dataset file is

selected:

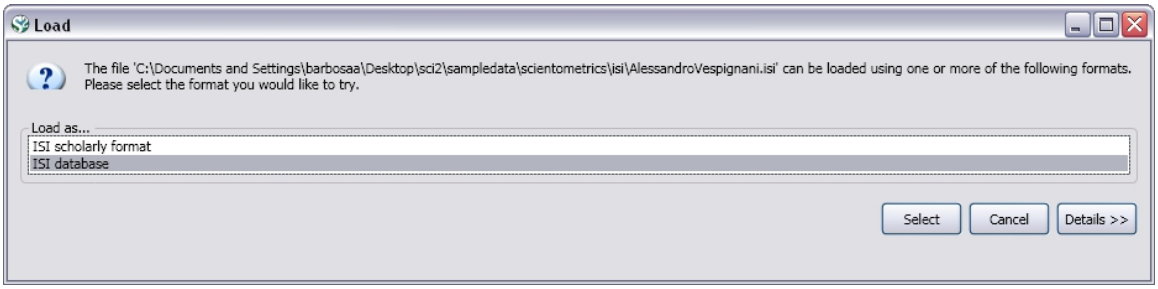


Figure 3.2: The 'Load' pop-up window

You can choose to load the file in either a text format (ex: ISI flat format or NSF csv format) or as a database. The format selected should match the format you want the tool to use to manipulate that dataset.

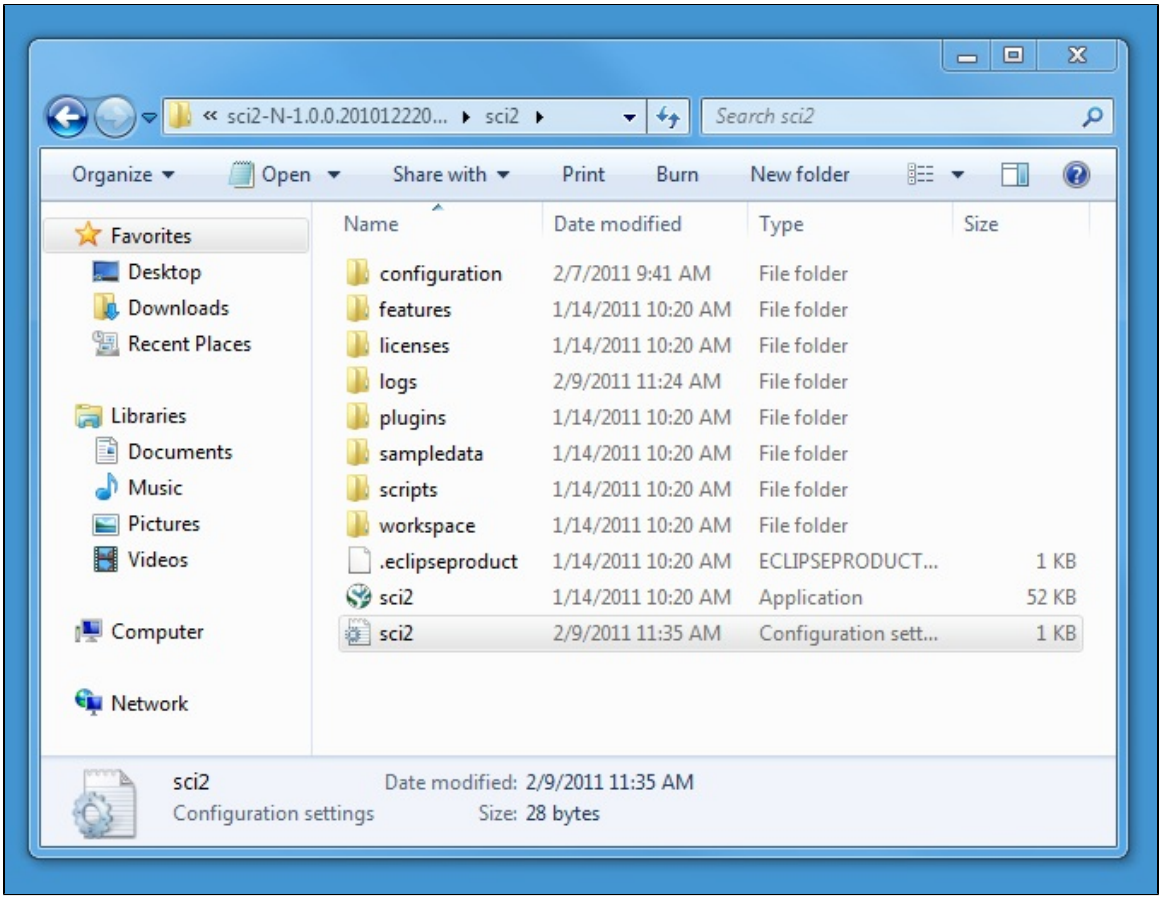
Datasets loaded as a database format, will appear with the icon in the Data Manager.

3.4 Memory Allocation

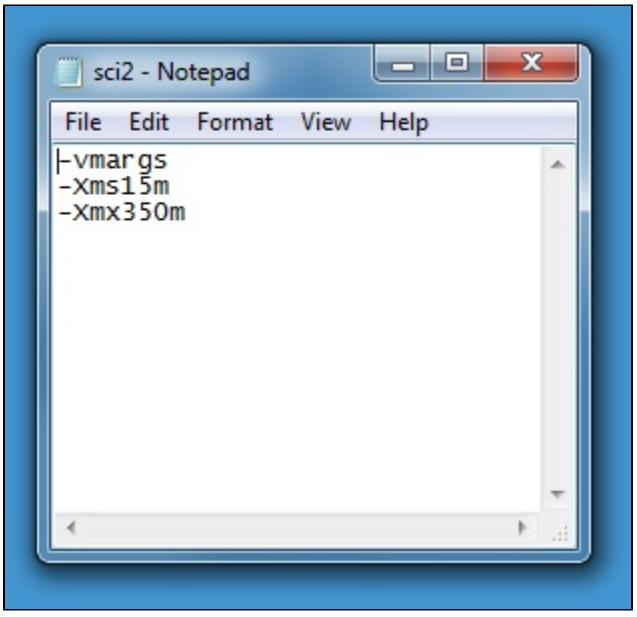
Due to the constraints of the Java virtual machine, the amount of memory available to a Java application must be determined before the application starts. The current default allotment of 350 Megabytes is a balance between providing enough memory for most uses of the tool, while not causing the Sci^2 Tool to crash on machines with too little memory. For most analyses this amount should be enough, but for larger scale operations this amount should be increased to make full use of your systems available memory

3.4.1 Windows and Linux

Open the file "scipolicy.ini" in your Sci^2 directory, using a simple text editor such as Notepad:



The file should contain the following three lines (if not, add them):



The first number (15 here) represents how much memory is allocated to Sci^2 when it first starts up. This number isn't particularly important, but should not be set to anything below 10m.
The second and more important number (350 here) represents the maximum amount of memory that can be allocated to Sci^2. This can be up to roughly 3/4ths the total available memory on your machine, but should not be set any higher, or Sci^2 will fail to start.
Make sure that the formatting is exactly as displayed above, as the .ini file can be finicky about extra spaces in the

file, or multiple arguments on a single line.
After changing these two numbers, save the sci2.ini file, and your new memory settings should be used the next time Sci^2 starts.

3.4.2 Mac

Configure the nwb.ini file inside the Sci^2 Tool application bundle. Open the Sci^2 Tool application folder and control + click on the sci2 icon. Select *'Show Package Contents'* from the menu.

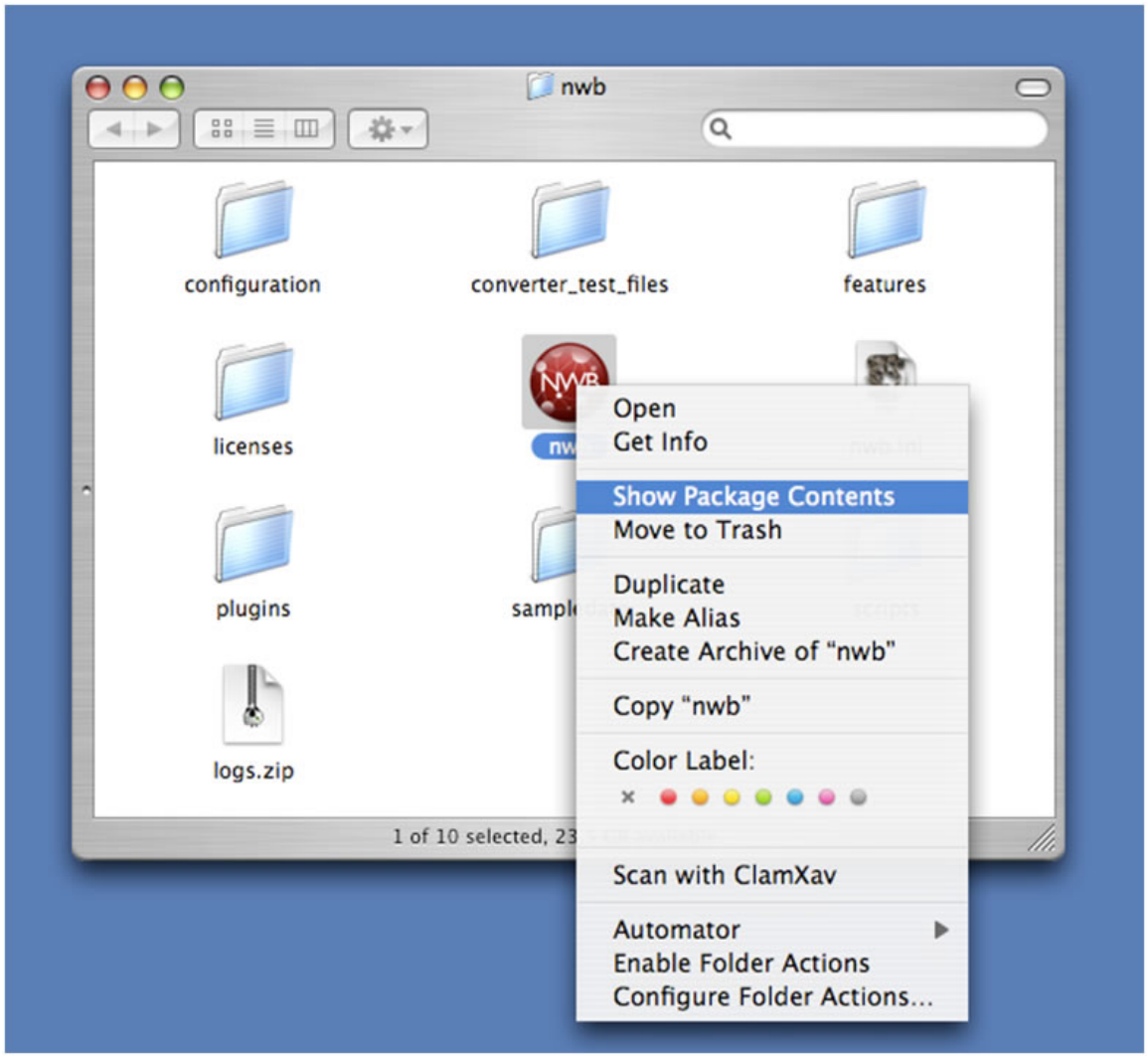


Figure 3.3: Find sci2 icon in the Sci^2 directory

This should bring up another window with a folder labeled '*Contents*.' Open the *'Contents'* folder and then open the *' MacOS'* folder found inside. Open the *'sci2.ini'* file in TextEdit or other text editor that will leave the contents as plain text.

Figure 3.3: Original sci2.ini file

To enable the Sci^2 Tool to access more memory, increase the number following *'-Xmx'* in the '*sci2.ini*' file. If the value is increased too much, the Sci^2 Tool will not function.

Figure 3.4: Updated sci2.ini file

3.5 Memory Limits

The Sci^2 Tool's database functionality greatly improves the volume of data which can be loaded into and analyzed by the tool. Whereas most scientometric tools available as of March 2010 require powerful computing resources to perform large scale analyses each time a network needs to be extracted, the pre-loaded database runs network extraction algorithms quickly and allows the users to format their own custom queries. This functionality has some front-heavy memory requirements in order to initially load the data into the database, the upper limits of which can be seen in Tables 3.1 and 3.2.

Entries	Load	Merge People	Merge Journals	Match	Extract Authors	Extract Documents	Extract Co-Authors	Extract Document Citation (with outer)	Extract Author Citation	Extract Document Co-Citation	Extract Document Co-Citation (core)	Author Co-Citation

50	5	1	1	1	1	1	1	1	1	10	4	2
500	10	4	4	1	1	1	1	1	1	20	4	5
5000	100	48	180	3	1	1	3	18	2	3300	20	35
10000	210	120	480	5	2	1	10	30	7		60	100
20000	400	300	1200	15	5	1	22	70	20		164	250

Table 3.1: The number of seconds to perform each action on a dataset of the given size, on a computer with 12GB of memory

Entries	**Load**	**Merge People**	**Merge Journals**	**Match**	**Extract Authors**	**Extract Documents**	**Extract Co-Authors**	**Extract Document Citation (with outer)**	**Extract Author Citation**	**Extract Document Co-Citation**	**Extract Document Co-Citation (core)**	**Author Co-Citation**
50	50	55	9	4	3	1	1	1	1	1	40	13
500	500	420	52	13	3	1	1	1	2	1	75	13
5000	5000	3600	1080	480	25	1	1	12	90	2		180

Table 3.2: The number of seconds to perform each action on a dataset of the given size, on a computer with 1.2GB of memory

3.6 Property Files

The properties files, sometimes referred to as aggregate function files, facilitate analysis by allowing visualization according to certain attributes of nodes and edges. These files can be used where aggregation of data is to be performed based on certain unique values. This ultimately enhances the power of visualization. The properties files are located in_'Sci2/sampledata/scientometrics/properties'_

All properties files follow the same pattern:
{**node|edge**}.*new_attribute* = *table_column_name*.[{**target|source**}].*function*

table_column_name is the name of the attribute we are going to operate on to create a new value for the final node, this can be the name of any of the node attributes in the network

function determines how we aggregate, or combine, the values for the new node

Aggregate functions:

- **arithmeticmean -** finds the average of an independent node attribute
- **geometricmean** - finds the average of a dependent node attribute
- **count -** counts the instances of appearance of a node attribute
- **sum** - the sum of each node's attribute values. Example use: When you have two author nodes who are really the same author and you want to combine the number of citations they have accumulated under both names.
- **max** - the maximum value of each node's attribute values. Example use: When you have two author nodes who are really the same author, which have two differing author ages, you might want to assume that the younger age was based on an old record, and keep the older age
- **min** - the maximum value of each node's attribute values.
- **mode -** reports the most common value for an attribute

Below is an example of a property file:

```
node.numberOfWorks = numberOfWorks.sum
node.timesCited = timesCited.sum
edge.numberOfCoAuthoredWorks = numberOfCoAuthoredWorks.sum
edge.weight = weight.ignore
```

The idea here is that for each set of nodes that are each others' duplicate, we combine each of the old attributes values for those nodes using some mathematical function, in order to create the new attribute values of the final merged node. If an attribute is not mentioned, the algorithm will pick one of attribute values from the duplicate nodes to use in the resulting node. This is acceptable if the two values will always be the same, but not if they could differ.

Below is a list of some of the various property files available for Sci2, grouped by the type of files they can used to operate on. Along with a brief description of each property file there are instructions on how to use the file. Note: this list is not complete. Be sure to check back as more property files are documented.

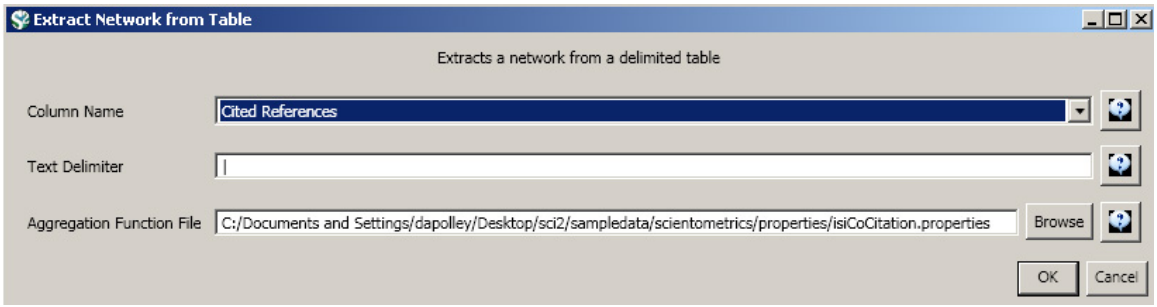
ISI Files

isiCoCitation.properties

This property file can be used with a co-occurrence extraction of cited references. The isiCoCitation.properties file will add the properties: "times cited", and "times appearing". It is useful for visualizing citation networks.

Select any ISI file to load in Sci2. Load the file in ISI flat format.

Choose *'Data Preparation > Extract Co-Occurrence Network'* and apply the following parameters, making sure to select the aggregation function file indicated below:



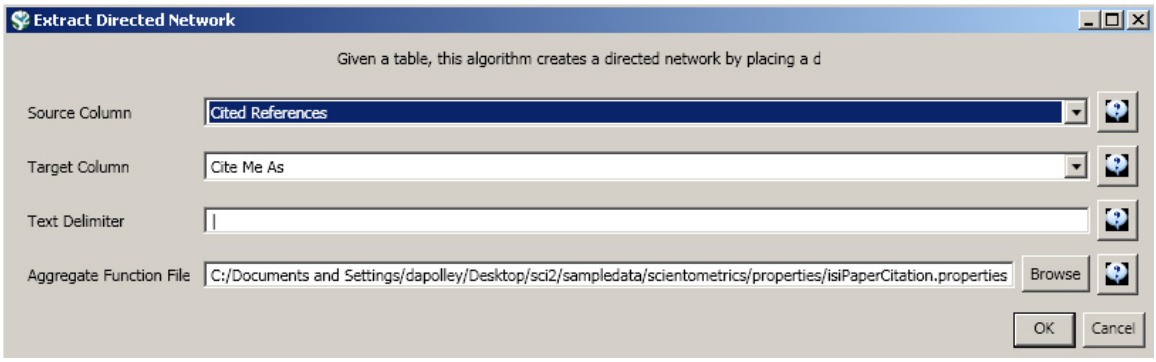
When the resulting network is visualized in GUESS two properties will be appended to the nodes and accessible

from the Graph Modifier pane. The isiCoCitation.properties file it will add "timesappearing" and "timescited" properties that will allow users to edit nodes according to these properties, allowing for more advanced visualizations.

isiPaperCitation.properties

This property file can be used in ISI paper citation data extractions. The file will add the properties: "global citation count", "in original data set", and "location citation count".

Load ISI file in Sci2. To use this file, users will want to extract a directed network, *'Data Preparation* > *Extract Directed Network',* and use the following parameters:



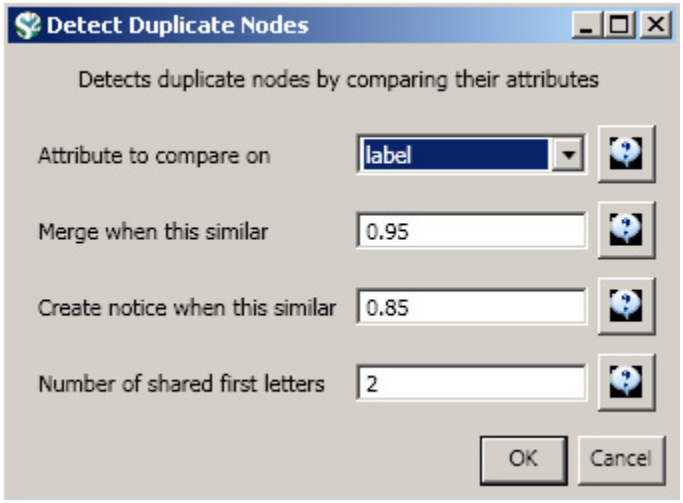
The resulting network will have the properties "globalcitationcount", "inoriginaldataset", and "localcitationcount" to the nodes in the network, allowing the nodes to be edited by these properties.

mergeIsiAuthors.properties

This property file allows for the merging of duplicate nodes in an author table and updates the corresponding network.

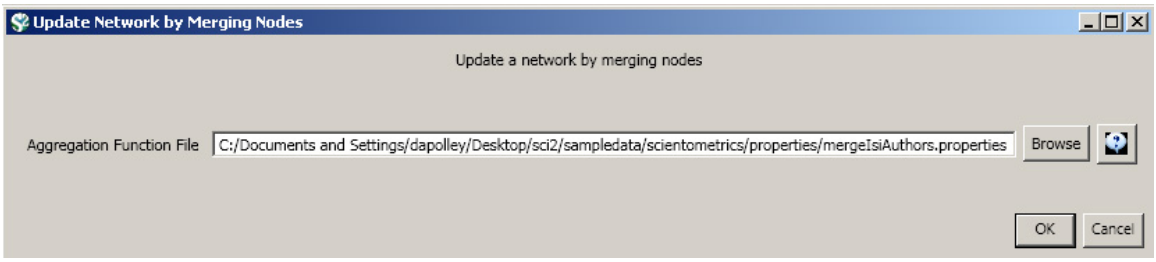
Load any ISI file in Sci2. This property file will be used when you extract a co-author network. Run *'Data Preparation* > *Extract Co-Author Network'*.

Next, you will want to detect any duplicate nodes in the network. Run, *'Data Preparation* > *Detect Duplicate Nodes'* u sing the following parameters:



Running this algorithm will result in a merge table and two text files. Specifically, the first log file provides information regarding which nodes will be merged, while the second log file lists nodes which are similar but will not be merged.

The automatically generated merge table can be further modified as needed.
To merge identified duplicate nodes, select both the "Network with directed edges from Cited References to Cite Me As" network and "Merge Table: based on label" by holding down the 'Ctrl' key. Run *'Data Preparation* > *Update Network by Merging Nodes'*. This will produce an updated network as well as a report describing which nodes were merged. To complete this workflow, an aggregation function file must also be selected from the pop-up window:



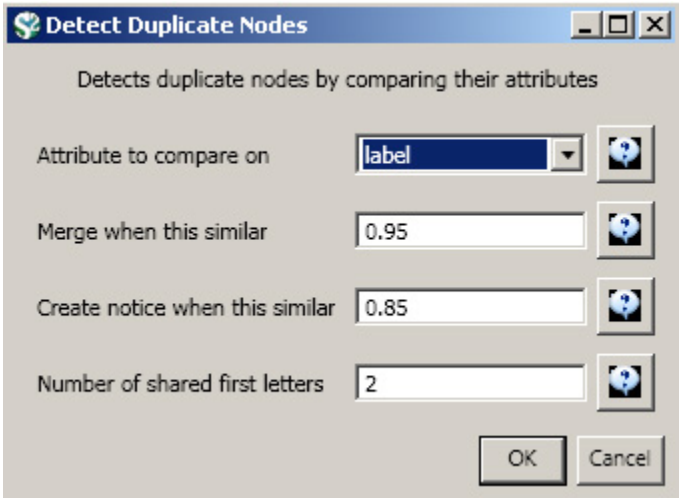
The resulting ISI co-author network will have duplicate author nodes merged.

mergeIsiPaperCitation.properties

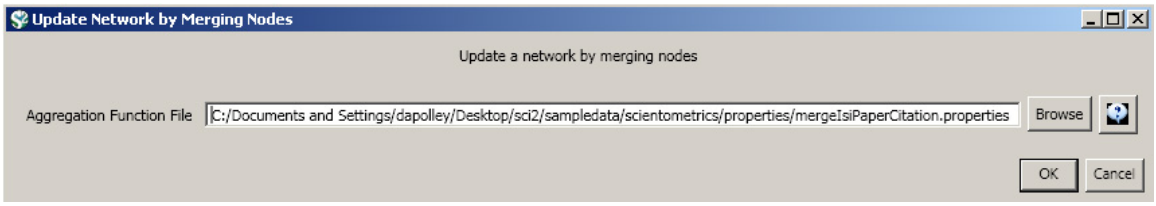
The property file allows for the merging of duplicate nodes in a paper citation table and updates the corresponding network.

Load any ISI file in Sci2. To use this file, users will want to extract a directed network, *'Data Preparation* > *Extract Directed Network'*, and use the following parameters:

Next, you will want to detect any duplicate nodes in the data set. Run, *'Data Preparation* > *Detect Duplicate Nodes'* u sing the following parameters:



Running this algorithm will result in a merge table and two text files. Specifically, the first log file provides information regarding which nodes will be merged, while the second log file lists nodes which are similar but will not be merged.
The automatically generated merge table can be further modified as needed.
To merge identified duplicate nodes, select both the "Network with directed edges from Cited References to Cite Me As" network and "Merge Table: based on label" by holding down the 'Ctrl' key. Run *'Data Preparation* > *Update Network by Merging Nodes'*. This will produce an updated network as well as a report describing which nodes were merged. To complete this workflow, an aggregation function file must also be selected from the pop-up window:



The resulting ISI paper citation network will have the properties "globalcitationcount", "inoriginaldataset", and "localcitationcount" to the nodes in the network and the duplicate nodes will have been merged.

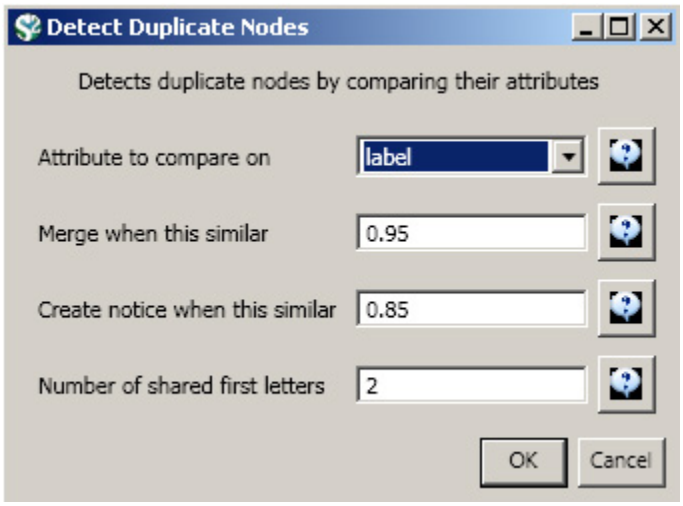
Endnote Files

mergeEndnoteAuthors.properties

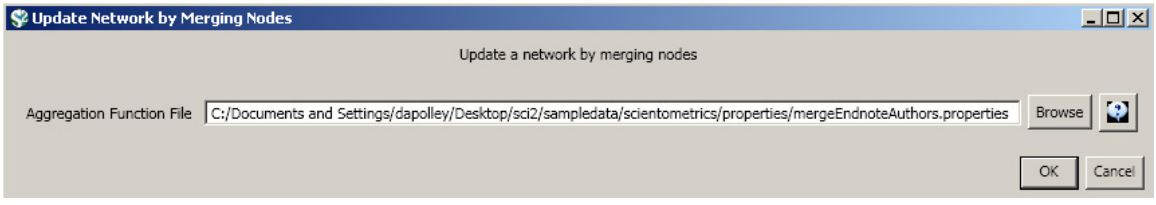
This property file allows for the merging of duplicate nodes in the author table and updates the corresponding network

Load any EndNote file in Sci2. Extract a co-author network by running *'Data Preparation >* *Extract Co-Author Network'*.

Run, *'Data Preparation >* *Detect Duplicate Nodes'* with the following parameters:



To merge identified duplicate nodes, select both the "Extracted Co-Authorship Network" network and "Merge Table: based on label" by holding down the 'Ctrl' key. Run *'Data Preparation >* *Update Network by Merging Nodes'*. This will produce an updated network as well as a report describing which nodes were merged. To complete this workflow, an aggregation function file must also be selected from the pop-up window:



The resulting network will have duplicate author nodes merged.

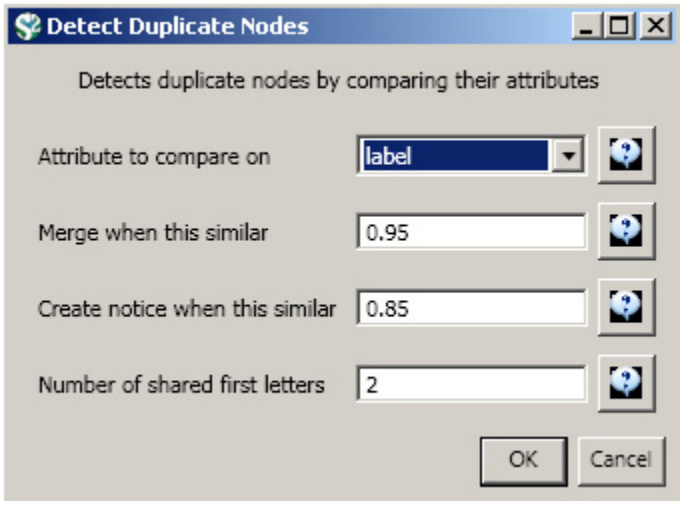
Scopus Files

mergeScopusAuthors.properties

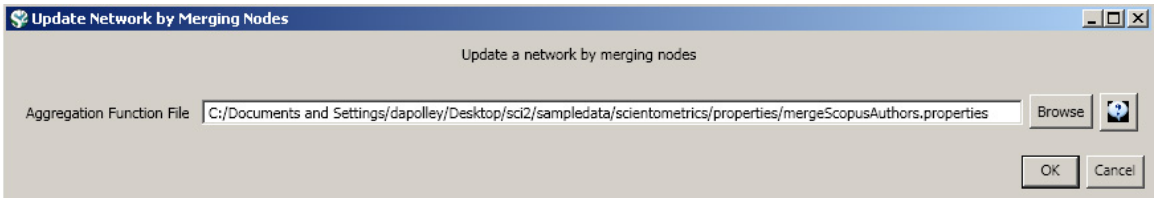
This property file allows for the merging of authors and updates the corresponding network.

Load any Scopus in Sci2. Extract a co-author network by following *'Data Preparation* > *Extract Co-Author Network'*

To detect any duplicate authors, select _'Data Preparation >_ *Detect Duplicate Nodes'* with the following parameters:



To merge identified duplicate nodes, select both the "Extracted Co-Authorship Network" and "Merge Table: based on label" by holding down the 'Ctrl' key. Run *'Data Preparation* > *Update Network by Merging Nodes'*. This will produce an updated network as well as a report describing which nodes were merged. To complete this workflow, an aggregation function file must also be selected from the pop-up window:



The resulting network will have duplicate author nodes merged.

4 Workflow Design

4.1 Overview

A typical science study starts with a N**eeds Analysis** of a selected stakeholder group that informs the subsequent workflow design, involving **Data Acquisition and Processing**; **Data Analysis, Modeling, and Layout**; and **Data**

Communication--Visualization Layers. All datasets, algorithms, and parameter values used in a study have to be documented in detail in support of replication and interpretation. The resulting **Validation and Interpretation** should then proceed in collaboration with domain experts and stakeholders. Insights gained might generate additional insight needs or inspire changes to the workflow. The process is highly incremental, often demanding many cycles of revision and refinement to ensure the best datasets are used, optimal algorithm parameters applied, and clearest insight achieved.

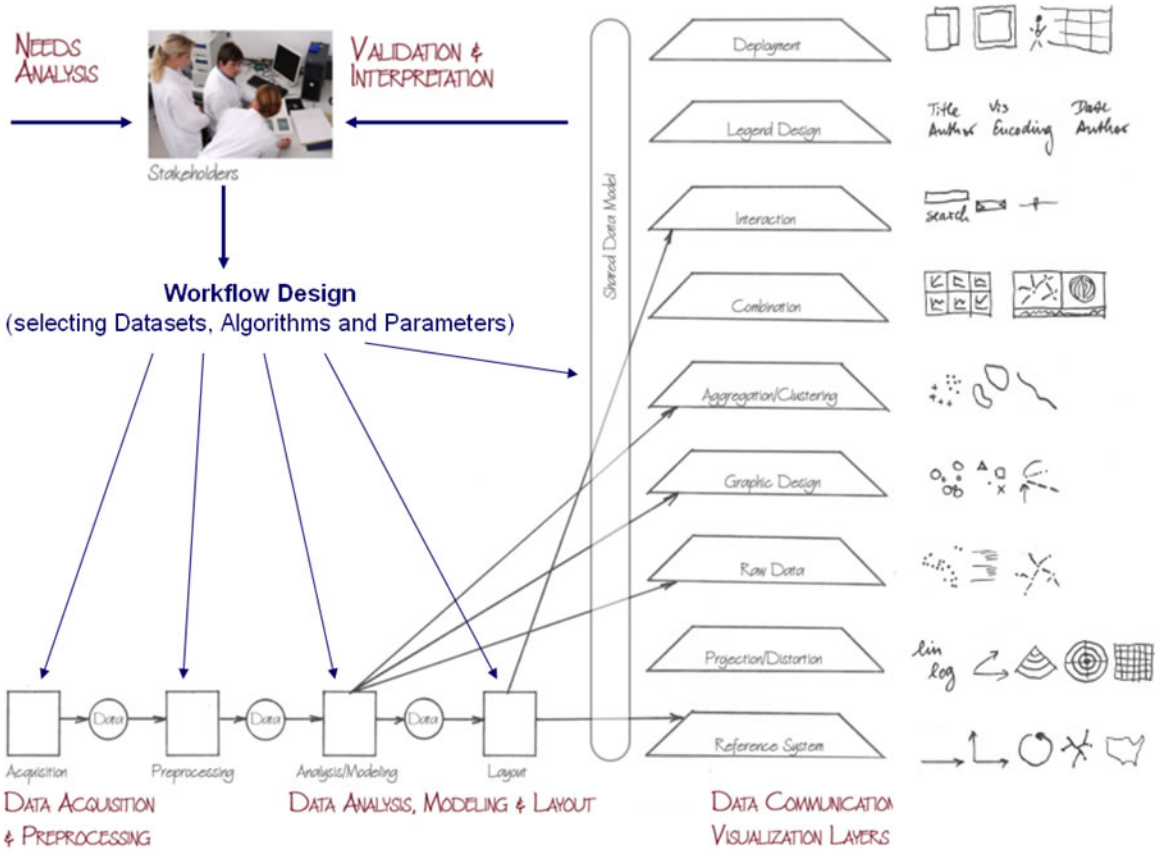


Figure 4.1: Needs-driven workflow design using a modular data acquisition/analysis/modeling/visualization pipeline as well as visualization layers

Note that the visualization layers interact with other workflow elements such as analysis algorithms (e.g., network analysis algorithms that compute additional node/edge attributes for graphic design, clustering techniques that indentify cluster boundaries), or layout algorithms (e.g., network layouts that compute a spatial reference system).

4.2 Data Acquisition and Preparation

- 4.2.1 Datasets: Publications
  - 4.2.1.1 Refer/BibIX/enw
  - 4.2.1.2 BibTeX
  - 4.2.1.3 ISI Web of Science
  - 4.2.1.4 Scopus
  - 4.2.1.5 Google Scholar
  - 4.2.1.5.1 Google Citation
- 4.2.2 Datasets: Funding
  - 4.2.2.1 NSF Award Search
  - 4.2.2.2 NIH RePORTER
- 4.2.3 Datasets: Scholarly Database

Typically, about 80 percent of the total project effort is spent on data acquisition and preprocessing; yet well prepared data is mandatory to arrive at high-quality results. Datasets might be acquired via questionnaires, crawled from the Web, downloaded from a database, or accessed as continuous data stream. Datasets differ by their coverage and resolution of time (days, months, years), geography (languages and/or countries considered), and topics (disciplines and selected journal sets). Their size ranges from several bytes to terabytes (trillions of bytes) of data. They might be high-quality materials curated by domain experts or random content retrieved from the Web. Based on a detailed needs analysis and deep knowledge about existing databases, the best suited yet affordable datasets have to be selected, filtered, integrated, and augmented. It may also be necessary for networks to be extracted (see section 4.7 Network Analysis for details).

4.2.1 Datasets: Publications

4.2.1.1 Refer/BibIX/enw

Refer was one of the first digital reference managers, developed by Bell labs in 1978. Refer's file output format has since been adopted by many tools and web services, including BibIX for UNIX, early versions of EndNote, CiteSeerX, Zotero.
Data in refer-formatted files can be used for the following types of analyses:

- Statistical Attributes** %1 (Times Cited)
- Temporal Analysis** %8 (Date)** %V (Volume)** %D (Year Published)
- Geospatial Analysis** %+ (Author Address)** %C (Place Published)
- Topical Analysis** %X (Abstract)** %J (Journal)** %K (Keywords)** %F (Label)
    - %! (Short Title)
    - %T (Title)
- Network Analysis** %A (Author)

4.2.1.2 BibTeX

Like Refer, BibTeX provides a standard reference file format used by many tools and web services, including CiteSeerX, citeulike, BibSonomy, and Google Scholar.
Data in BibTex files can be used for the following types of analyses:

- Temporal Analysis** date** bibdate** date-added** date-modified
    - issue
    - month
    - timestamp
    - volume
    - year
- Geospatial Analysis** address** location
- Topical Analysis** abstract** booktitle** conference** description
    - journal
    - keywords
- Network Analysis** author** organization

4.2.1.3 ISI Web of Science

ISI Web of Science (WoS) is a leading citation database cataloging over 10,000 journals and over 120,000 conferences. Access it via the "Web of Science" tab at http://www.isiknowledge.com (**note:** access to this database requires a paid subscription). Along with Scopus, ISI WoS provides some of the most useful datasets for scientometric analysis.

To find all publications by an author, search for the last name and the first initial followed by an asterisk in the author field. To find papers by Eugene Garfield, enter Garfield E* in the author field. The search yielded 1,529 results on November 11th 2009, 500 of which can be downloaded at a time, see Figure 4.2.

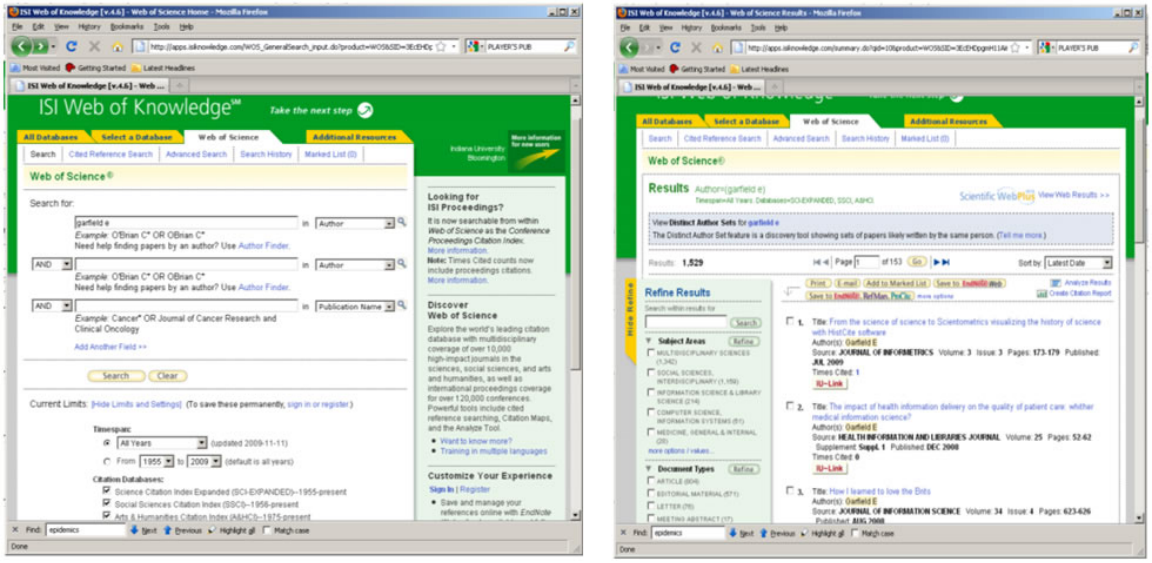
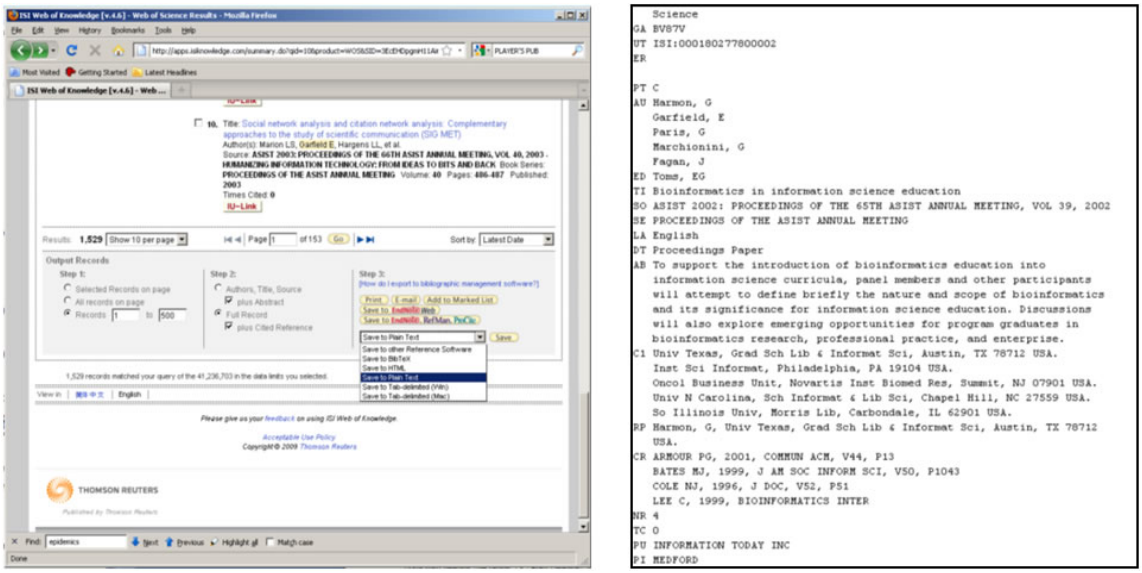


Figure 4.2: ISI Web of Science search interface and ISI Web of Science search results

Download the first 500 records using the output box at the bottom of the page. Enter records '1' to '500', select 'Full Record' and 'plus Cited Reference', select 'Save to Plain Text' in the drop down menu, and then click save. Wait for the processing to complete, and then save the file as *GarfieldE.isi.* The resulting file can be seen in Figure 4.3.



```
   Science
GA BV87V
UT ISI:000180277800002
ER

PT C
AU Harmon, G
   Garfield, E
   Paris, G
   Marchionini, G
   Fagan, J
ED Toms, EG
TI Bioinformatics in information science education
SO ASIST 2002: PROCEEDINGS OF THE 65TH ASIST ANNUAL MEETING, VOL 39, 2002
SE PROCEEDINGS OF THE ASIST ANNUAL MEETING
LA English
DT Proceedings Paper
AB To support the introduction of bioinformatics education into
   information science curricula, panel members and other participants
   will attempt to define briefly the nature and scope of bioinformatics
   and its significance for information science education. Discussions
   will also explore emerging opportunities for program graduates in
   bioinformatics research, professional practice, and enterprise.
C1 Univ Texas, Grad Sch Lib & Informat Sci, Austin, TX 78712 USA.
   Inst Sci Informat, Philadelphia, PA 19104 USA.
   Oncol Business Unit, Novartis Inst Biomed Res, Summit, NJ 07901 USA.
   Univ N Carolina, Sch Informat & Lib Sci, Chapel Hill, NC 27599 USA.
   So Illinois Univ, Morris Lib, Carbondale, IL 62901 USA.
RP Harmon, G, Univ Texas, Grad Sch Lib & Informat Sci, Austin, TX 78712
   USA.
CR ARMOUR PG, 2001, COMMUN ACM, V44, P13
   BATES MJ, 1999, J AM SOC INFORM SCI, V50, P1043
   COLE NJ, 1996, J DOC, V52, P51
   LEE C, 1999, BIOINFORMATICS INTER
NR 4
TC 0
PU INFORMATION TODAY INC
PI MEDFORD
```

Figure 4.3: Saving records from Web of Science and viewingGarfieldE.isi

ISI files are loosely based on the RIS file format, and data in this format can be used for the following types of analyses:

- Statistical Attributes** NR (Cited Reference Count)** TC (Times Cited)
- Temporal Analysis** RC (Date / Date Modified)** PD (Date Published)** IS (Issue)** CY (Meeting Date)
  - VL (Volume)
  - PY (Year)
- Geospatial Analysis** AD (Address)** C1 (Author Address)** CL (Meeting Location)** PA (Publisher Address)
  - PI (Publisher City)

    - RP (Reprint Address)
- Topical Analysis** AB (Abstract)** BS (Book Series Subtitle)** SE (Book Series Title)** CT (Conference Title)
    - ID (Index Keywords)
    - CT (Meeting Title)
    - MH (MeSH Terms)
    - A2 (Other Abstract)
    - SO (Source)
    - TI (Title)
    - FT (Vernacular Title)
- Network Analysis** AU (Author)** CR (References)** IV (Investigators)** AN (PubMed ID)

4.2.1.4 Scopus

Elsevier's Scopus, like ISI Web of Science, has an extensive catalog of citations and abstracts from journals and conferences. Subscribers to Scopus can access the service via http://www.scopus.com.
To find all articles whose abstract, title, or keywords include the terms 'Watts Strogatz Clustering Coefficient', simply enter those terms in the Article/Abstract/Keywords field. Twenty-five results were found as of November 11th, 2009. Download up to 2,000 references by checking the 'Select All' box and clicking 'Output'.

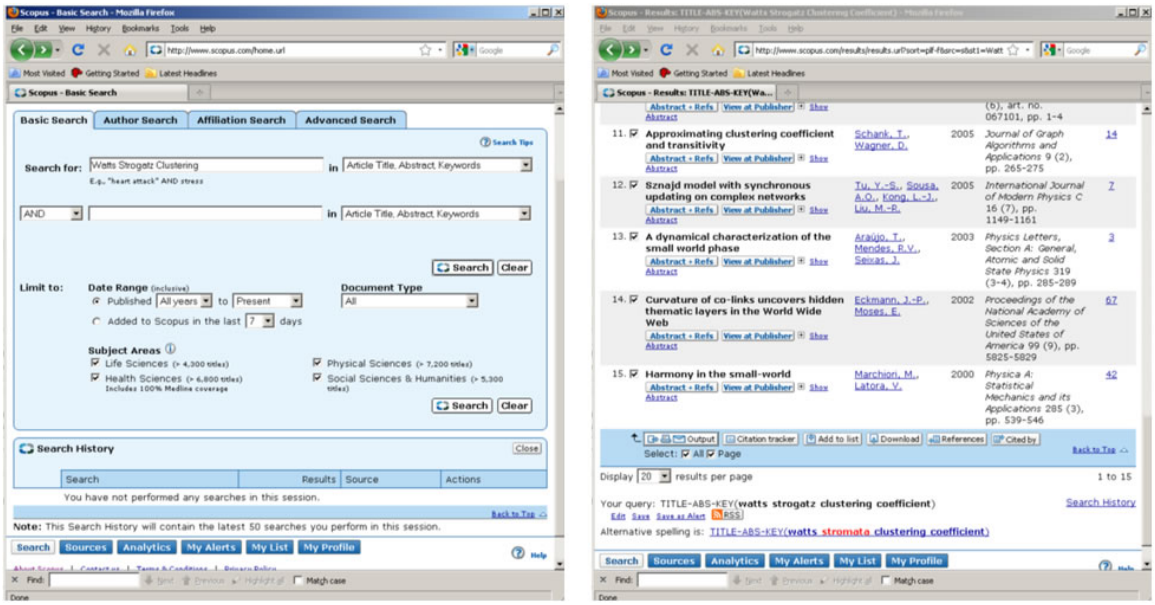
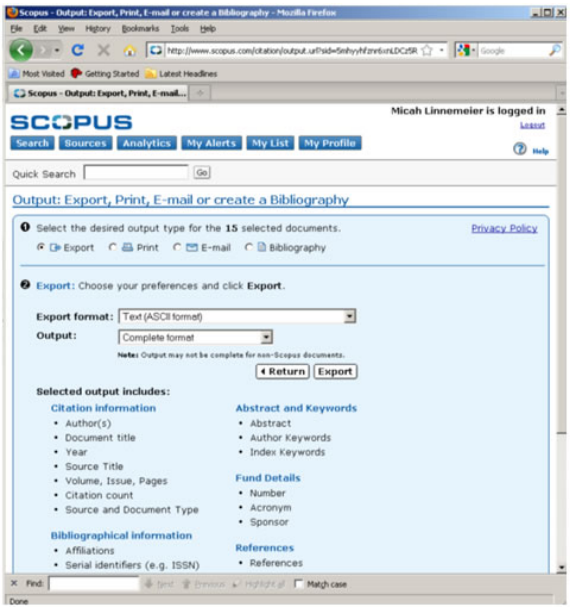


Figure 4.4: Scopus search interface and Scopus search results

At the output window, select *'Comma separated file, .csv'* (e.g. Excel) and 'Complete format' from the drop-down menus and choose *'Export'*. Save the file as *WattsStrogatz.scopus*. The resulting file can be seen in Figure 4.5.




```
Authors,Title,Year,Source title,Volume,Issue,Art
"Li K., Small M., Wang K., Fu X.","Three structu
"Yang X., Wang B., Wang W., Sun Y.","A novel sma
"Kaiser M.","Mean clustering coefficients: The r
"Da Fontoura Costa L., Andrade R.F.S.","What are
"Yang J., Xie Z., Sun Y.","Clustering effect on
"Chen Y.W., Zhang L.F., Huang J.P.","The Watts-S
"Park S.M., Kim B.J.","Dynamic behaviors in dire
"Yair Y., Aviv R., Ravid G., Yaniv R., Ziv B., P
"Li Y., Fang J.-Q., Liu Q., Liang Y.","Small wor
"Yang L.H., Holland M.D.","Small-world propertie
"Schank T., Wagner D.","Approximating clustering
"Tu Y.-S., Sousa A.O., Kong L.-J., Liu M.-R.","S
"Araujo T., Mendes R.V., Seixas J.","A dynamical
"Eckmann J.-P., Moses E.","Curvature of co-links
"Marchiori M., Latora V.","Harmony in the small-
```

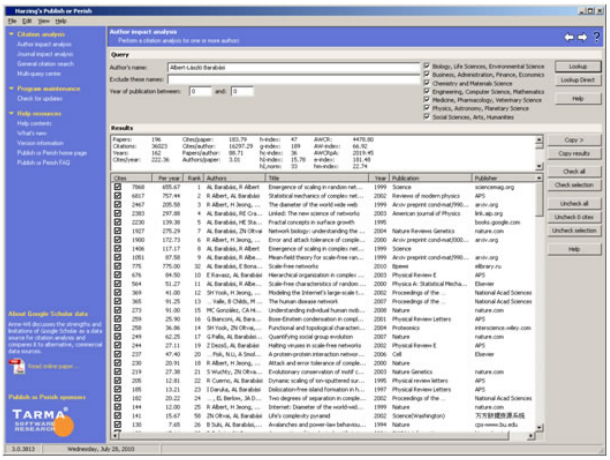
Figure 4.5: Saving records in Scopus and viewing WattsStrogatz.scopus

Data in Scopus files can be used for the following types of analyses:

- Temporal Analysis** Issue** Volume** Year
- Geospatial Analysis** Correspondence Address
- Topical Analysis** Abstract** Author Keywords** Conference Name** Index Keywords
  - Source Title
  - Source
  - Title
- Network Analysis** Authors** References

4.2.1.5 Google Scholar

Google Scholar data can be acquired using *Publish or Perish* (Harzing, 2008) that can be freely downloaded from http://www.harzing.com/pop.htm. A query for papers by Albert-László Barabási run on Sept. 21, 2008 results in 111 papers that have been cited 14,343 times, see Figure 4.6.



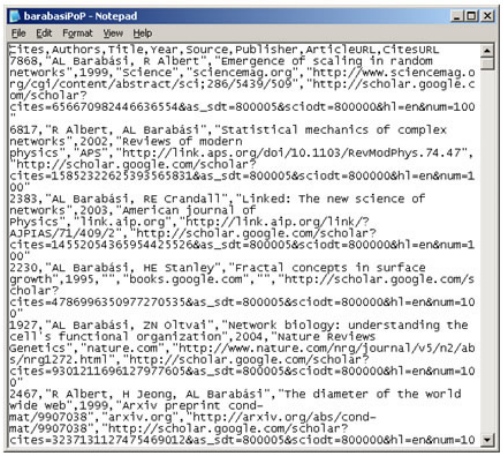


Figure 4.6: Publish or Perish search results for Albert-László Barabási and viewing barabasiPoP.csv

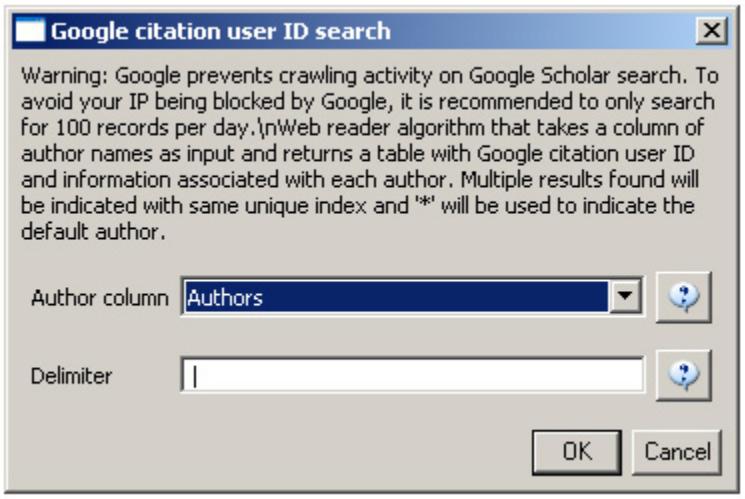
To save records, select '*File > Save'* from menu and then choose the appropriate file format (**.csv, *.enl, or *.bib) in the *'Choose File'* pop-up window. All three file formats can be read by the Sci^2 Tool. The result in all three**

formats named 'barabasi.' is also available in the respective subdirectories in '***yoursci2directory*** */sampledata/scientometrics/*' and will be used later in this tutorial.
Data from Google Scholar can be used for the following types of analyses:

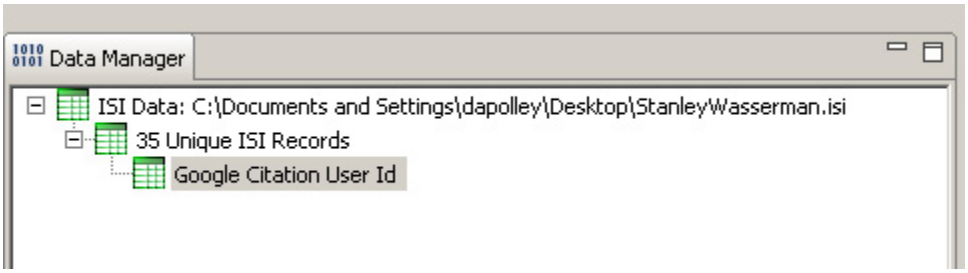
- Statistical Attributes** Cites
- Temporal Analysis** Year
- Topical Analysis** Source** Title
- Network Analysis** Authors

4.2.1.5.1 Google Citation

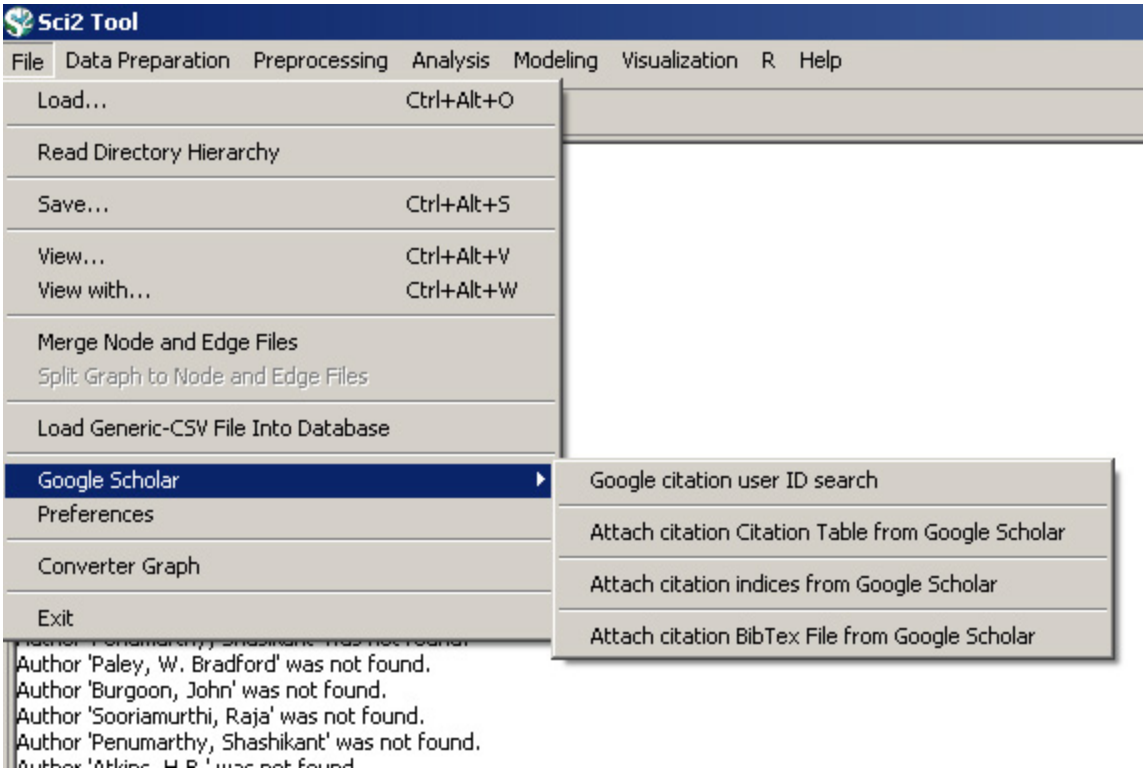
Data from Google Scholar can also be gathered based on Google Citation user IDs. Once any dataset with author information has been loaded into Sci2 (**.isi, .enl** etc.) run *File* > *Google Scholar* > *Google citation user ID search*' with the following parameter:



A list of authors for whom no Google Citation user ID could be found will print in the console. A table will be generated for those authors for whom Google Citation IDs could be found in the data manager:



Based on this table a variety of algorithms that require the Google Citation User ID as an input parameter can be run:



Please note that before any of the following algorithms can be run, the Google Citation User ID Search algorithm must be run to retrieve the Google Citation User IDs.

Attach citation Citation Table from Google Scholar

This algorithm creates a series of tables, one for each Google User ID, and creates tables with citation information for each Google Citation ID. Run '*File > Google Scholar >* *Attach Citation Table from Google Scholar*' with the default parameter.

- Statistical Attributes** Cites
- Temporal Analysis** Year
- Topical Analysis** Title
- Network Analysis** Authors

Attach citation indices from Google Scholar

This algorithm will create a citation index of the authors from the original file based on their Google Citation user ID (for those who have one). Run *'File > Google Scholar >* *Attach citation indices from Google Scholar'* with the default parameter.

- Statistical Attributes** Cites**h-index **i10-index
- Network Analysis** Citation User ID

Attach citation BibTex File from Google Scholar

This algorithm will create a BibTex file for each unique Google Citation user ID. Run *'File > Google Scholar >* *Attach citation BibTex File from Google* Scholar' with the default parameter.

- Temporal Analysis** date** bibdate** date-added** date-modified
  - issue
  - month
  - timestamp

    - volume
    - year
- Geospatial Analysis** address** location
- Topical Analysis** abstract** booktitle** conference** description
    - journal
    - keywords
- Network Analysis** author** organization

4.2.2 Datasets: Funding

4.2.2.1 NSF Award Search

Funding data provided by the National Science Foundation (NSF) can be retrieved via the *Award Search* site (http://www.nsf.gov/awardsearch). Search by PI name, institution, and many other fields, see Figure 4.7.

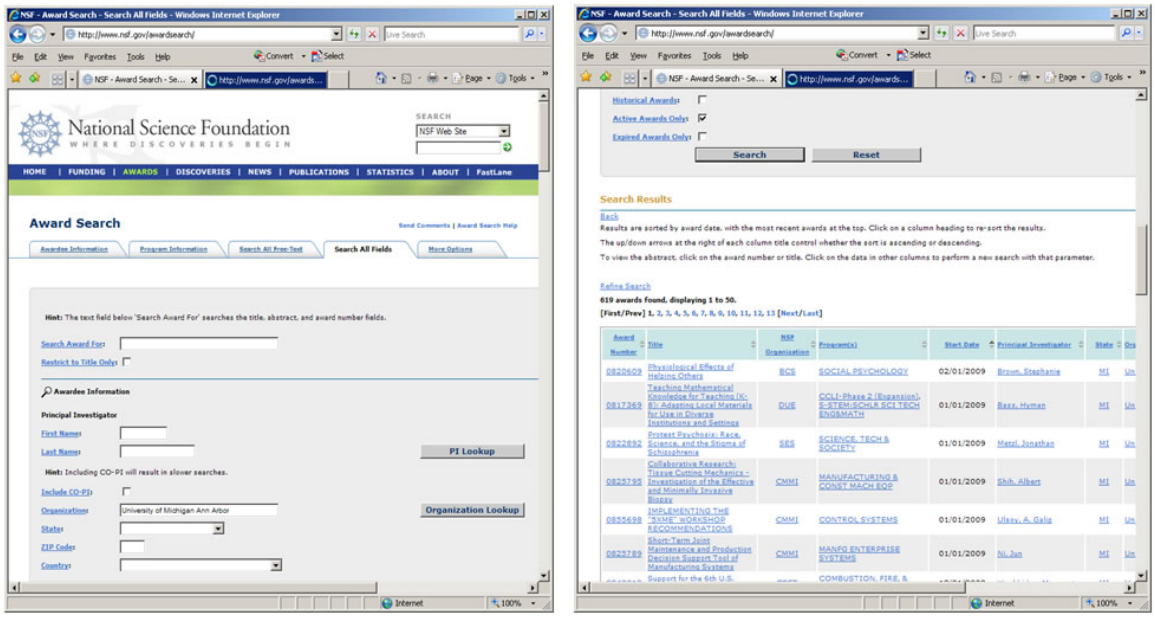


Figure 4.7: NSF 'Award Search' interface and search results page

To retrieve all projects funded under the Science of Science and Innovation Policy (SciSIP) program, simply select the 'Program Information' tab, do an 'Element Code Lookup', enter '7626' into the 'Element Code' field, and click the 'Search' button. On Sept 21st, 2008, exactly 50 awards were found. Award records can be downloaded in csv, Excel, or XML format. Save file in csv format, and change the file extension from .csv to .nsf. A sample .nsf file is available in '***yoursci2directory*** */sampledata/scientometrics/nsf/BethPlale.nsf*'. In the Sci^2 Tool, load the file using *'File > Load File'.* Select "NSF csv format" in the "Load" pop-up window. A table with all records will appear in the Data Manager. View the file in Excel.
Data in NSF files can be used for the following types of analyses:

- Network Analysis** Principle Investigator** Co-PI Name(s)** Organization
- Temporal Analysis** Expiration Date** Start Date
- Geospatial Analysis** Organization City** Organization State** Organization Street Address** Organization Zip
- Topical Analysis** Abstract** NSF Organization** Title

4.2.2.2 NIH RePORTER

Funding data provided by the National Institutes of Health (NIH), and associated publications and patents, can be

retrieved via the *NIH RePORTER* site (http://projectreporter.nih.gov/reporter.cfm). The database draws from eRA, Medline, PubMed Central, NIH Intramural, and iEdison. Search by location, PI name, category, etc., see Figure 4.8.

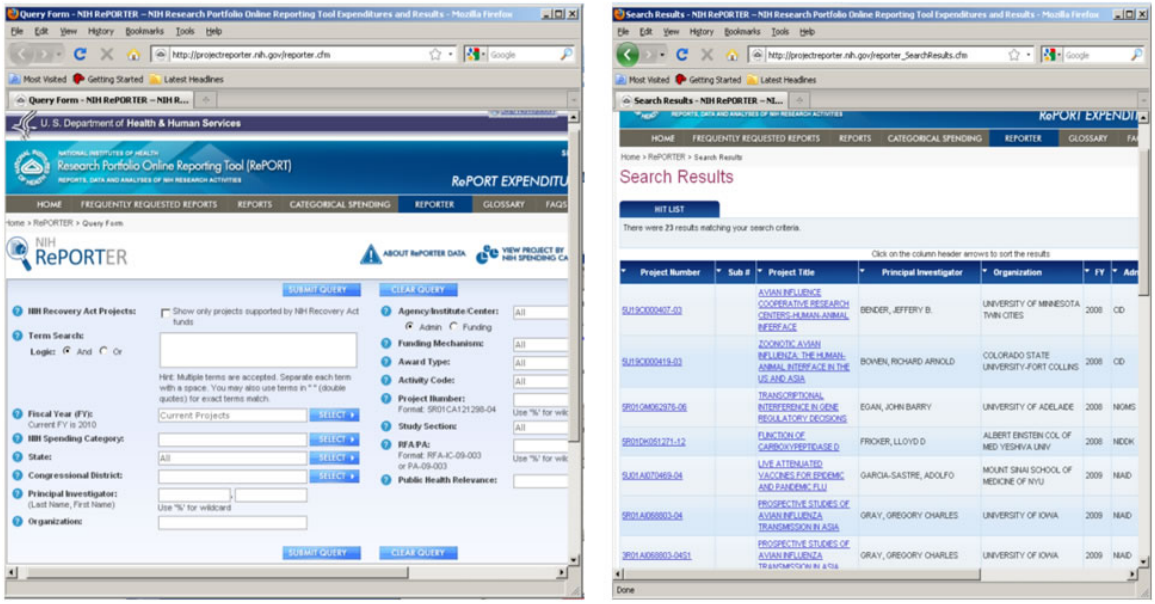


Figure 4.8: NIH RePORTER search interface and search results page

A sample search of "Epidemic" in the 'Public Health Relevance' field displays 205 results as of November 11th, 2009. Up to 500 results can be exported into csv or Excel format using the "Export" button at the top of the page. Save the file as a .csv and load it into the Sci^2 Tool using *'File > Load File'* to perform temporal or topical analyses. Data in NIH files can be used for the following types of analyses:

- Statistical Attributes** Type
- Temporal Analysis** Year of award
- Topical Analysis** Abstract** Project Title
- Network Analysis** Principle Investigator** Organization** Project Number

4.2.3 Datasets: Scholarly Database

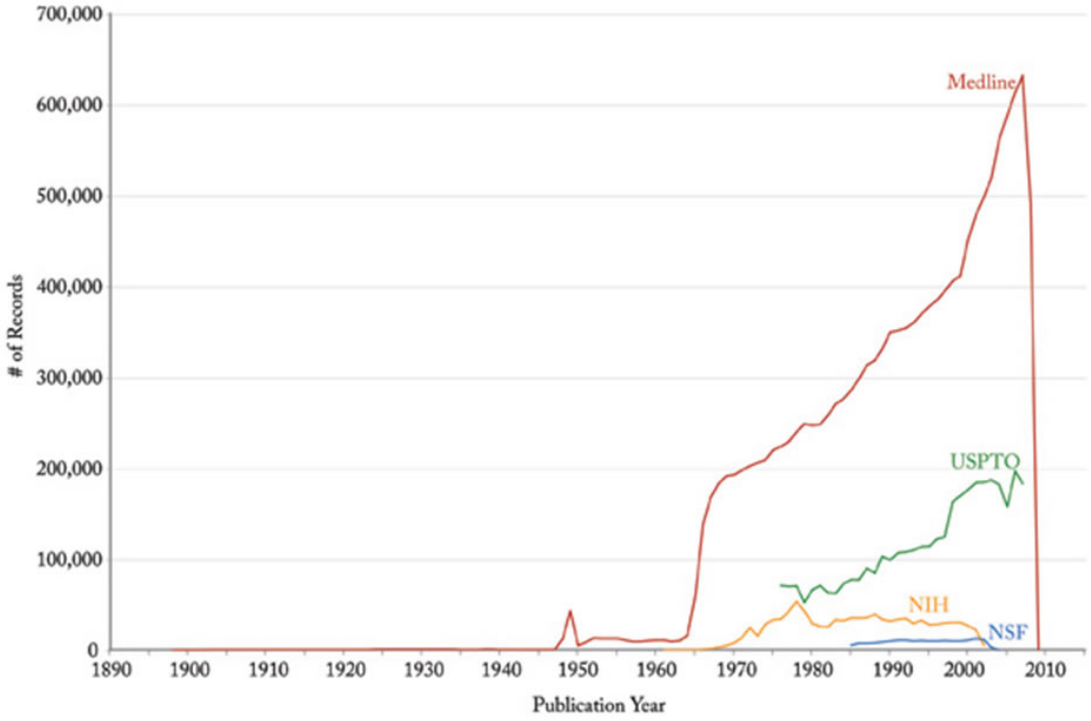


Figure 4.9: Graph of the numbers of records published each year by various organizations

Medline, U.S. patent, as well as funding data provided by the National Science Foundation and the National Institutes of Health can be downloaded from the Scholarly Database (SDB) at Indiana University. SDB supports keyword based cross-search of the different data types and data can be downloaded in bulk, see Figures 4.10 and 4.11 for interface snapshots.

Register to get a free account or use '*Email: nwb@indiana.edu*' and '*Password: nwb'* to try out functionality. Search the four databases separately or in combination for 'Creators' (authors, inventors, investigators) or terms occurring in 'Title,' 'Abstract,' or 'All Text' for all or specific years. If multiple terms are entered in a field, they are automatically combined using the Boolean operator 'OR.' Entering 'breast cancer' will match any record with 'breast' or 'cancer' in that field. Using the Boolean operator AND (for example, 'breast AND cancer') would only match records that contain both terms. Double quotations can be used to match compound terms, e.g., "breast cancer" retrieves records with the phrase "breast cancer," but not records where 'breast' and 'cancer' are present in isolation. The importance of a particular term in a query can be increased by putting a ^ and a number after the term. For instance, 'breast cancer^10' would increase the importance of matching the term 'cancer' by ten compared to matching the term 'breast.'

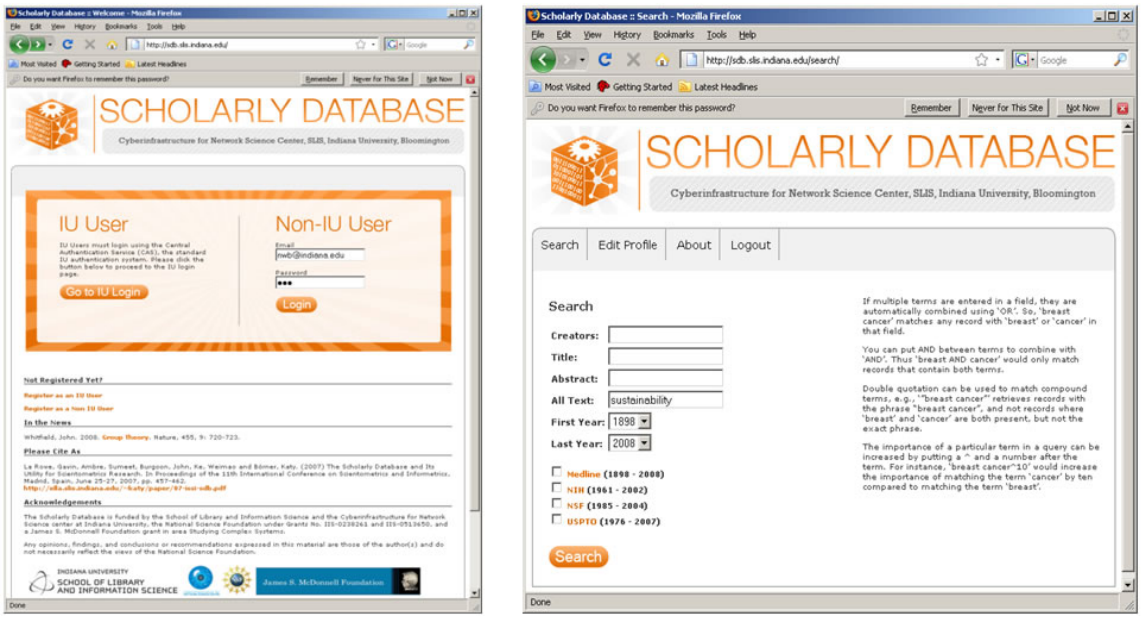


Figure 4.10: Scholarly Database 'Home' page and 'Search' interface

Results are displayed in sets of 20 records, ordered by a Solr internal matching score. The first column represents the record source, the second the creators, third comes the year, then title and finally the matching score. Datasets can be downloaded in different subsets and formats for future analysis.

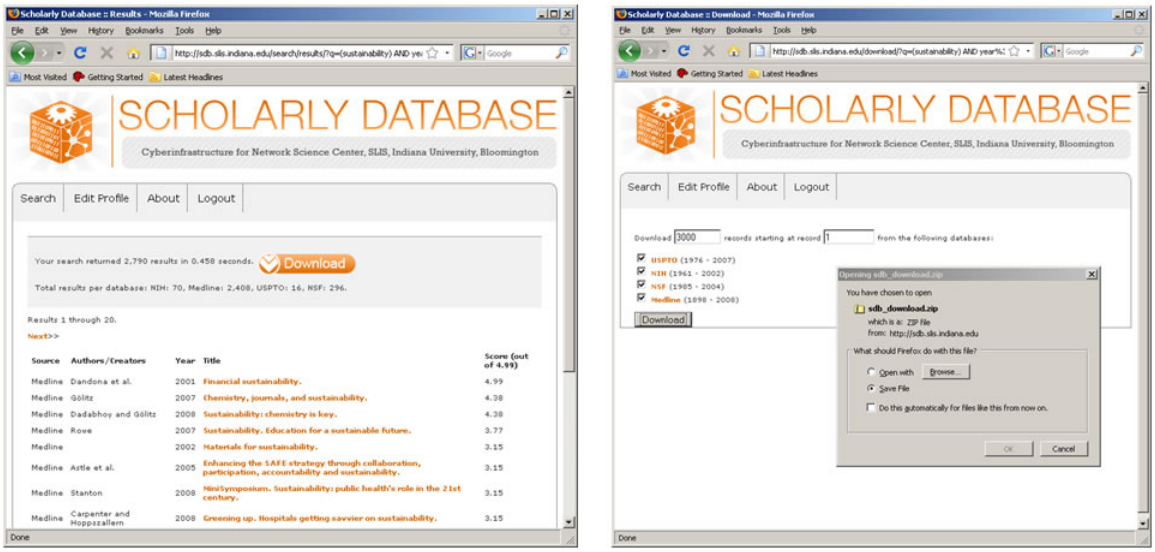


Figure 4.11: Scholarly Database search results and download interfaces

Data from the SDB can be used in a great number of ways. The following is an abridged list of suggested uses:

- Statistical Attributes** expected_total_amount** times_cited** citing_patents
- Temporal Analysis** issue_date** year** date_expires** project_end
    - issue
    - date_started
    - project_start
    - volume
    - published_year
- Geospatial Analysis** address** street** city** state
    - country
    - zipcode

    - residence
- Topical Analysis** abstract** descriptorname** nsf_org** title
    - article_title
    - Title
- Network Analysis** name** inventor** authors** cited_patents
    - investigators
    - pi_title

4.3 Database Loading and Manipulation

Coming soon.

4.4 Summaries and Table Extractions

Coming soon.

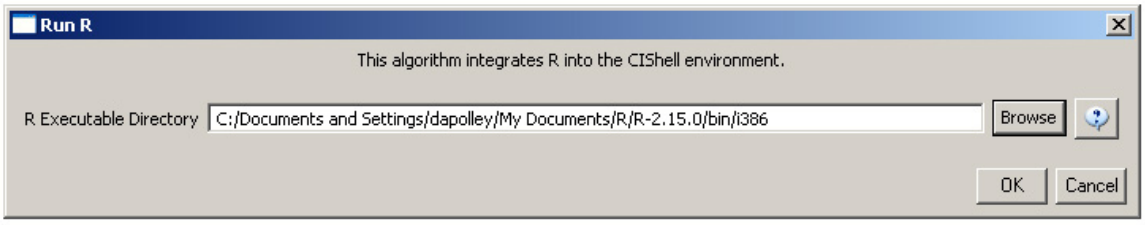
4.5 Statistical Analysis and Profiling

4.5.1 Using Sci2 with R

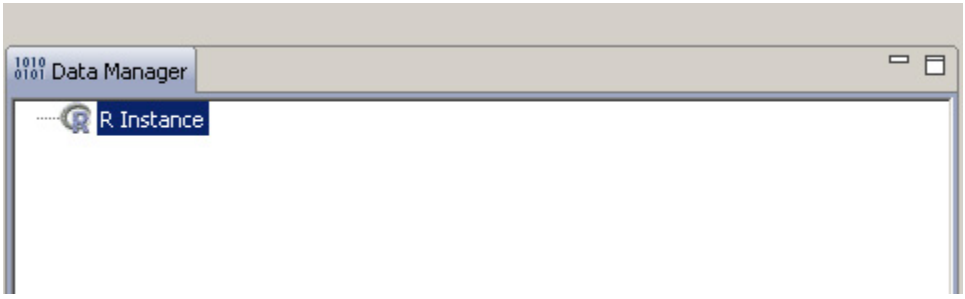
R is an environment and language for statistical analysis and modeling. R also enables graphical visualization. It is freely available under the GNU General Public Licenses. For more information on R and to download click here. You can use Sci to to create an R instance and send tables to R for analysis as well as extract tables from R.

Create an R Instance

Once R has been downloaded, an instance will need to be generated in order to import and export tables using Sci2. Run *'R > Create an R Instance'* to create an R instance. The R executable directory will need to be selected (on a Windows machine it will most likely be located here: `C:\Users\username\Programs\R-2.14.1\bin\i386`):



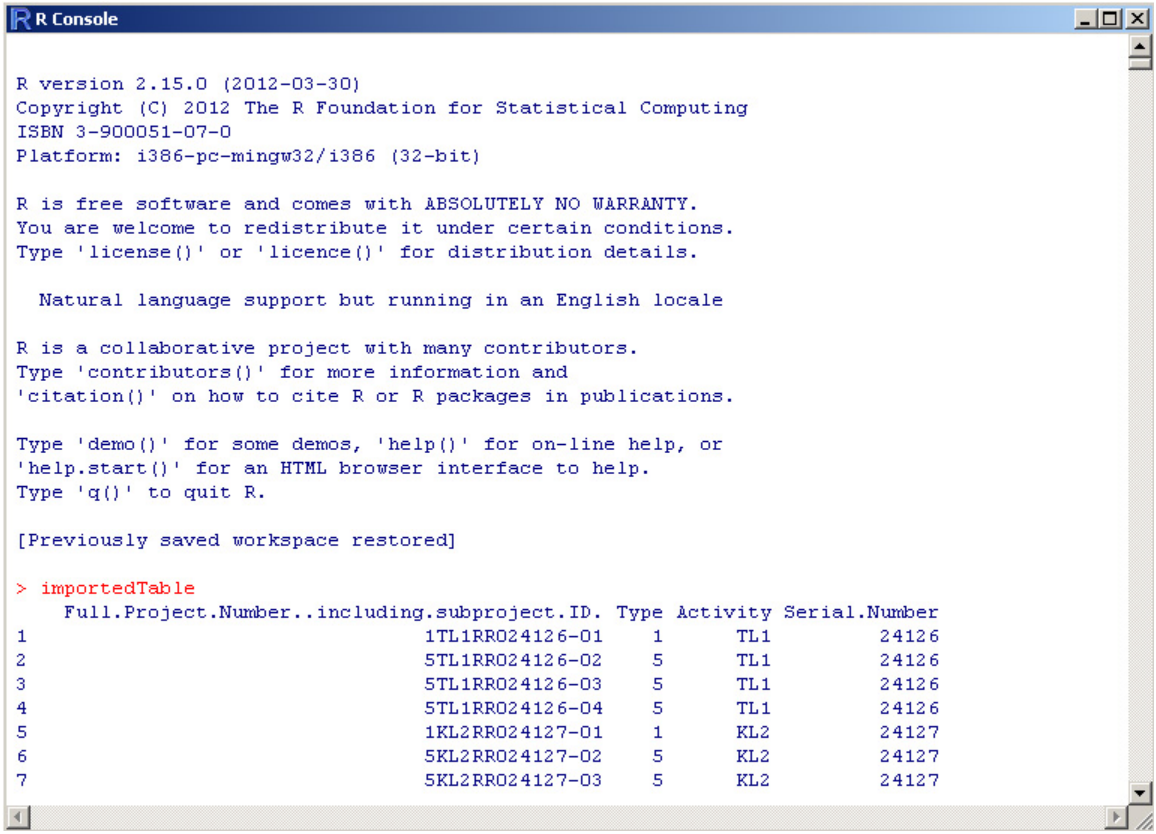
The R instance will then display in the console as follows:



Send Table to R

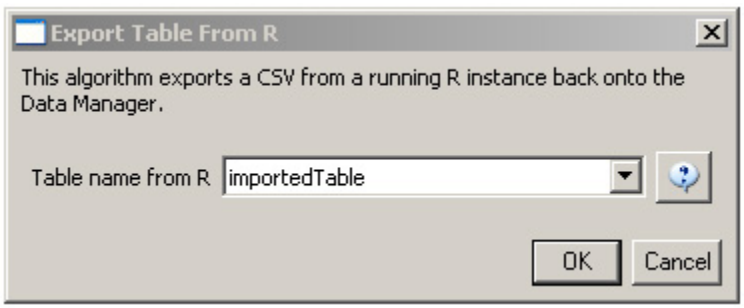
Tables in csv format cab be sent from Sci2 to the R instance for analysis and visualization. First, load a csv file in Sci2. Select both the loaded csv file and the R instance in the data manager (in Windows, hold the `ctrl` key as you click with the mouse). Run 'R > Send Table to R' and choose name for the imported table:

R can now be opened by running *'R > Run Rgui'* and the table is now available in the R environment using the variable name specified as a parameter in the "Import Table Into R". You can call the table in the R instance:



Get a Table from R

To get a table that has been imported to R run *'R > Get table from R'* and select the table from the drop-down menu:



The table will then show up in the data manager.

4.6 Temporal Analysis (When)

Science evolves over time. Attribute values of scholarly entities and their diverse aggregations increase and decrease at different rates and respond with different latency rates to internal and external events. Temporal analysis aims to identify the nature of phenomena represented by a sequence of observations such as patterns, trends, seasonality, outliers, and bursts of activity.
A time series is a sequence of events or observations that are ordered in time. Time-series data can be continuous (i.e., there is an observation at every instant of time) or discrete (i.e., observations exist for regularly or irregularly spaced intervals). Temporal aggregations – over journal volumes, years, or decades – are common.
Frequently, some form of filtering is applied to reduce noise and make patterns more salient. Smoothing (i.e., averaging using a smoothing window of a certain width) and curve approximation might be applied. The number of scholarly records is often plotted to get a first idea of the temporal distribution of a dataset. It might be shown in total values or as a percentage of those. One may find out how long a scholarly entity was active; how old it was at a certain point; what growth, latency to peak, or decay rate it has; what correlations with other time series exist; or what trends are observable. Data models such as the least squares model – available in most statistical software packages – are applied to best fit a selected function to a data set and to determine if the trend is significant.
Kleinburg's burst detection algorithm is commonly applied to identify words that have experienced a sudden change in frequency of occurrence.

4.6.1 Burst Detection

A scholarly dataset can be understood as a discrete time series, i.e., a sequence of events/observations which are ordered in one dimension – time. Observations (e.g., papers) come into existence at regularly spaced intervals, (e.g., each week, month, issue, volume, or year).
Kleinberg's burst detection algorithm (Kleinberg, 2002) identifies sudden increases in the frequency of words. Rather than using simple frequencies of the occurrences of words, the algorithm employs a probabilistic automaton whose states correspond to increasing frequencies of individual words. State transitions correspond to points in time around which the frequency of the word changes significantly. The algorithm generates a list of the word bursts in the document stream, ranked according to the burst weight, together with the intervals of time in which these bursts occurred. The burst weight depicts the intensity of the burst, i.e., how great the change in the word frequency that triggered the burst. The user must choose if s/he wants to detect bursts related to author names, journal names, country names, references, ISI keywords, or terms used in the title and/or abstract of a paper. This can serve as a means of identifying relevant topics, terms, or concepts that increased in usage, were more active for a period of time, and then faded away.
Kleinberg's burst detection algorithm has four important parameters: the gamma value, first ratio, general ratio, and the number of bursting states. The gamma parameter controls the ease with which the automaton can change states. The higher the gamma value, the smaller the list of bursts generated. The first ratio and general ratio control how great the change of frequency of a word must be to be considered a burst. Usually, only one bursting state is used, since one bursting state is sufficient to indicate whether or not a burst occurred for a specific character string within a specific time period. However, if the user wishes to identify bursts inside bursts, s/he must use more than one bursting state to capture such a hierarchical structure. In Sci^2, default values are provided for all parameters.
The Sci^2 tool currently uses this algorithm in *'Analysis > Temporal > Burst Detection'*.
Because the algorithm itself is case-sensitive, care must be taken if the user desires 'KOREA' and 'korea' and

'Korea' to be identified as the same word. It is recommended that the user normalize the target data before applying the burst algorithm. The normalization process separates text into word tokens, normalizes word tokens to lower case, removes "s" from the end of words, removes dots from acronyms, deletes stop words, and applies the English Snowball stemmer (http://snowball.tartarus.org/algorithms/english/stemmer.html. The normalization of an entry column can be done using the menu '*Preprocessing > Topical > Normalize Text*'.
For an example on how to use the burst detection algorithm, see Section 5.2.5 Burst Detection in Physics and Complex Networks.

4.6.2 Slice Table by Time

Slicing a table allows the user to see the evolution of a network over time. Time slices can be cumulative, i.e., later tables include information from all previous intervals, or fully sliced, i.e., each table only includes data from its own time interval. Cumulative slices can be useful for seeing growth over time, whereas fully sliced tables should be used for displaying changing structure over time.

4.6.3 Horizontal Bar Graph

Horizontal bar graphs also allow the user to visualize numeric data over time. This algorithm accepts csv (tabular) datasets, including NSF grant data and the output of burst detections.
In this visualization, the x-axis is time, and the y-axis is amount. The output of this visualization consists of labeled, color-coded horizontal bars that correspond to records in the original dataset.
This algorithm accepts the following input parameters:

- Label: The field used to label the bar lines
- Start Date: The field used for the starting point (on the x-axis) of the records that are visualized as horizontal bars
- End Date: The field used for the ending point (on the x-axis) of the records that are visualized as horizontal bars
- Size Bars By: The field used to size the horizontal bars
- Minimum Amount Per Day For Bar Scaling: The smallest value necessary to allow for scaling the numeric data visualized in the horizontal bars
- Bar Scaling: Can be either linear or logarithmic
- Date Format: The date format to use for parsing the Start Date and End Date.
- Year Label Font Size: The font size to use for year labels. NOTE: If the font size is too large, the labels may overlap.
- Bar Label Font Size: The font size to use for bar labels. NOTE: Extremely large values here may result in poor results.
- Colorized by: The field values used to color code the horizontal bars

4.7 Geospatial Analysis (Where)

Geospatial analysis has a long history in geography and cartography. Geospatial analysis aims to answer the question of where something happens and with what impact on neighboring areas.
Geospatial analysis requires spatial attribute values or geolocations for authors and their papers, extracted from affiliation data or spatial positions of nodes, generated from layout algorithms. Geospatial data can be continuous (i.e., each record has a specific position) or discrete (i.e., each set of keywords has a position or area-shape file –

e.g., number of papers per country). Spatial aggregations (e.g., merging via ZIP codes, counties, states, countries, and continents) are common.
Cartographic generalization refers to the process of abstraction such as (1) graphic generalization: the simplification, enlargement, displacement, merging, or selection of entities without enhancing their symbology; and (2) conceptual symbolization: the merging, selection, and symbolization of entities, including enhancement – such as representing high-density areas with a new (city) symbol.
Geometric generalization aims to solve the conflict between the number of visualized features, the size of symbols, and the size of the display surface. Cartographers dealt with this conflict intuitively in part until researchers like Friedrich Töpfer attempted to solve them with quantifiable expressions.

4.7.1 Extract ZIP Code

Description

This algorithm parses the address information provided and extracts ZIP codes from it. Currently it accepts ZIP codes which are in United States of America format i.e. either XXXXX (short form) or XXXXX-XXXX (long form).

Pros & Cons

This algorithm facilitates quick Spatial analysis by extracting ZIP codes from a given address, which can be further processed. Its only limitation is that currently it only supports parsing of USA ZIP Codes or countries which have USA based ZIP Code format.

Implementation Details

The algorithm works as follows,

1. Get the address for each row & begin parsing for zip codes in following manner,
    a. Save all groups of digits along with their start position, end position & length of the group of digits.
    b. Since the zip code, presumably, will be in the later portion of the address string, traverse the collected zip code candidates in the reverse fashion.
    c. If there is a 5 digit group then,
        i. Consider it as the primary zip code. If the user wants the ZIP code in truncated form then the algorithm will skip the checking for extension ZIP code.
        ii. The extension of the ZIP code follows the primary zip code, so check if the previous group has length 4 and if so, then check if its distance from primary zip is less than or equal to 2. If yes, than consider this as the extension of the zip code.
        iii. If there is no 4 digit group satisfying the above conditions then return null as the extension value.
        iv. If there is no 5 digit group then return null for the primary zip code value. In this case, display a warning to the user that no ZIP code was found for this particular address string.

Usage Hints

The user has to provide 3 inputs; a file containing the addresses for which ZIP code parsing is required, whether to truncate the parsed ZIP code or not and name of the address column. If the plugin was unable to find any ZIP code then it will print a warning message and set the ZIP code to empty string. The data for ZIP codes can be in either short form i.e. XXXXX or long form i.e. XXXXX-XXXX. It will also accept ZIP code information in the following format,

XXXXX<Any Character(s) of Max Length = 2>XXXX.
The output of this algorithm will be the original input table with 1 column added containing the parsed ZIP code.

4.7.2 Generic Geocoder

Description

This algorithm provides a general-purpose geocoding functionality that is expanded on by other more specific geocoding algorithms (see Yahoo! Geocoder). It supports four types of geocoding: address, country, U.S. states and U.S. ZIP codes.

Pros & Cons

1. Increase of code re-used with abstract classes that defines the common behaviors such as GUI layout, data handling, etc.
2. Standard GUI layout gives a professional look to the application. Once user learns to use one geocoder, it will be same for using other geocoder plugins.
3. Problems can arise if the inherited behaviors are not well defined, since a single change on the abstract class will cause changes on all sub-class.

Applications

Plugins that use this interface provide geographical coordinate information for the geomap application. Scientists can then visualize their data geographically.

Implementation Details

This algorithm provides a common front-end behaviors algorithm for multiple geocoder plugins. It uses MVC (Model-View-Controller) idea to facilitate the in-dependency and code reused implementation.

1. View
   - ***AbstractGeocoderFactory*** defined GUI layout, data validation and geocoder type selection. It contains a *FamilyOfGeocoder* member that refer to the related geocoder family (Generic, Yahoo, etc)
2. Controller
   - ***GeocoderAlgorithm*** processes the geolocation look up using the given *Geocoder*. The first look up is through invoking *geocodingFullForm*. If the look up failed, the second look up will be performed through invoking *geocodingAbbreviation*. It also provide error handling which analyzes the look up failures and provides appropriate warning message to user. In success, it generates a CSV output file with two additional columns that hold latitude and longitude values.
   - ***FamilyOfGeocoder*** contains four type of geocoders from the smae family. There are address, country, U.S. states and U.S. ZIP codes.
3. Model
   - It uses the common geocoder model that defined in edu.iu.scipolicy.model.geocode. It uses *Geolocation* that represents geographical coordinate and *USZipCode* that contains uzip (the first 5 digit ZIP code) and postbox number (the last 4 digits number in 9-digits ZIP code).
   - Each geocoder might holds its own model if needed

Usage Hints

The usage is provided on each geocoder wiki page. For example, Yahoo! Geocoder.

Links

- Source Code
- External Package

Acknowledgments

The geocoding algorithm was authored, modified, integrated and documented by Chin Hua Kong. Many thanks to Chintan Tank first Generic Geocoder implementation that provide a based code to start from.

4.7.3 Yahoo Geocoder

Description

This algorithm converts place names or addresses into Latitude, Longitude co-ordinates. It accepts international addresses, countries, States of United States of America and ZIP codes of United States of America. All co-ordinates are obtained by querying Yahoo! PlaceFinder service. Internet access must be available during geocoding.

Pros & Cons

1. The performance is slower than theGeocoder and may vary due to the network latency since the queries are requested through internet service. The benchmark test geocoded 470 unique locations per minute
2. Yahoo! Geocoder supports address geocoding with international coverage which is not supported by Geocoder.
3. To use Yahoo! Geocoder, user has to obtain an application id through Yahoo! registration. Save your application id and provide it when requested by the Yahoo! Geocoder. Since each application id is allowed to geocode 50,000 locations per 24 hours, the user is encouraged to test on a small set of data first.

Applications

The plugin is useful for scientists who would like to visualize their data on a geographical map (geomap). User can obtain the geographical coordinates (Latitude and Longitude values) and feed them to the visualization plugin.

Implementation Details

The algorithm receives a list of input data (locations) and queries their locations one by one through Yahoo! PlaceFinder geocoding service. The results will temporarily be cached in memory so that the same query for duplicated locations can be avoided. The cache is deleted after each user request is completed. This plugin is included with the Sci2 application. Performance of this algorithm is O(n).

The detail of the algorithm is shown as following,

1. Yahoo! Gecoder is favored by MVCs (Model-View-Controller)
2. View
   - ***YahooGeocoderFactory*** is extending *AbstractGeocoderFactory*. It defines all the display options and user input wrapping.. The detail of the implementation can be accessed through here.
3. Controller
   - ***GeocoderAlgorithm*** is the shared geocoder controller which is documented with *AbstarctGeocoderFactory*. It invoked *YahooFamilyOfGeocoder's* methods to retrieve the geocoder based on the selected type and performs the geocoding operation.
   - ***YahooFamilyOfGeocoder*** contains four geocoders: *YahooCountryCoder*, *YahooStateCoder*, *YahooZipCodeCoder*, *YahooAddressCoder*. Each coder will only be created when it is invoked
   - ***Geocoder*** contains the geocodingFullForm and geocodingAbbreviation methods which will invoked *PlaceFinderClient* to request Yahoo! PlaceFinder service.
   - ***PlaceFinderClient*** will performs the service query to Yahoo! PlaceFinder. Every request has three times retry for network failure. The result will be returned as ResultSet which is defined in the model
4. Model
   - The model classes were generated by using placeFinder.xsd and located in *edu.iu.scipolicy.preprocessing.geocoder.coders.yahoo.placefinder.beans*.
   - The JAXB technology is chosen as the un-marshaller since Yahoo! PlaceFinder have a simple and standard service response definition at here. JAXP is more suitable for complicated data model which provides more control in data preprocessing.
5. Dependency: javax.xml.*

Usage Hints

Here is a 7 steps guide to using the plugin:

1. Make sure you are connected to the internet.
2. Load an input data table that contains locations to be geocoded.
3. Select **Analysis > Geospatial > Yahoo! Geocoder** from the menu bar. A window will pop up.
4. Enter your application id. You can obtain one from here.
5. Choose place type that represents your input location data. The place type can be address, country, U.S. state or U.S. ZIP code.
6. Choose place name column that represents the location field in your data file.
7. Select Include address details if you want Yahoo! to return the parsed address information.
8. Press Ok to start geocoding.

All rows of the data will be geocoded one by one using Yahoo! PlaceFinder. Emtpy entries and invalid locations that failed to be geocoded are listed in the console.

The output of this algorithm is the original input table with two additional for latitude and longitude. Locations that failed to be geocoded will have blank entries.

Performance varies by machine and network latency. Our benchmark is 470 unique locations per minutes.

Links

- Source Code
- Root Source
- External Package
- Yahoo! PlaceFinder

Acknowledgments

The geocoding algorithm was authored, implemented, integrated and documented by Chin Hua Kong. Many thanks to Micah Linnemeier for providing guidance of the plugin integration. I would also like to acknowledge the usage of Yahoo! PlaceFinder's geocoding service.

4.7.4 Congressional District Geocoder

Description

This algorithm converts the given **9-digits U.S. ZIP codes (ZIP+4 codes)** into its congressional districts and geographical coordinates (latitude and longitude). The Benchmark is 50,000 ZIP codes per second. Download the plugin here.

Pros & Cons

1. The algorithm is using a local database mapping with 25MB file size. It will increase the application size dramatically. So it is build as an external plugin
2. For first execution in the same application window, the plugin required 5 seconds to load the database. The consequent execution will not required the pre-loading phase.
3. Since some 5-digits ZIP codes contain multiple districts, the 9-digits ZIP codes is required for the conversion. Warning message will be printed to notice user if the given 5-digits ZIP codes contain multiple districts
4. Congressional district might be varied by each election. The database would need to be maintained and updated relatively.

Applications

This plugin only support U.S. ZIP codes. It convert 9-digits ZIP codes to their belonging congressional district. It is an external plugin since the data size is so large. The dataset is based on the year 2008 election.

Implementation Details

Words for developers: Please do take a look at the ZIP code wiki at here *to have a better understand on how U.S. ZIP+4 code system works. The first 5-digits number in ZIP code is called Uzip. The last 4-digits number in the ZIP+4 code is Post Office box number which can refer to* here*.*

The challenge of the implementation is the design of the mapping model that used to look up congressional districts from ZIP+4 codes. To understand the metadata file (provided by GovTrack), create a mapping model with constant (O(1)) look up time and easy to managed. The implementation detail is documented in the source code.

The following will provide a high level view of the design.

1. The algorithm is facilitated by the Model-View-Controller idea
2. Model
   - The core of this implementation. Formed by *ZipCodeToDistrictMap*, *PostBoxToDistrictMap* and DistrictRegistry.
   - ***ZipCodeToDistrictMap*** hold a map of uzip to *USDistrict* and a map of uzip to *PostBoxToDistrictMap*.
   - ***PostBoxToDistrictMap*** hold a map of postBox to *USDistrict* and a map of wildcard to *USDistrict* map.
   - ***DistrictRegistry*** contains non-duplicated of *USDistrict* objects. It holds entire U.S congressional districts information.
   - ***USDistrict*** contains district label and geolocation. The class is imported from *edu.iu.scipolicy.model.geocode* package
3. View - ***ZipToDistrictAlgorithmFactory*** contains all the view setup implementation, including title, windows and options
4. Controller
   - ***ZipToDistrictAlgorithm*** prepares the model; parses the input ZIP codes to *USZIPCode* objects; performs the district look up, handles exceptions and saves the result to a CSV file.
   - The **Look up** is performed through *ZipCodeToDistrictMap*. If there isn't found a direct match of uzip to *USDistrict*, it will performed a look up through *PostBoxToDistrictMap* that holds by the uzip. Return *USDitrict* in success while throws *ZipToDistrictException* if no matched found
5. Dependency: dist2geolocation.txt and zip4dist-prefix.txt

The output table contains all columns of the input table with three new columns (Congressional district, latitude and longitude).

Usage Hints

Here is a four steps guide to use the plugin:

1. Load your input data file that contains 9-digits U.S. ZIP codes to be geocoded.
2. Select **Analysis > Geospatial > Congressional District Geocoder** from menu bar. A window will be pop up
3. Choose place name column that represents the ZIP code field in your data file.
4. Press Ok button to start the geocoding

5-digits ZIP codes with multiple congressional districts, empty entries and invalid ZIP codes that failed to be geocoded will list in warning messages on the console.

The output of this algorithm is the original input table with additional 3 columns (Congressional district column, latitude column and longitude column). ZIP codes that failed to be geocoded will have blank entries.

Our benchmark is 50,000 ZIP codes per second.

Geomap the congressional districts

1. Firstly, you might want to aggregate your data based on congressional district. To do this, you can follow user hints at here.
2. You are ready to plot your aggregated result to geomap. It is recommended to plot the congressional district results on a country map due to some U.S. districts are located outside of the America Continents. To geomap the congressional districts, please follow the user hints at here.

Enjoy!!!

Links

- Source Code
- External Package
- Home Page
- GovTrack
- WATCHDOG.NET
- ZIP code's wiki

Acknowledgments

The geocoding algorithm was authored, implemented, integrated and documented by Chin Hua Kong. Many thanks to the Sprint team for providing advices and suggestions. Many thanks to GovTrack that provides ZIP to district mapping data and district's geolocation information. Thanks to Carl Malamud and Aaron Swartz, that make the data available on WATCHDOG.NET for GovTrack.

4.7.5 Geo Map (Circle Annotations) and Geo Map Colored Region Annotations

Choropleth Map

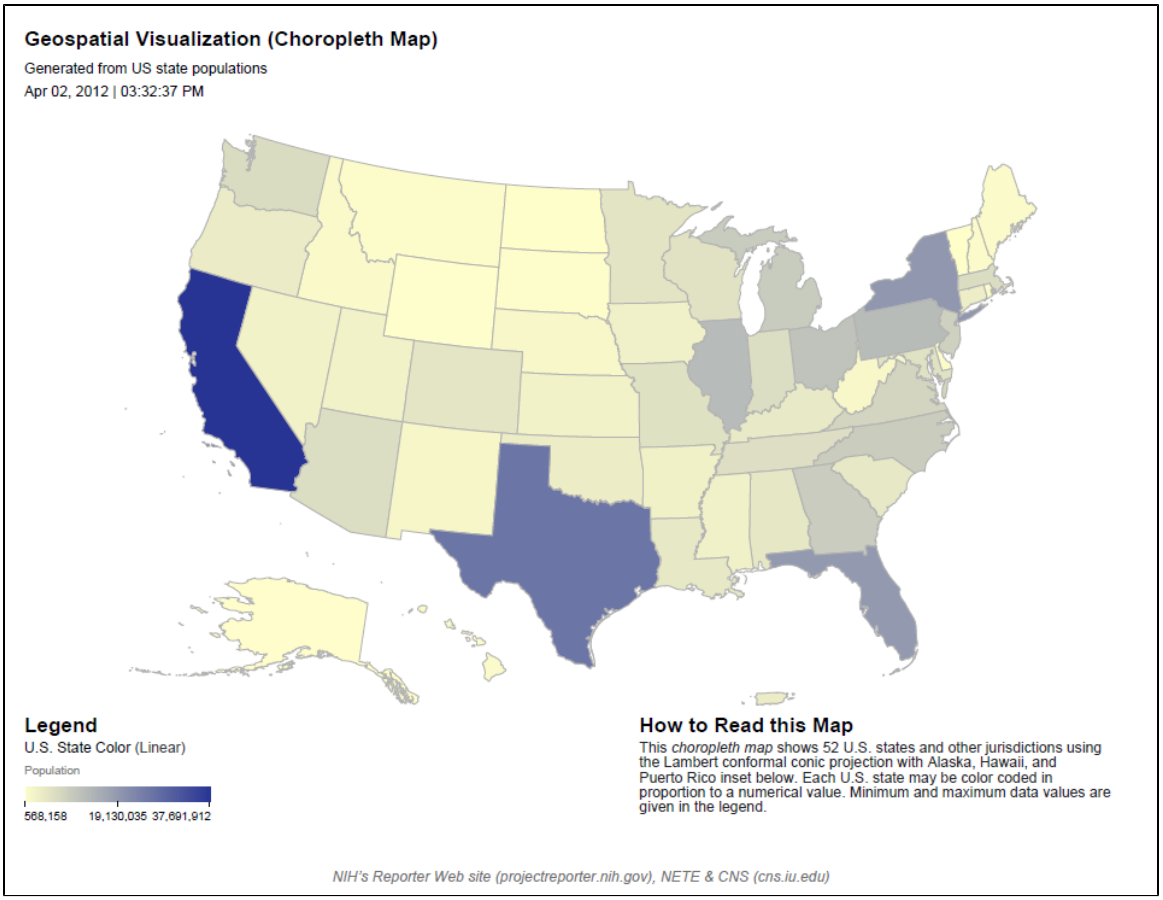
Color countries of the world or states of the US in proportion to numeric data.

- When using the **world** base map you input data must include a text-valued column that identifies **countries** using the names listed in Recognized country names.
- When using the **United States** base map your input data must include a text-valued column that identifies **states** using the names listed in Recognized state names.

Features

- Optionally scale each individual dimension of numeric data logarithmically or exponentially.
- Legends for each dimension show how extrema of the data correspond to visual representations.

Examples



Download source CSV
Download PDF

Usage Hints

- It may be difficult to see any color coding applied to comparatively small regions on the map. In this case you may wish to geocode your region names as (longitude, latitude) coordinates and use Proportional Symbol Map, perhaps disabling circle size coding and exterior color coding, then using for interior color coding whichever column you would have used with Choropleth Map.

Usage Hints

- Source Code

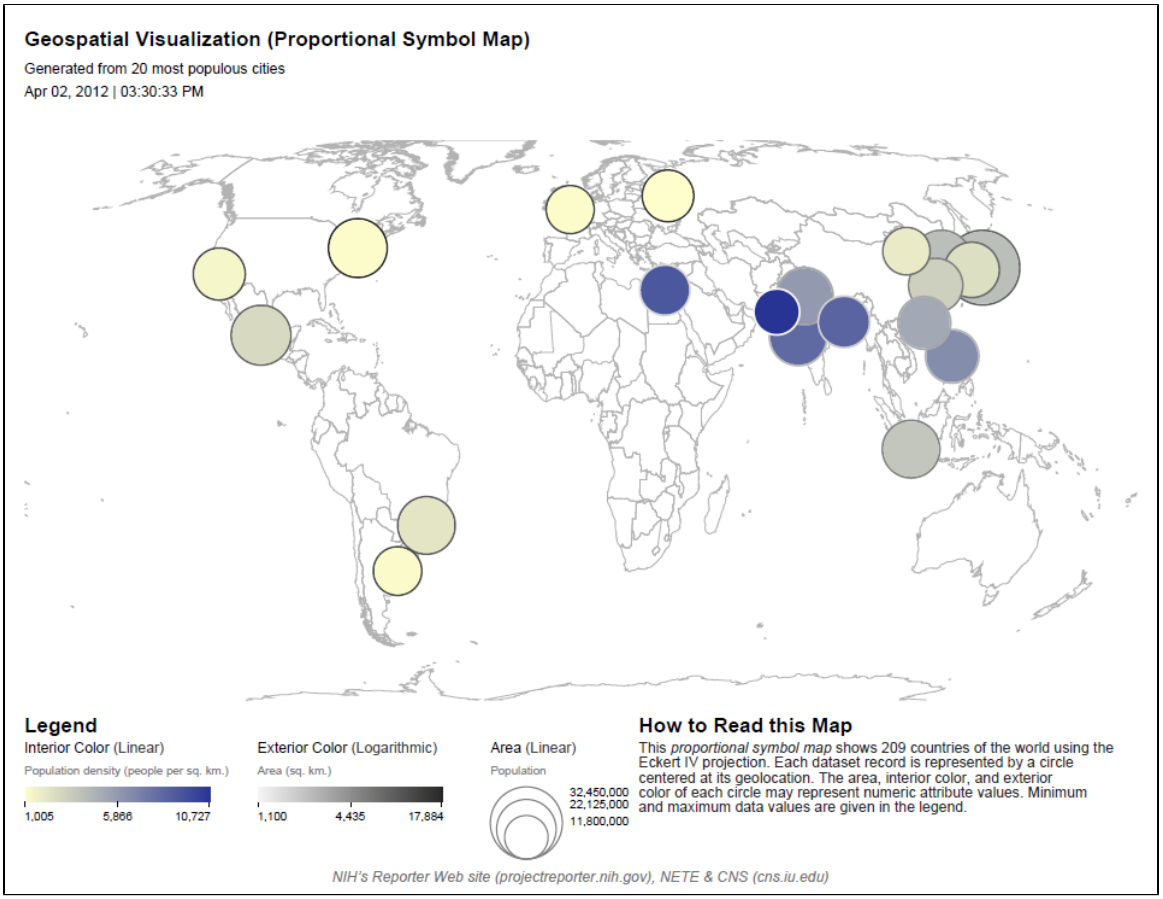
Proportional Symbol Map

Takes a table of geospatial coordinates associated with up to 3 numeric attributes and visualizes them as symbols overlaid on a world or United States base map. The sizes and colors of the symbols are proportional to the associated numeric data.

Features

- Optionally scale each individual dimension of numeric data logarithmically or exponentially.
- Legends for each visualization dimension show how extrema of the data are represented visually.

Examples



Download PDF

Implementation Details

Expects a table with numeric attributes that:

- Describe that datum's longitude in degrees (from -180 to +180).
- Describe that datum's latitude in degrees (from -90 to +90).
- Determine the size of the circle at that coordinate.
- Determine the color of the circle at that coordinate.

Usage Hints

- If you have region names but not latitude and longitude data in your table, you can still easily produce a circle-annotated map using the Geocoder algorithm, which will find the coordinate data for each region name and add it to the table.
- There is no perfect map projection – the choice depends on your application. We recommend:
- Albers equal-area conic if preserving area is important.
- Lambert conformal conic if preserving angles (or shapes) is important.
- Mercator if presenting a familiar and fairly standardized projection is important.
- The color-region Geomap is not suitable if your data that contain small countries. For example, Singapore is too small a geographic region to be visible on the map when the region is colored. We recommend the user select the circle Geomap if data on smaller geographic regions is important to analysis.
- The color-region Geomap requires that each country be identified using its name from the list Country names recognized by Geo Maps.
- The color-region Geomap requires that each U.S. state be identified using its name from the list U.S. state names recognized by Geo Maps.

Links

- Source Code

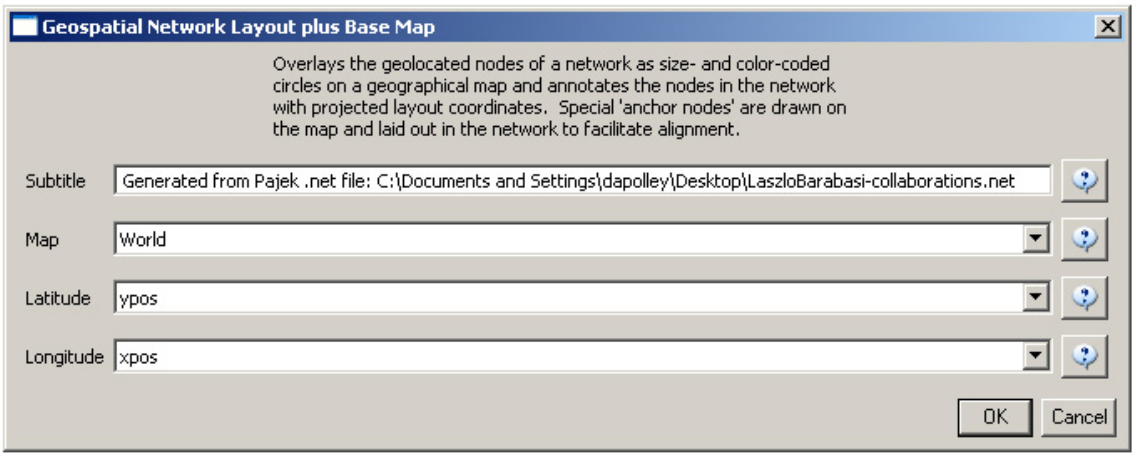
4.7.6 Using Gephi to Render Networks Overlaid on Geo Maps

Loading and Saving Geovisualization Files in Sci2

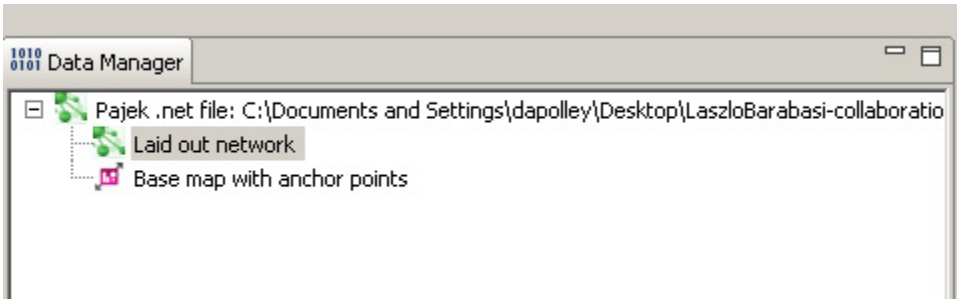
This algorithm allows for the geospatial visualization of network data. The algorithm produces a network file and corresponding blank map. Gephi is used to edit the network produced by Sci2. Once the network has been edited in Gephi it can be exported in a format that will allow it to be overlaid on the map, facilitating visualization of the geospatial data. The following is a brief workflow explaining the process, beginning to end. For this visualization the LaszloBarabasi-collaborations.net file will be used. This network maps Albert-László Barabási and his collaborators.

1. Load the LaszloBarabasi-collaborations.net network in Sci2.

2. Once the network had been loaded in Sci2 run *'Visualization > Geospatial > Geospatial Network Layout with Base Map'* and set the following parameters:



3. One this algorithm has been run the result will be two files in the data manager. A network (Laid out network) and a blank world map (base map with anchor points):



You will want to save both files. First, right click on the map file (top one) and save the file as a "PostScript" and select desired location. Now, right click on the network file (second one) and save the file as GraphML (Prefuse) and select the desired location. Note, you may still need to change the network file type to a .graphml file.

Showing File Extensions

Windows XP: You can have the file extensions shown in the folder where you have saved the network file. In the folder, follow this path: "*Tools > Folder Options*" and select the view tab. Once you have done this you can deselect the "Hide extensions for know file types" box.
Windows 7: Open folder options by following this path: "*Start > Control Panel > Appearance and Personalization > Folder Options*" Now click the view tab. Under the Advanced Settings you can deselect the "Hide extensions for known file types" box.

Once the network file has been saved as a .graphml it can be loaded into Gephi. (If you do not currently have Gephi, it is freely available here and Gephi tutorials have been made available here.)

Manipulating the Network File in Gephi

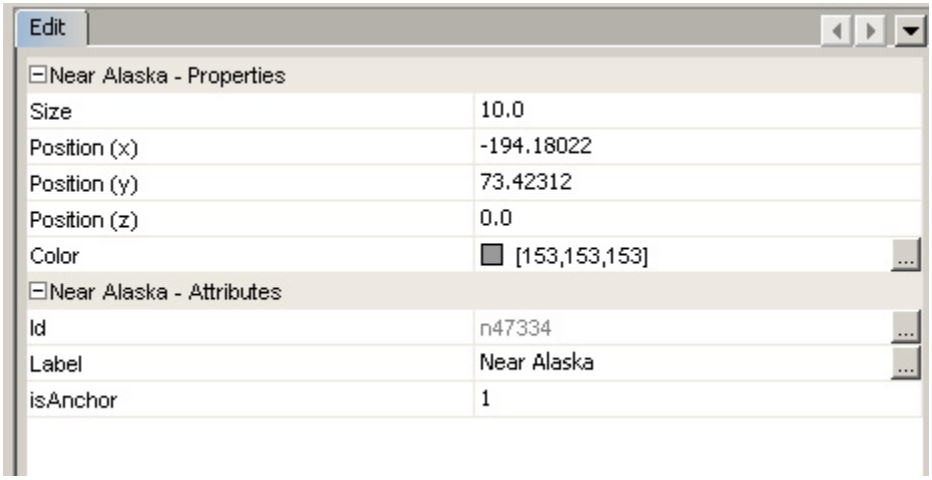
1. When you open Gephi select "Open a graph file" from under the "New Project" heading. Find the folder where you save the network file and load the .graphml version of the network file. The "Important Report" pop-up will display, informing you that they network will be loaded as an undirected network. Click OK.

2. Once the network has been loaded you can view the graph in the "Overview" panel. The network had two nodes that correspond to two dots on the map file. These nodes will help you when you overlay the network on the map. In order to make these nodes more visible in the network file you may want to change their color.

3. Go to the "Data Laboratory" tab in the top left-hand corner of the Gephi tool.

4. Make sure you have the "Nodes" tab selected, you will see a list of all the nodes in the network. By scrolling down to the bottom of the node list you will notice that two nodes are labeled "Near Alaska" and "Near Antarctica" and are both anchor nodes (look at the isAnchor column).

5. Right click on the first anchor node and select the "Edit node" option. This will bring up the Edit function on the left-hand side of the screen:

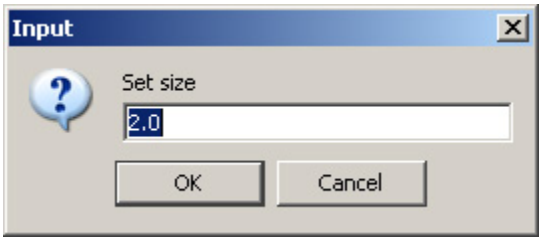


6. You can now adjust the color to make the nodes more visible. Repeat the same process for the node near Antarctica.

7. Return to the "Overview" tab and you will notice that the nodes have been colored based on your specifications.

8. Next you will want to resize the nodes and decrease the edge weight to make the network more visible in the resulting visualization. You can change the size by right-clicking on the size button in the lower left-hand side of the

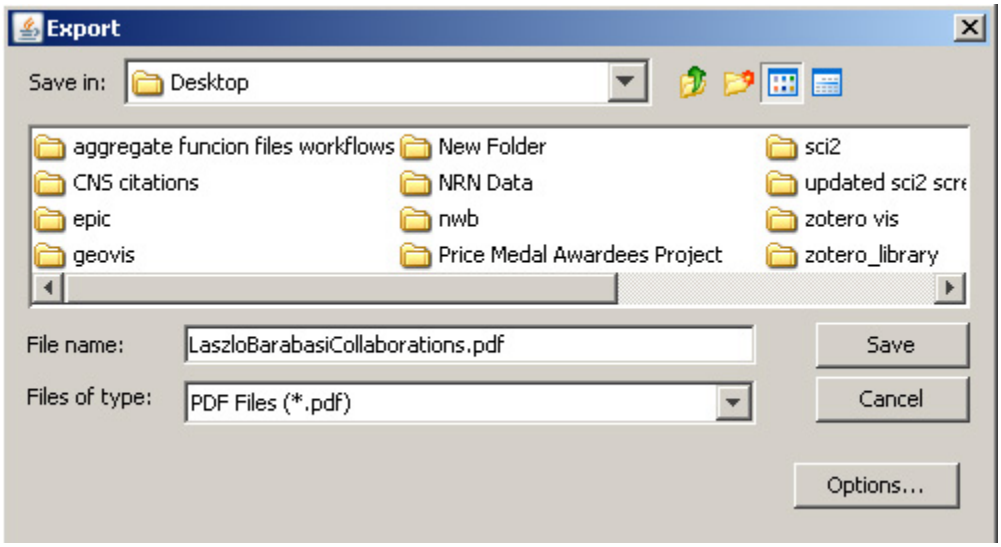
"Graph" screen:



9. Decrease the Edge weight using the slider bar at the bottom of the 'Graph' screen:

Saving the Network in Gephi

1. There are several ways to overlay the network file on the map file. It can be done entirely in Photoshop or it can be done by using a combination of Adobe Illustrator and Photoshop. The easiest way is to first export the network file you have edited with Gephi to a PDF format.



Creating the Visualization in Photoshop

1. Open the map file (blank map generated in Sci2) in photoshop. You may need to rotate the image. This can be done by selecting the layer and clicking *'Edit > Transform > Rotate'.*

2. Next, open the PDF saved from Gephi in Adobe Illustrator. You can delete the path that borders the entire image and use the select arrow to select the entire network. Then click *'Edit > Copy'.*

3. Now, in the map file select *'Edit > Paste''*. You will want to select paste as pixels. This will create a new layer in the map Photoshop file. The network will appear as a new layer on top of the map.

4. Any resizing that needs to be done in order to line up the colored nodes on the network file and their corresponding dots on the map file can be done by selecting the network image layer and using *'Edit > Transform > Scale'*. Tip: hold down shift while doing using the transform and it will be done to scale.

5. You can remove the colored nodes that are used to line up the images by using the eraser tool.

6. When you are finished editing the image you will want to merge both layers prior to saving the file. You can select both layers in the layers window to the right by using Cntrl click. The right click and select "Merge Layers".

7. In order to save the visualization once it has been created in Photoshop go to *'File > Save As'* and then select an appropriate file name and file format, such as JPEG.

The resulting image will look like this:

Creating the Visualization in Gimp

Gimp offers an open source alternative to costly software like Adobe Photoshop. Gimp works really well for the purpose of overlaying network images generated by Sci2 and Gephi on geomaps. For more information or to download the tool see the Gimp

website

.

Here is the process for overlaying the network on the geomap created from the LaszloBarabasi-collaborations.net:

> You will want to follow the same steps exporting the network from Gephi in PDF format, as this will easily open Gimp.

1. Open the map PDF file exported from Sci2 in one Gimp file.

2. Open the network PDF file exported from Gephi in another Gimp file.

3. You will most likely need to crop the network image. You can either use *'Tools > Transform > Crop'* or the crop shortcut in the Toolbox:

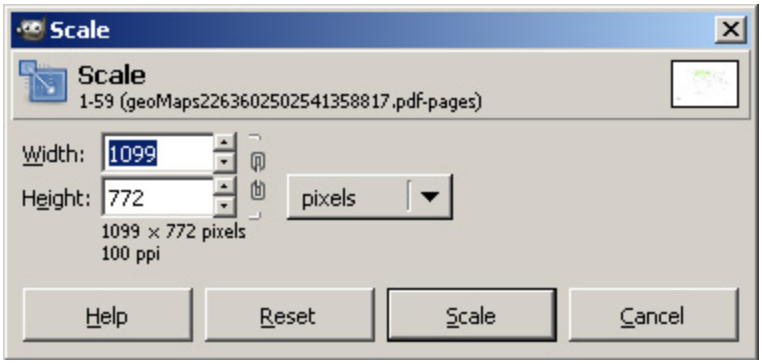
4. Once the image is cropped, copy the image, *'Edit > Copy'.*

5. Paste the image as a new layer in the map file, *'Edit > Paste as > New Layer'*.

6. The new layer (network overlay) will need to be made transparent, *'Layer > Transparency > Color to Alpha...'*. This will make the geomap layer visible under the network overlay.

7 The network overlay will need to be scaled, allowing for the anchor nodes to line up with their corresponding positions on the map (near Alaska and near Antarctica. The easiest way is select the Scale tool from the Toolbox:

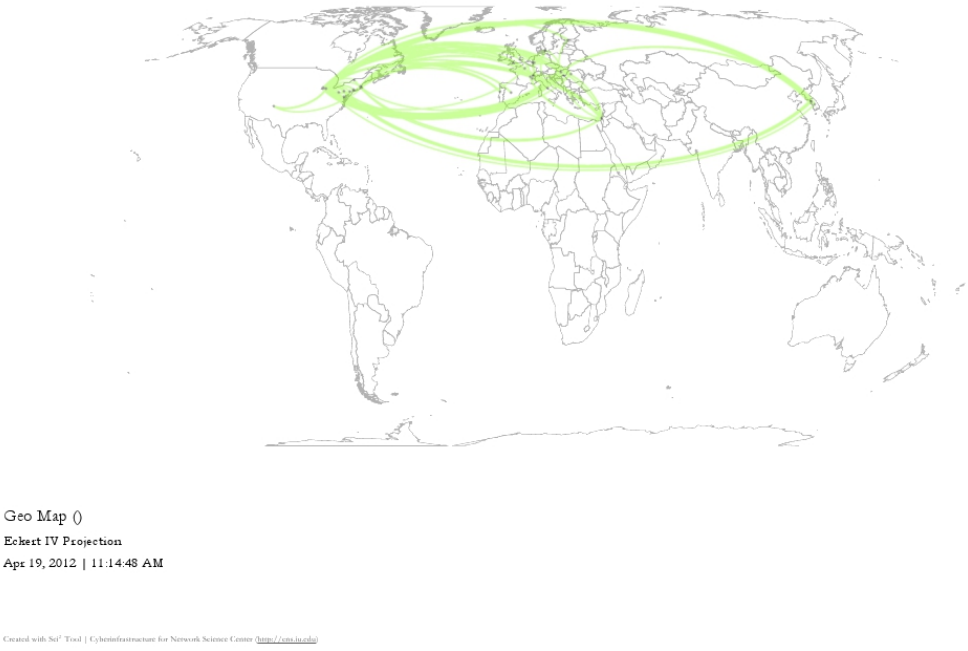
Click on the network overlay layer and the scale tool pop-up will appear. The layer to be scaled will appear with rectangles at each corner which allow users to re-size the layer with the mouse by clicking and dragging. Line up the anchor nodes on the network image with the corresponding points on the map image. Once the desired scaling has been achieved click the "Scale" button on the scaling pop-up:



8. Next you will need to merge the layers, *'Layer > Merge Down'.*

9. Before the image can be saved it will need to be flattened, *'Image > Flatten Image'*.

10. Save the visualization in desired format (recommended file format is .jpg). Below is an example of the LaszloBarabasi-collaborations.net file overlaid on a geomap. Note, the edges were colored in Gephi to make them more visible in the resulting visualization:



4.8 Topical Analysis (What)

The topic or semantic coverage of a unit of science can be derived from the text associated with it. Topical aggregations (e.g., over journal volumes, scientific disciplines, or institutions) are common.
Topical analysis extracts the set of unique words or word profiles and their frequency from a text corpus. Stop words, such as 'the' and 'of' are removed. Stemming can be applied. Co-word analysis identifies the number of times two words are used in the title, keyword set, abstract and/or full text of a paper. The space of co-occurring words can be mapped providing a unique view of the topic coverage of a dataset. Similarly, units of science can be grouped according to the number of words they have in common.
Salton's term frequency inverse document frequency (TFIDF) is a statistical measure used to evaluate the

importance of a word in a corpus. The importance increases proportionally to the number of times a word appears in the paper but is offset by the frequency of the word in the corpus.
Dimensionality reduction techniques are commonly used to project high-dimensional information spaces (i.e., the matrix of all unique papers multiplied by their unique terms, in a low, typically two-dimensional space).

4.8.1 Word Co-Occurrence Network

The topic similarity of basic and aggregate units of science can be calculated via an analysis of the co-occurrence of words in associated texts. Units that share more words in common are assumed to have higher topical overlap and are connected via linkages and/or placed in closer proximity. Word co-occurrence networks are weighted and undirected.

4.9 Network Analysis (With Whom?)

The study of networks aims to increase our understanding of natural and man-made networks. It builds on social network analysis, physics, information science, bibliometrics, scientometrics, econometrics, informetrics, webometrics, communication theory, sociology of science, and several other disciplines.
Authors, institutions, and countries, as well as words, papers, journals, patents, and funding are represented as nodes and their complex interrelations as edges. Nodes and edges can have (time-stamped) attributes.
Figure 4.12 shows a sample dataset of five papers, A through E, published over three years together with their authors' x, y, z, references (blue references are papers outside this set) and citations (green ones go to papers outside this set) as well as some commonly derived networks. The extraction and analysis of these and other scholarly networks is explained subsequently.

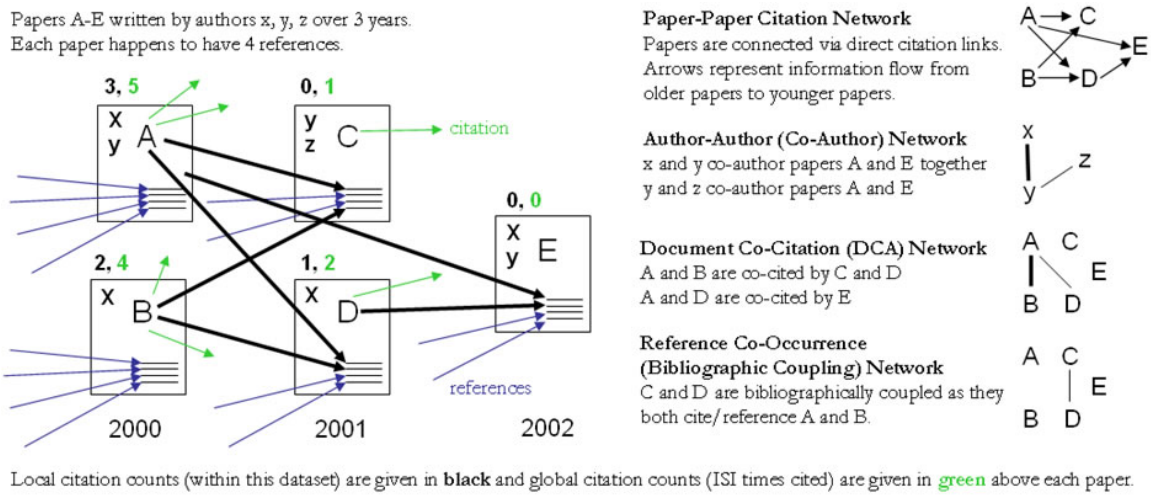


Figure 4.12: Sample paper network (left) and four different network types derived from it (right)

Diverse algorithms exist to calculate specific node, edge, and network properties. Node properties comprise degree centrality, betweenness centrality, or hub and authority scores. Edge properties include durability, reciprocity, intensity (weak or strong), density (how many potential edges in a network actually exist), reachability (how many steps it takes to go from one "end" of a network to another), centrality (whether a network has a "center" point), quality (reliability or certainty), and strength. Network properties refer to the number of nodes and edges, network density, average path length, clustering coefficient, and distributions from which general properties such as small-world, scale-free, or hierarchical can be derived. Identifying major communities via community detection algorithms and calculating the "backbone" of a network via pathfinder network scaling or maximum flow algorithms helps to communicate and make sense of large scale networks.

4.9.1 Network Extraction

4.9.1.1 Direct Linkages

4.9.1.1.1 Document-Document (Citation) Network

Papers cite other papers via references forming an unweighted, directed paper citation graph. It is beneficial to indicate the direction of information flow, in order of publication, via arrows. Such a representation allows for the tracking of citation networks chronologically, yielding a better understanding of the influence of previous research on subsequent research, which more clearly describes the scholarly relationship between individual publications. Citations to a paper support the forward traversal of the graph. Citing and being cited can be seen as roles a paper possesses (Nicolaisen, 2007).

4.9.1.1.2 Author-Author (Citation) Network

Authors cite other authors via document references forming a weighted, directed author citation graph. Like document-document networks, author citation networks represent the flow of information over time. Unlike document citations, however, these networks have weighted edges representing the volume of citations from one author to the next.

4.9.1.1.3 Source-Source (Citation) Network

For larger scale studies, it is often useful to explore citation patterns between entire journals and other varieties of publications. These networks can reveal both the relative importance of certain publications, and the underlying connections between disciplines. These networks are directed and weighted by volume of citations between journals.

4.9.1.1.4 Author-Paper (Consumed/Produced) Network

There are active and passive units of science. Active units, e.g., authors, produce and consume passive units, e.g., papers, patents, datasets, software. The resulting networks have multiple types of nodes, e.g., authors and papers. Directed edges indicate the flow of resources from sources to sinks, e.g., from an author to a written/produced paper to the author who reads/consumes the paper.

4.9.1.2 Co-Occurrence Linkages

4.9.1.2.1 Author Co-Occurrence (Co-Author) Network

Having the names of two authors (or their institutions, countries) listed on one paper, patent, or grant is an empirical manifestation of scholarly collaboration. The more often two authors collaborate, the higher the weight of their joint co-author link. Weighted, undirected co-authorship networks appear to have a high correlation with social networks that are themselves impacted by geospatial proximity (Börner, Penumarthy, Meiss, & Ke, 2006; Wellman, White, & Nazer, 2004).

4.9.1.2.2 Document Cited Reference Co-Occurrence (Bibliographic Coupling) Network

Papers, patents or other scholarly records that share common references are said to be coupled bibliographically (Kessler, 1963),

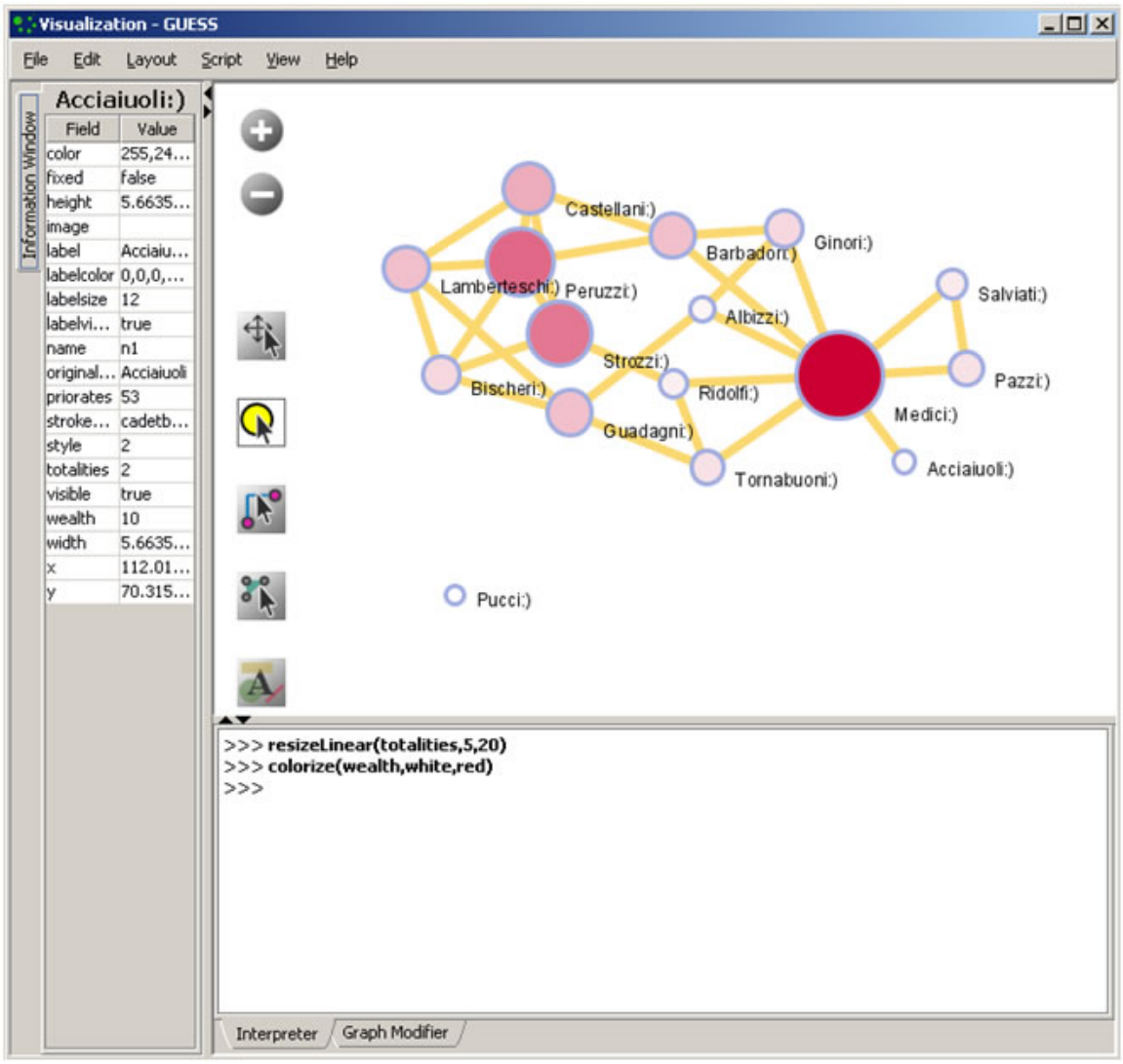


Figure 4.13: An example of a bibliographic coupling network

The bibliographic coupling (BC) strength of two scholarly papers can be calculated by counting the number of times that they reference the same third work in their bibliographies. The coupling strength is assumed to reflect topic similarity. Co-occurrence networks are undirected and weighted.

4.9.1.2.3 Author Cited Reference Co-Occurrence (Bibliographic Coupling) Network

Authors who cite the same sources are coupled bibliographically. The bibliographic coupling (BC) strength between two authors can be said to be a measure of similarity between them. The resulting network is weighted and undirected.

4.9.1.2.4 Journal Cited Reference Co-Occurrence (Bibliographic Coupling) Network

Like document and author bibliographic coupling networks, journal cited reference co-occurrences provide a measurement of similarity between journals. Edge strength between two journals is determined by summing the number of unique references both journals cite.

4.9.1.3 Co-Citation Linkages

Two scholarly records are said to be *co-cited* if they jointly appear in the list of references of a third paper. The more often two units are co-cited the higher their presumed similarity.

4.9.1.3.1 Document Co-Citation Network (DCA)

DCA was simultaneously and independently introduced by Small and Marshakova in 1973 (Marshakova, 1973; Small, 1973; Small & Greenlee, 1986). It is the logical opposite of bibliographic coupling. The co-citation frequency equals the number of times two papers are cited together, i.e., they appear together in one reference list.

4.9.1.3.2 Author Co-Citation Network (ACA)

Authors of works that repeatedly appear together in lists of references are assumed to be related. Clusters in ACA networks often reveal shared schools of thought or methodology, common subjects of study, collaborative and student-mentor relationships, ties of nationality, etc. Some regions of scholarship are densely crowded and interactive. Others are isolated or nearly vacant.

4.9.1.3.3 Journal Co-Citation Network (JCA)

JCA networks offer wide-angle views of scholarly disciplines. Slicing these networks by time can reveal the evolution of disciplinary similarity. Like author and document co-citation networks, these are undirected and weighted.

4.9.2 Compute Basic Network Characteristics

It is often advantageous to know for a network

- Whether it is directed or undirected
- Number of nodes
- Number of isolated nodes
- A list of node attributes
- Number of edges
- Whether the network has self loops, if so, lists all self loops
- Whether the network has parallel edges, if so, lists all parallel edges
- A list of edge attributes
- Average degree
- Whether the graph is weakly connected

- Number of weakly connected components
- Number of nodes in the largest connected component
- Strong connectedness for directed networks
- Graph density

In the Sci^2 Tool, use *'Analysis >* *Network Analysis Toolkit (NAT)'* to get basic properties, e.g., for the network of Florentine families available in '*yoursci2directory/sampledata/socialscience/florentine.nwb*'. The result for this dataset is:

> This graph claims to be undirected.
> Nodes: 16
> Isolated nodes: 1
> Node attributes present: label, wealth, totalities, priorates
> Edges: 27
> No self loops were discovered.
> No parallel edges were discovered.
> Edge attributes:
> Nonnumeric attributes:
> Example value
> marriag[tab] T
> busines[tab] F
> Did not detect any numeric attributes
> This network does not seem to be a valued network.
> Average degree: 3.375
> This graph is not weakly connected.
> There are 2 weakly connected components. (1 isolates)
> The largest connected component consists of 15 nodes.
> Did not calculate strong connectedness because this graph was not directed.
> Density (disregarding weights): 0.225

4.9.3 Network Analysis

In the analysis menu, certain algorithms append values to each node, or delete groups of nodes and edges entirely:

- Weak Component Clustering extracts the N largest weakly connected components of a network
- Node Degree calculates the number of edges adjacent to a node
- Pathfinder Network Scaling prunes a network to find its underlying structure
- Node Betweenness Centrality appends a value to each node which correlates to the number of shortest paths that node resides on. The more "shortest paths" between node-pairs a certain node resides on, the higher its betweenness centrality. To learn about each algorithm, see section 3.1 'Sci2 Tool Plugins' or visit https://nwb.cns.iu.edu/community/ for greater detail.

4.9.4 Network Visualization

4.9.4.1 GUESS Visualizations

Load the sample dataset '*yoursci2directory/sampledata/socialscience/florentine.nwb'* and calculate an additional node attribute 'Betweenness Centrality' by running *'Analysis > Networks > Unweighted and Undirected > Node Betweenness Centrality'* with default parameters. Then select the network and run *'Visualization > Networks > GUESS'* to open GUESS with the file loaded. It might take some time for the network to load. The initial layout will be random. Wait until the random layout is completed and the network is centered before proceeding.
The GUESS window is divided into three sections:

1. Information - Examine node and edge attributes, see Figure 4.14, left

2. Visualization - View and manipulate network, see Figure 4.14, top right.
3. Interpreter/Graph Modifier - Analyze/change network properties, see figure 4.14, bottom right.

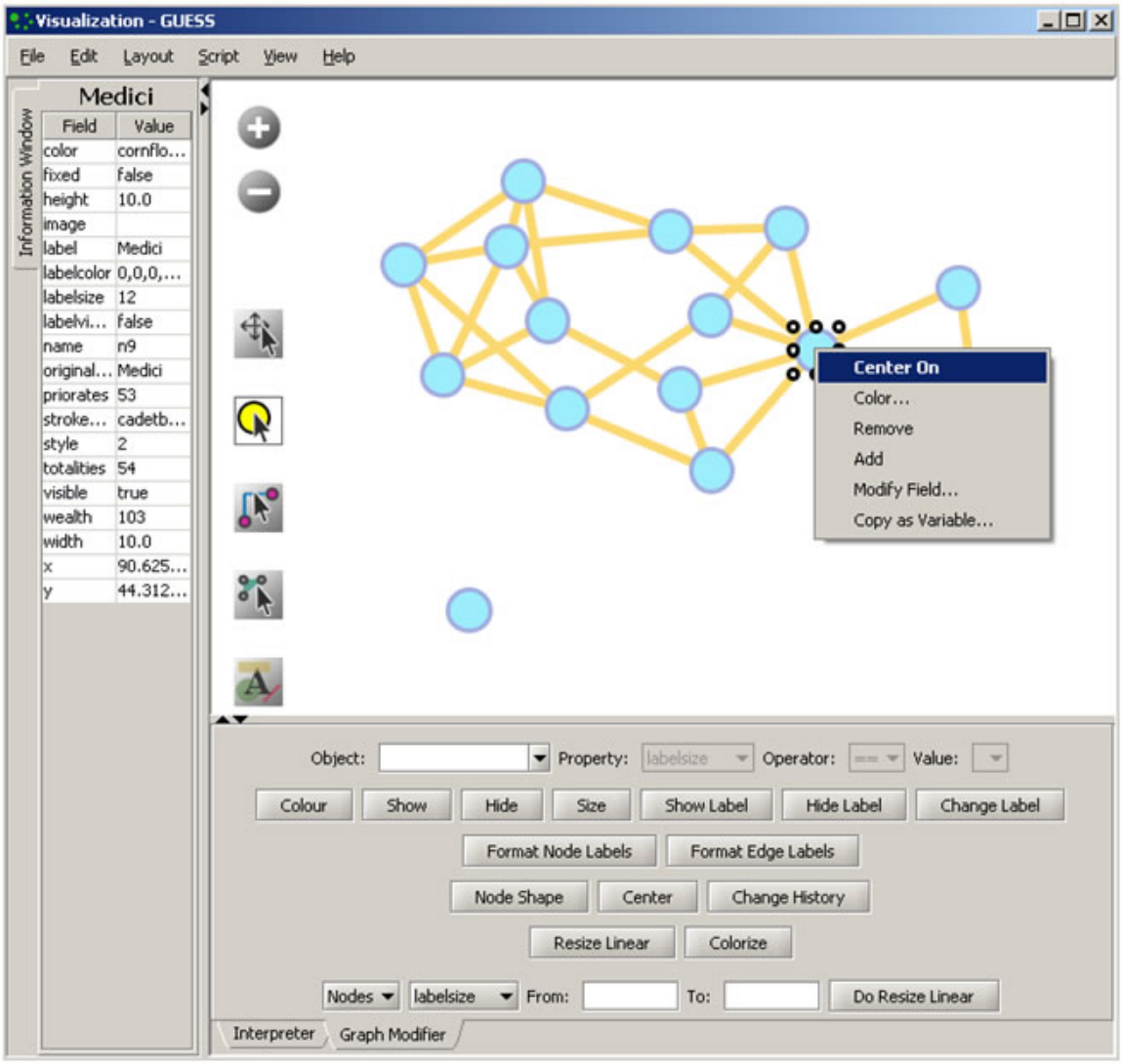


Figure 4.14: GUESS 'Information Window', visualization window, and 'Graph Modifier' window

4.9.4.1.1 Network Layout and Interaction

GUESS provides different network layout algorithms under menu item 'Layout'. Apply *'Layout > GEM'* to the Florentine network. Use *'Layout > Bin Pack'* to compact and center the network layout. These layout algorithms often employ some degree of randomness, and layouts may look different every time they are used. Also note that running GEM and/or Bin Pack on the same network multiple times will continue to change the visualization each and every time they are used.
Using the mouse pointer, hover over a node or edge to see its properties in the Information window. GUESS has several methods of interaction:

- Pan – simply 'grab' the background by clicking and holding down the left mouse button, and move it using the mouse.
- Zoom – Using the scroll wheel on the mouse OR press the "+" and "-" buttons in the upper-left hand corner of the visualization window OR right-click and move the mouse left or right. Center graph by selecting '*View > Center*'.
- Click to select/move single nodes. Hold down 'Shift' to select multiple.
- Right click: Right clicking on a node gives the options to '*Center on*', '*Color*', '*Toggle Label*', '*Remove*', '*Add*', '

Modify Field', and '*Copy as Variable*', see Figure 4.14.

Use the Graph Modifier to change node attributes, e.g.,

- Select '*all nodes*' in the Object drop-down menu and click 'Show Label' button.
- Select *'Resize Linear'* and choose *'Nodes'* and *'totalities'* from the drop-down menus. Enter the parameters 'From: 5 To: 20.' Then select '*Do Resize Linear*.'
- Select *'Colorize'* and choose *'Nodes'* and *' totalities*' from the drop-down menus. Click the small box next to '*Fr om*' and in the pop-up '*Choose the First Color*' window, under the '*Swatches*' tab, select white, and set the color options under the '*RGB*' tab to: (204,0,51) Click '*OK*' and then click '*Do Colorize*' In the Graph Modifier.
- To view the entities associated with each node, click the '*Show Label*' button. To modify node labels, select '*F ormat Node Labels*,' and replace the default text with your own label in the pop-up box . The labels shown in Figure 4.15 have been modified to '{originallabel and '.

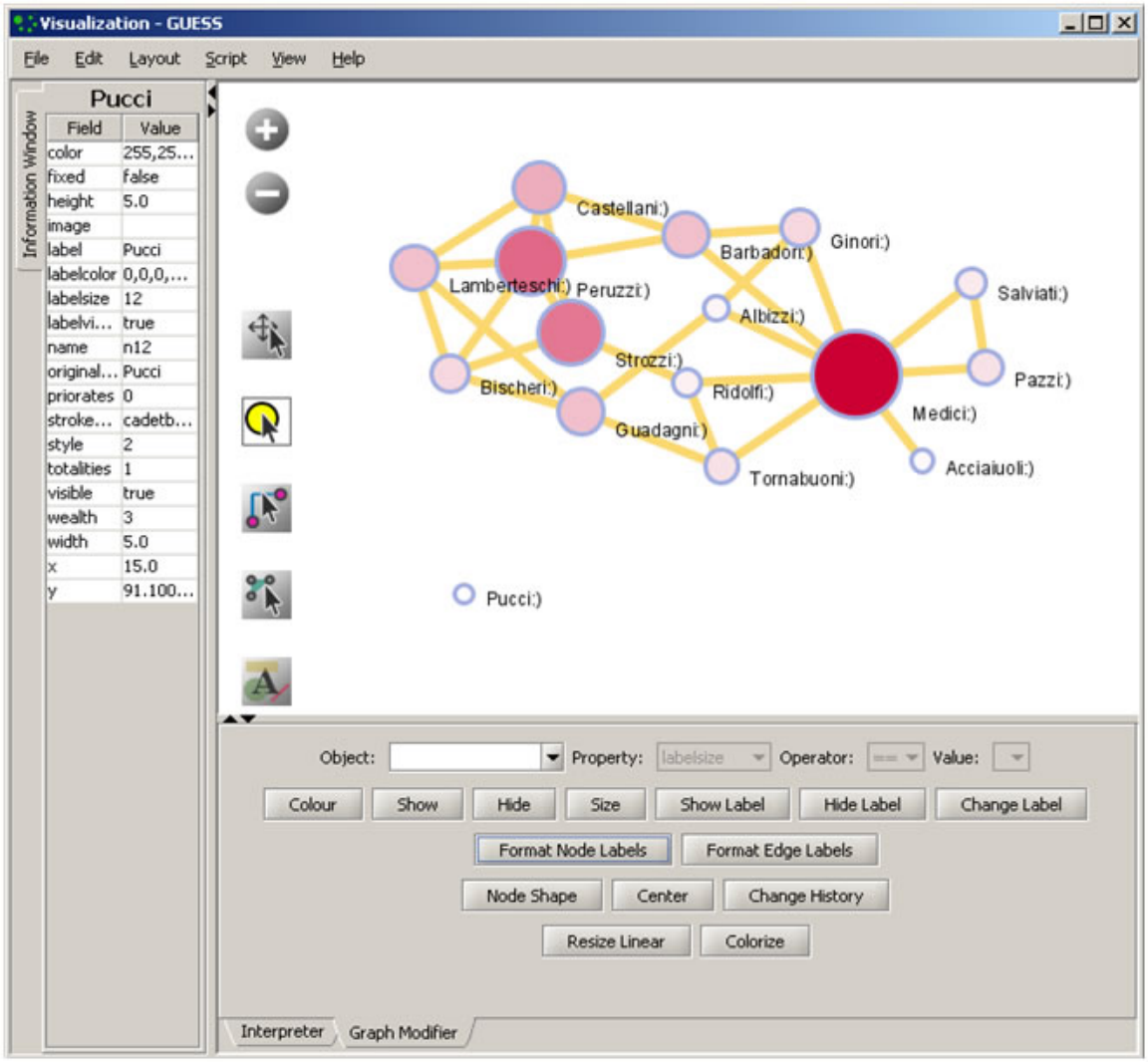


Figure 4.15: Using the GUESS 'Graph Modifier'

4.9.4.1.2 Interpreter

Use Jython, a version of Python that runs on the Java Virtual Machine, to write code in the interpreter. Here we list some GUESS commands which can be used to modify the layout.
Color all nodes *uniformly*

```
g.nodes.color =red
```

```
    g.nodes.strokecolor =red
    g.nodes.labelcolor =red
    colorize(numberofworks,gray,black)
    for n in g.nodes:
        n.strokecolor = n.color

Size code nodes
    g.nodes.size = 30
    resizeLinear(numberofworks,.25,8)

Label
    for n in g.nodes:
n.labelvisible = true
Print
for n in g.nodes:
print n.label + ":" + str(n.indegree)
Edges
g.edges.width=10
g.edges.color=gray
Color and resize nodes based on their betweenness:
colorize(wealth,white,red)
resizeLinear(sitebetweenness,5,20)
```

The result is shown in Figure 4.16. Read https://nwb.cns.iu.edu/community/?n=VisualizeData.GUESS for more information on how to use the interpreter.

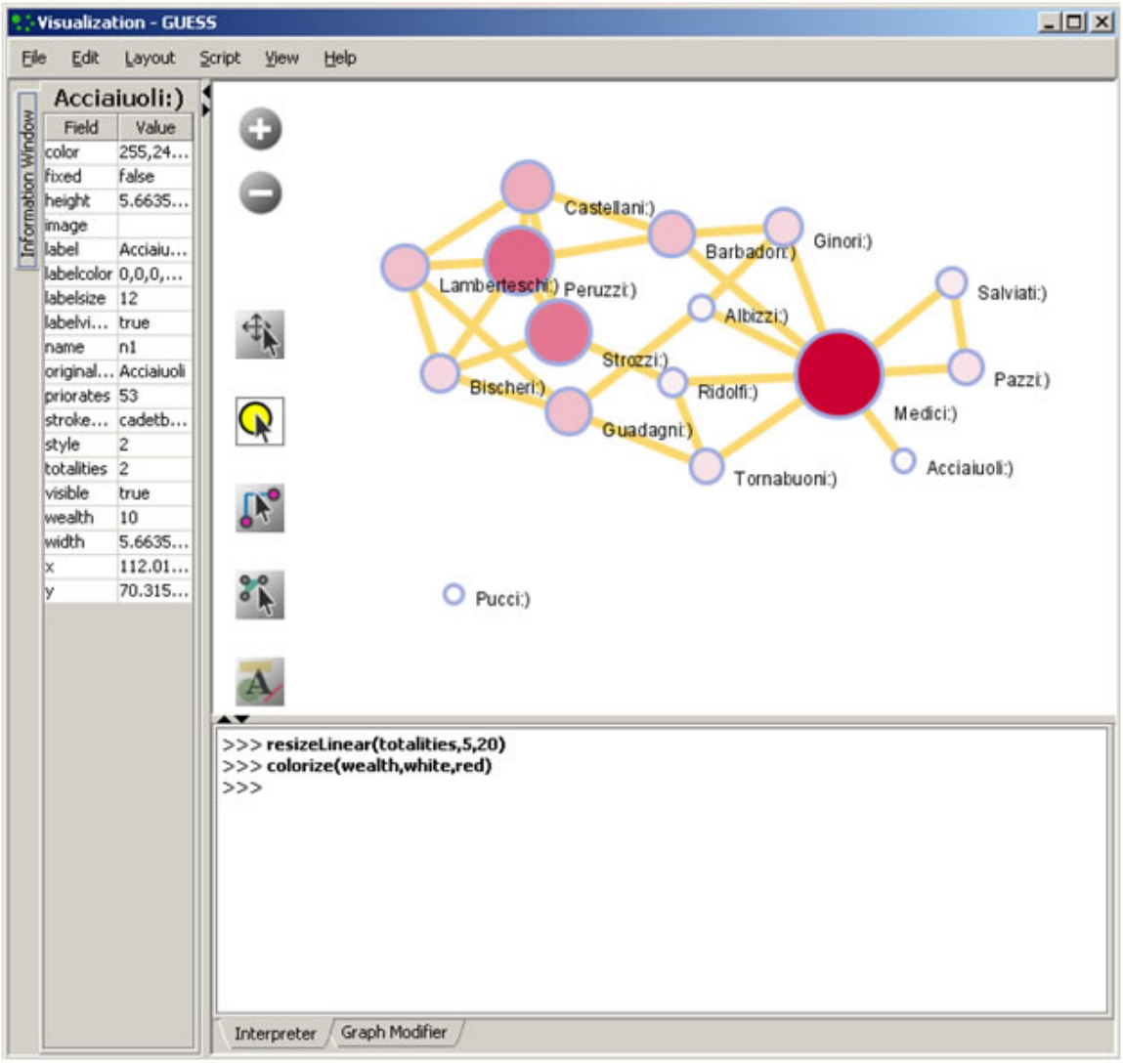


Figure 4.16: Using the GUESS 'Interpreter'

4.9.4.2 DrL Large Network Layout

DrL is a force-directed graph layout toolbox for real-world large-scale graphs up to 2 million nodes (Davidson, Wylie, & Boyack, 2001; Martin, Brown, & Boyack,unpublished). It includes:

- Standard force-directed layout of graphs using an algorithm based on the popular VxOrd routine (used in the VxInsight program).
- Parallel version of force-directed layout algorithm.
- Recursive multilevel version for obtaining better layouts of very large graphs.
- Ability to add new vertices to a previously drawn graph.

This is one of the few force-directed layout algorithms that can scale to over 1 million nodes, making it ideal for large graphs. However, the algorithm doesn't always render small graphs (less than a hundred records) well. The algorithm expects similarity matrices as input. Distance and other networks will have to be converted before they can be laid out. For article/citation networks, feed the network into either cocitation or bibliographic coupling for computing similarity. Use this network for laying out DrL.
The version of DrL included in Sci^2 only does the standard force-directed layout (no recursive or parallel computation). DrL expects the edges to be weighted, directed edges where the weight (greater than zero) denotes how similar the two nodes are (higher is more similar). The Sci^2 version has several parameters. The edge cutting parameter expresses how much automatic edge cutting should be done. 0 means as little as possible, 1 as much as

possible. Around .8 is a good value to use. The weight attribute parameter lets users choose which edge attribute in the network corresponds to the similarity weight. The X and Y parameters let users choose the attribute names in the returned network which correspond to the X and Y coordinates computed by the layout algorithm for the nodes. DrL is commonly used to lay out large networks such as co-citation and co-word analyses. In the Sci^2 Tool, the results can be viewed in either GUESS or *'Visualization > Specified (prefuse alpha)'*. For more information see https://nwb.cns.iu.edu/community/?n=VisualizeData.DrL.
For an application of DrL, see 5.1.4.5 Word Co-Occurrence Network.

4.10 Modeling (Why?)

Data models are grouped into two major types: descriptive models and process models. *Descriptive models* aim to illustrate the major features of a (typically static) data set, such as statistical patterns of article citation counts, networks of citations, individual differences in citation practice, the composition of knowledge domains, or the identification of research fronts as indicated by new yet highly cited papers. *Process models* or predictive models aim to simulate, statistically describe, or formally reproduce statistical and dynamic characteristics of interest.

4.10.1 Random Graph Model

The random graph model generates a graph that has a fixed number of nodes which are connected randomly by undirected edges, see Figure 4.17 (left). The number of edges depends on a specified probability. The edge probability is chosen based on the number of nodes in the graph. The model most commonly used for this purpose was introduced by Gilbert (Gilbert, 1959). This is known as the *G(n,p)* model with *n* being the number of vertices and *p* the linking probability. The number of edges created according to this model is not known in advance. Erd?s-Rényi introduced a similar model where all the graphs with *m* edges are equally probable and *m* varies between *0* and *n(n-1)/2* (Erd?s & Rényi, 1959). This is known as the *G(n,m)* model. The degree distribution for this network is Poissonian, see Figure 4.17 (right).

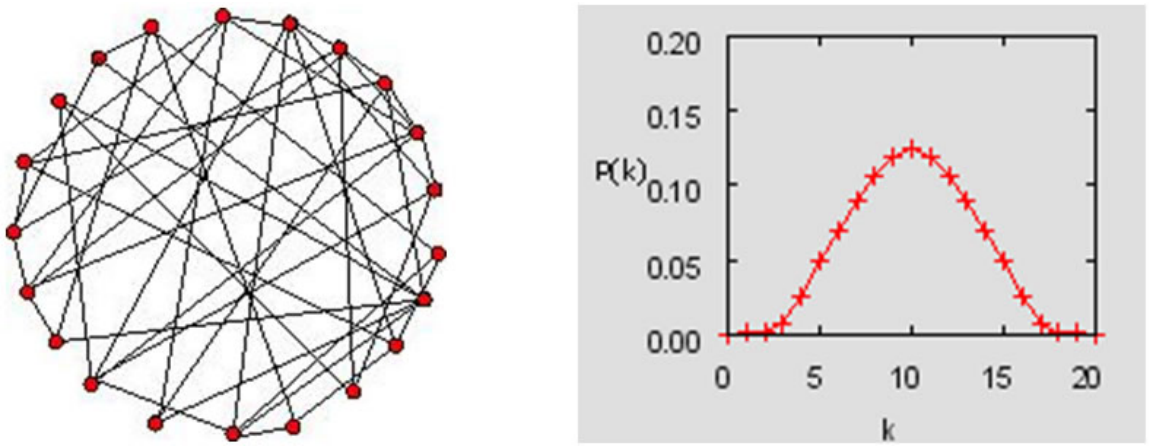
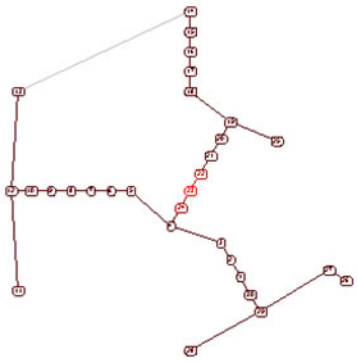


Figure 4.17: Random graph and its Poissonian node degree distribution

Very few real world networks are random. However, random networks are a theoretical construct that is well understood and their properties can be exactly solved. They are commonly used as a reference, e.g., in tests of network robustness and epidemic spreading (Batagelj & Brandes).
In the Sci^2 Tool, the random graph generator implements the *G(n,p)* model by Gilbert. Run *'Modeling > Random Graph'* and input the total number of nodes in the network and their wiring probability. The output is a network in which each pair of nodes is connected by an undirected edge with the probability specified in the input.
A wiring probability of 0 would generate a network without any edges and a wiring probability of 1 with *n* nodes will generate a network with *(n-1)* edges. The wiring probability should be chosen dependent on the number of vertices. For a large number of vertices the wiring probability should be smaller.

4.10.2 Watts-Strogatz Small World

A small-world network is one whose majority of nodes are not directly connected to one another, but still can reach any other node via very few edges. It can be used to generate networks of any size. The degree distribution is almost Poissonian for any value of the rewiring probability (except in the extreme case of rewiring probability zero, for which all nodes have equal degree). The clustering coefficient is high until beta is close to 1, and as beta approaches one, the distribution becomes Poissonian. This is because the graph becomes increasingly similar to an Erd?s-Rényi Random Graph, see Figure 4.18. (Watts & Strogatz, 1998; Inc. Wikimedia Foundation, 2009).



$$P(k) = \sum_{n=0}^{f(k,K)} C_{K/2}^{n} (1-\beta)^{n} \beta^{K/2-n} \frac{(\beta K/2)^{k-K/2-n}}{(k-K/2-n)!} e^{-\beta K/2}$$

Figure 4.18: Small world graph (left) and its node degree distribution equation (right)

Small world properties are usually studied to explore networks with tunable values for the average shortest path between pairs of nodes and a high clustering coefficient. Networks with small values for the average shortest path and large values for the clustering coefficient can be used to simulate social networks, unlike ER random graphs, which have small average shortest path lengths, but low clustering coefficients.
The algorithm requires four inputs: the number *n* of nodes of the network, the number *k* of initial neighbors of each node (the initial configuration is a ring of nodes), the probability of rewiring the edges (which is a real number between 0 and 1), and the seed of a random number generator. The network is built following the original prescription of Watts and Strogatz, i.e., by starting from a ring of nodes each connected to the k nodes and by rewiring each edge with the specified probability. The algorithm run time is *O(kn).*
Run with *'Modeling > Watts-Strogatz Small World'* and input 1000 nodes, 10 initial neighbors, and a rewiring probability of 0.01 then compute the average shortest path and the clustering coefficient and verify that the former is small and the latter is relatively large.

4.10.3 Barabási-Albert Scale Free Model

The Barabási-Albert (BA) model is an algorithm which generates a scale-free network by incorporating growth and preferential attachment. Starting with an initial network of a few nodes, a new node is added at each time step. Older nodes with a higher degree have a higher probability of attracting edges from new nodes. The probability of attachment is given by

$$P(k_i) = \frac{k_i}{\sum_j k_j}$$

.

The initial number of nodes in the network must be greater than two and each of these nodes must have at least one connection. The final structure of the network does not depend on the initial number of nodes in the network. The degree distribution of the generated network is a power law with a scaling coefficient of *-3* (Barabási & Albert, 1999; Barabási & Albert, 2002). Figure 4.19 shows the network on the left and the probability distribution on a log-log scale on the right.

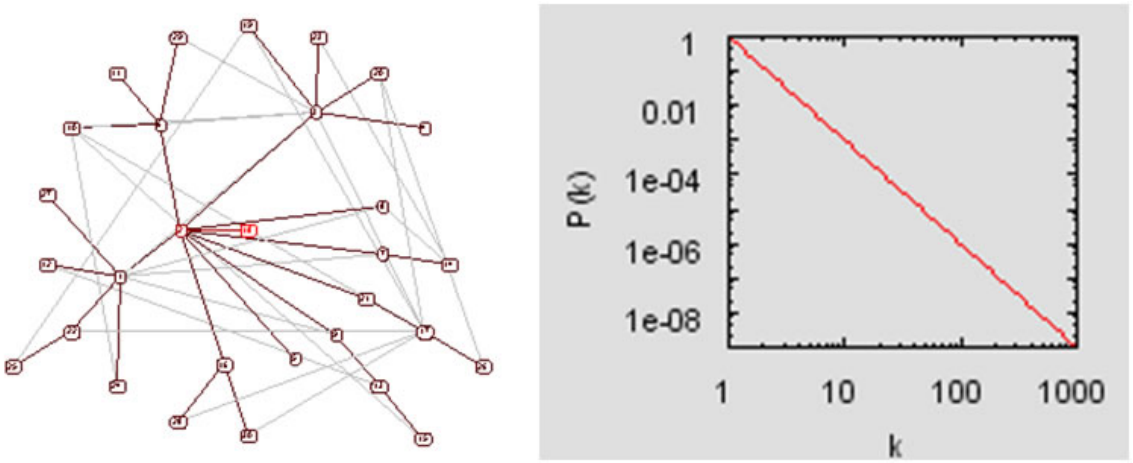


Figure 4.19: Scale free graph (left) and its node degree distribution (right)

This is the simplest known algorithm to generate a scale-free network. It can be applied to model undirected networks such as the collaboration network among scientists, the movie actor network, and other social networks where the connections between the nodes are undirected. However, it cannot be used to generate a directed network.

The inputs for the algorithm are the number of initial nodes, the number of initial edges for a new node, and the seed of a random number generator. The algorithm starts with the initial number of nodes that are fully connected. At each time step, a new node is generated with the initial number of edges. The probability of attaching to an existing node is calculated by dividing the degree of an existing node by the total number of edges. If this probability is greater than zero and greater than the random number obtained from a random number generator then an edge is attached between the two nodes. This is repeated in each time step.

Run with *'Modeling > Barabási-Albert Scale Free Model'* and a time step of around *1000*, initial number of nodes *2*, and number of edges *1* in the input. Lay out and determine the number and degree of highly connected nodes via *'Analysis > Unweighted and Undirected > Degree Distribution'* using the default value. Plot node degree distribution using Gnuplot.

5 Sample Workflows

- 5.1 Individual Level Studies - Micro
- 5.2 Institution Level Studies - Meso
- 5.3 Global Level Studies - Macro

Scientometric studies cover a wide array of datasets, methodologies, and results. Analysis can lead to several types of insights, as a result of asking and answering the questions "what?", "where?", "when?", and "with whom?" (topical, geospatial, temporal, and network analysis, respectively). Many studies also cover statistical surveys of scientometric datasets. For detailed descriptions of these types of scientometric analyses, see sections 4.5 Statistical Analysis through 4.9 Network Analysis.

Analysis can be performed on three major scales: micro/individual, meso/local, and macro/global. The Sci^2 Tool supports workflows in all fifteen varieties of scientometric studies, as well as combination and modeling studies. The following chapter describes the workflows used to conduct scientometric studies of each type and at each scale. Tables 5.1 and 5.2 show examples of studies in each category, several of which can be found in section 6 Sample Science Studies & Online Services.

Analysis Types and Sample Studies	Micro/Individual (1-100 records)	Meso/Local (101--10,000 records)	Macro/Global (10,000 < records)

Statistical Analysis/Profiling	Individual persons and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of US, all of science.
Temporal Analysis (When)	Funding portfolio of one individual	Mapping topic bursts in 20-years of PNAS	113 years of physics research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a state's intellectual landscape	PNAS publications
Topical Analysis (What)	Base knowledge from which one grant draws	Knowledge flows in Chemistry research	Topic maps of NIH funding
Network Analysis (With Whom?)	NSF Co-PI network of one individual	Co-author network	NSF's core competency

Table 5.1: Major analysis types and levels of analysis.

	Micro/Individual (1-100 records)	*Meso/Local (101–10,000 records)*	*Macro/Global (10,000 < records)*
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NS SA, all of sci
Temporal Analysis (When)	Funding portfolio of one individual	ic bursts f PNAS	113 Years of P Research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a s intellectual la	PNAS
Topical Analysis (What)	s.	research	VxOrd/Topic r NIH funding
Network Analysis (With Whom?)	NSI vork of one	k	NIH's cy

Table 5.2: Screenshots of major analysis types and levels of analysis.

5.1 Individual Level Studies - Micro

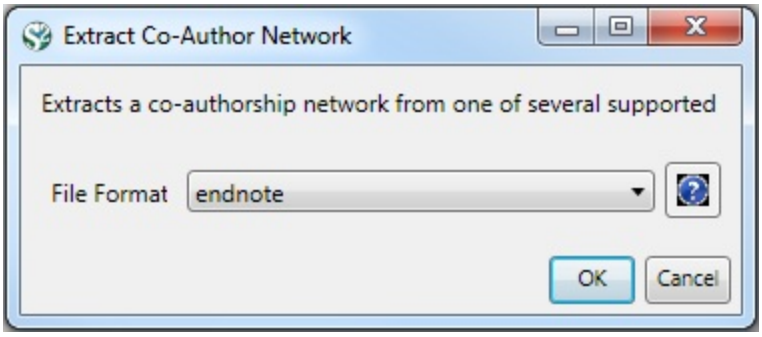
- 5.1.1 Mapping Collaboration, Publication, and Funding Profiles of One Researcher (EndNote and NSF Data)
  - 5.1.1.1 Endnote
    - 5.1.1.1.1 Using Cytoscape to Visualize Networks
  - 5.1.1.2 NSF
- 5.1.2 Time Slicing of Co-Authorship Networks (ISI Data)
  - New ISI File Format
  - Sci2 solution

5.1.1 Mapping Collaboration, Publication, and Funding Profiles of One Researcher (EndNote and NSF Data)

5.1.1.1 Endnote

KatyBorner.enw	
Time frame:	1992-2010
Region(s):	Indiana University, University of Technology in Leipzig, University of Freiburg, University of Bielefeld
Topical Area(s):	Network Science, Library and Information Science, Informatics and Computing, Statistics, Cyberinfrastructure, Information Visualization, Cognitive Science, Biocomplexity
Analysis Type(s):	Co-Authorship Network

Many researchers, tools, and online services use EndNote to organize their bibliographies. To analyze an individual researcher's collaboration and publication profile, load an EndNote file which includes the researcher's entire CV into the Sci^2 Tool. To generate a research profile for Katy Börner, load Katy Börner's EndNote CV at '***yoursci2directory****/sampledata/scientometrics/endnote/KatyBorner.enw*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Then run *'Data Preparation* > *Extract Co-Author Network'* using the parameter:



After generating Dr. Börner's co-authorship network, run *'Analysis > Networks > Unweighted & Undirected > Node Degree'* to append degree information to each node. To visualize the network, run *'Visualization > Networks > GUESS'* and select '*GEM*' in the Layout menu once the graph is fully loaded.. The resulting network in Figure 5.1 was modified using the following workflow:

1. Resize Linear > Nodes > totaldegree > From: 5 To: 30 > Do Resize Linear (Note: total degree is the number of papers)
2. Resize Linear > Edges > weight From: 1 To: 10 > Do Resize Linear (Note: weight is the number of co-authored papers)
3. Colorize > Nodes > totaldegree From : To: > Do Colorize
4. Colorize > Edges > weight From: To: > Do Colorize
5. Object: nodes based on -> > Property: totaldegree > Operator: >= > Value: 10 > Show Label
6. Type in Interpreter:

```
>for n in g.nodes:
     n.strokecolor = n.color
```

GUESS Graph Modifier

Make sure to maximize the window displaying the GUESS visualization tool in order to see all the options in the Graph Modifier.

Note: The Interpreter tab will have '>>>' as a prompt for these commands. It is not necessary to type '>" at the beginning of the line. You should type each line individually and press "Enter" to submit the commands to the Interpreter. For more information, refer to the GUESS tutorial at http://nwb.cns.iu.edu/Docs/GettingStartedGUESSNWB.pdf. The largest cluster in the network is outlined in black, and represents one single paper with many authors.

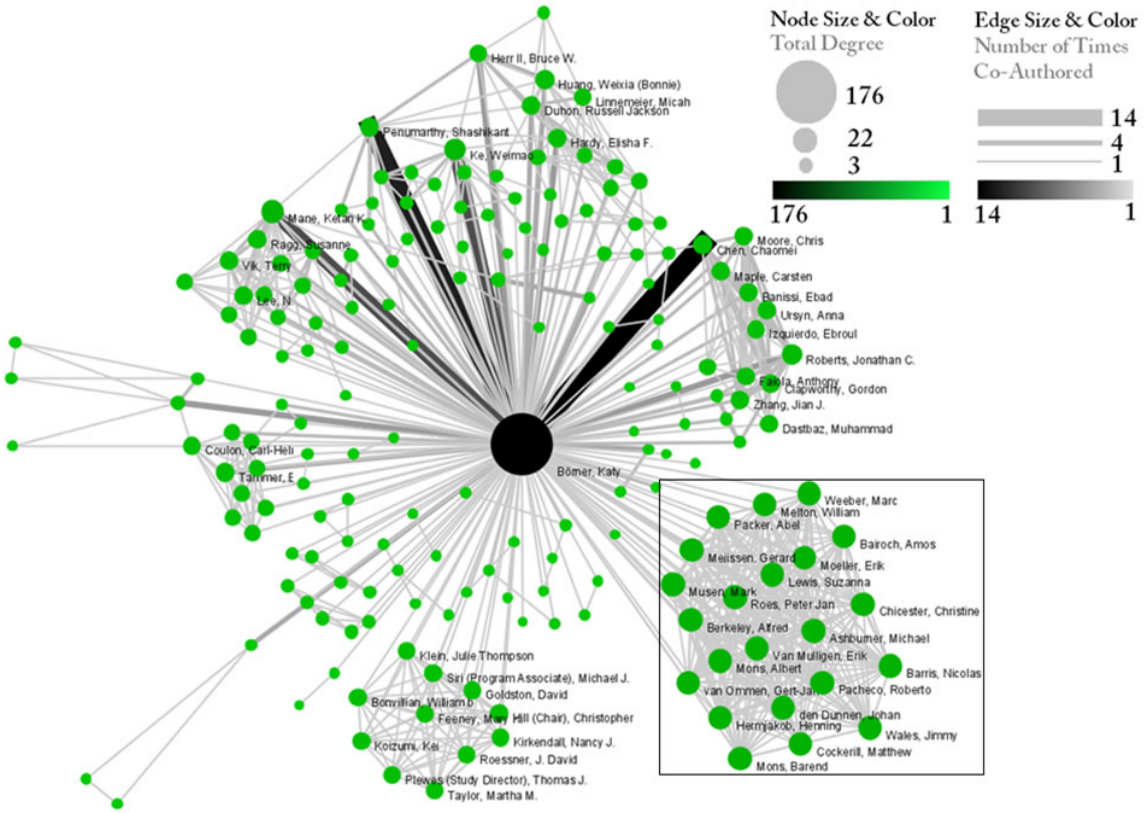


Figure 5.1: Co-authorship network of Katy Börner
GUESS supports the repositioning of selected nodes. Multiple nodes can be selected by holding down the 'Shift' key and dragging a box around specific nodes. The final network can be saved via *'GUESS: File > Export Image'* and opened in a graphic design program to add a title and legend. The image above was created using Adobe Photoshop. Node clusters were highlighted and increased in size, the label font size was increased for emphasis, and a legend was added to clarify the significance of node and edge size and color.

To see the log file from this workflow save the 5.1.1.1 Endnote log file.

5.1.1.1.1 Using Cytoscape to Visualize Networks

> **Extended Version**
>
> This workflow uses the extended version of the Sci2 Tool. To know how to extend Sci2 view Section 3.2 Additional Plugins.

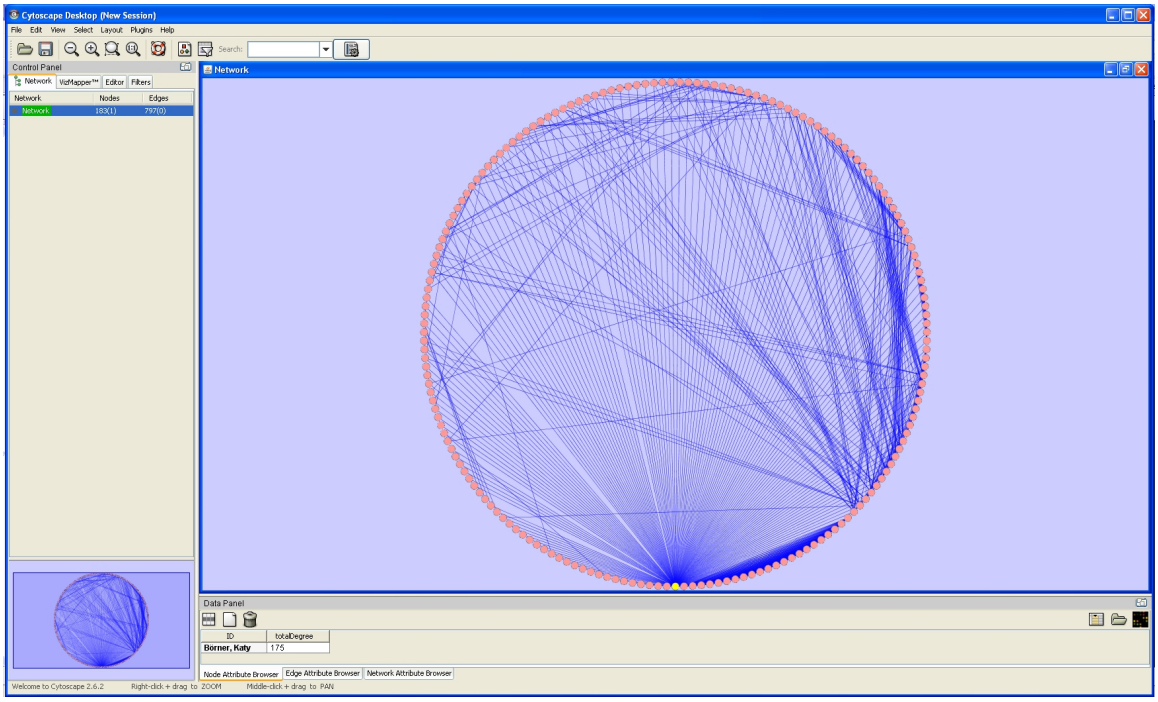
Cytoscape (http://www.cytoscape.org) is an open source software platform for visualizing networks and integrating these with any type of attribute data. Although Cytoscape was originally designed for biological research, now it is a general platform for network analysis and visualization. A variety of layout algorithms are available, including cyclic, tree, force-directed, edge-weight, and yFiles Organic layouts.

Cytoscape is available as a plugin in the extended version of the Sci^2 Tool. To visualize networks in Sci^2 using Cytoscape, first you have to extend the Sci^2 Tool according to the directions at Section 3.2 Additional Plugins.

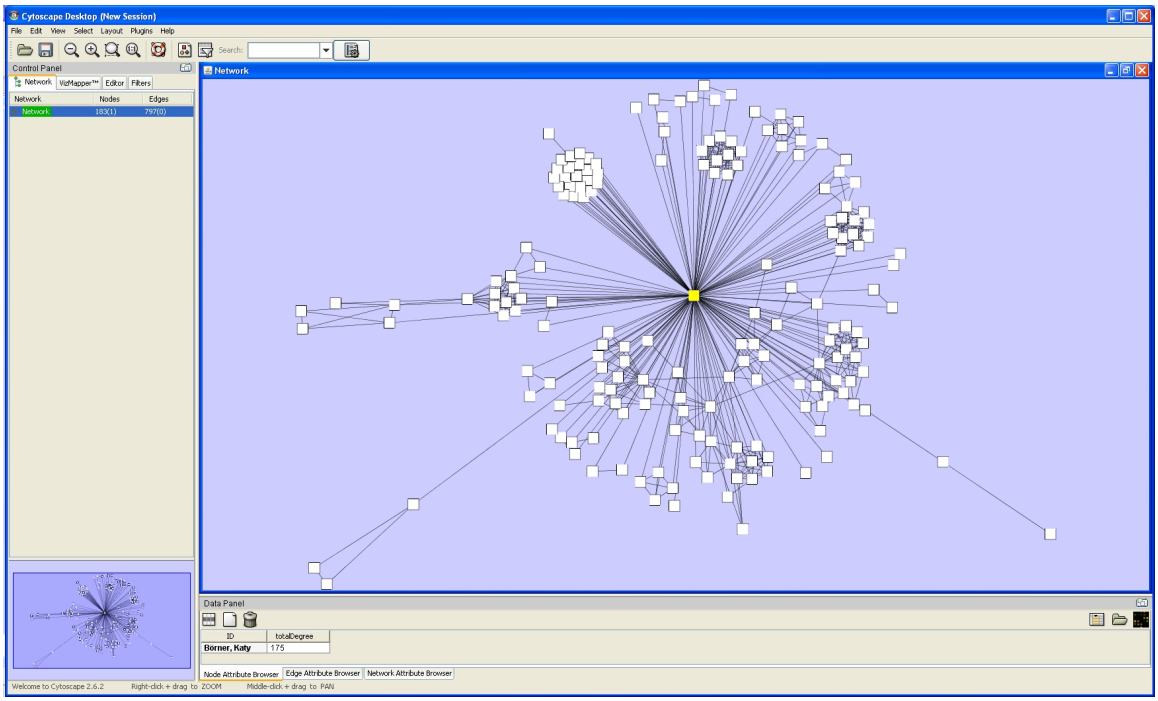
To exemplify how to use Cytoscape to visualize Networks, let's use Dr. Börner's co-authorship network generated above. To visualize this network, run '*Visualization > Networks > Cytoscape*'.

In Cytoscape, the Layout menu has an array of features for organizing the network visually according to one of several algorithms, aligning and rotating groups of nodes, and adjusting the size of the network.

For example, run 'Layout > Degree Sorted Circle Layout > All Nodes' once the network is fully loaded. This layout algorithm sort nodes in a circle by degree of the nodes. When the layout for the network finishes, Cytoscape will look similar to the image below:



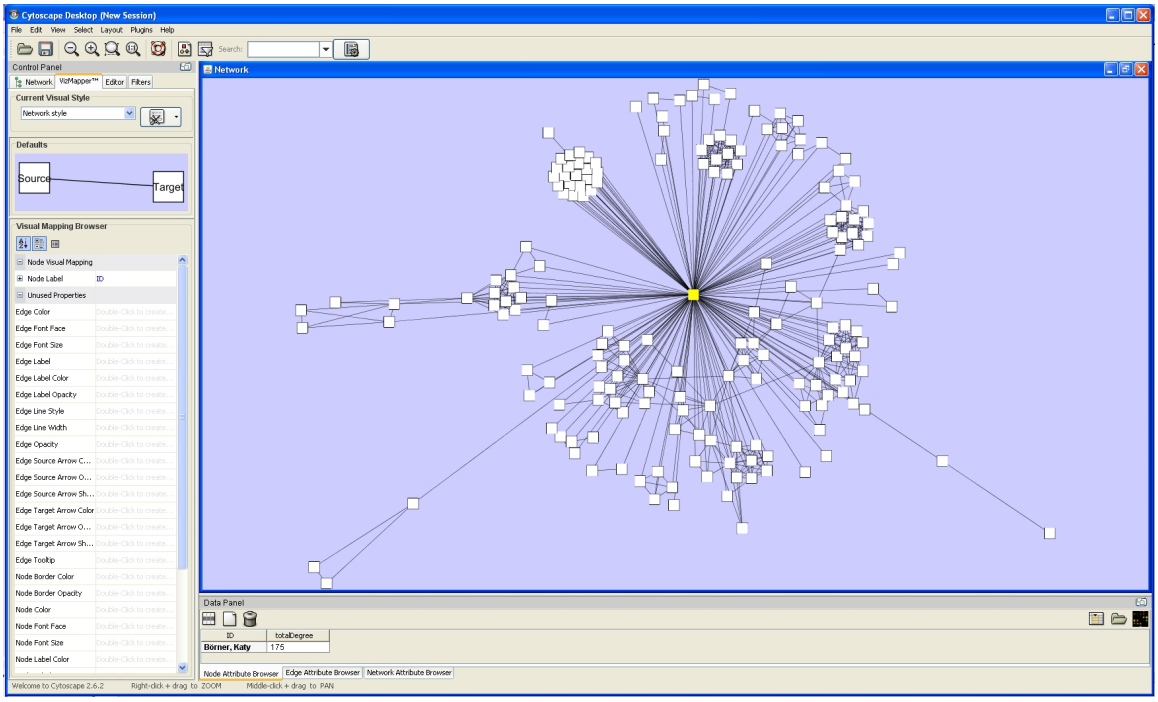
Now, run 'Layout > Cytoscape Layouts > Spring Embedded > All nodes'. This spring-embedded layout is based on a "force-directed" paradigm as implemented by Kamada and Kawai (1988). Network nodes are treated like physical objects that repel each other, such as electrons. The connections between nodes are treated like metal springs attached to the pair of nodes. These springs repel or attract their end points according to a force function. The layout algorithm sets the positions of the nodes in a way that minimizes the sum of forces in the network. To see a description of all layouts and functionalities available in Cytoscape see Cytoscape User Manual.



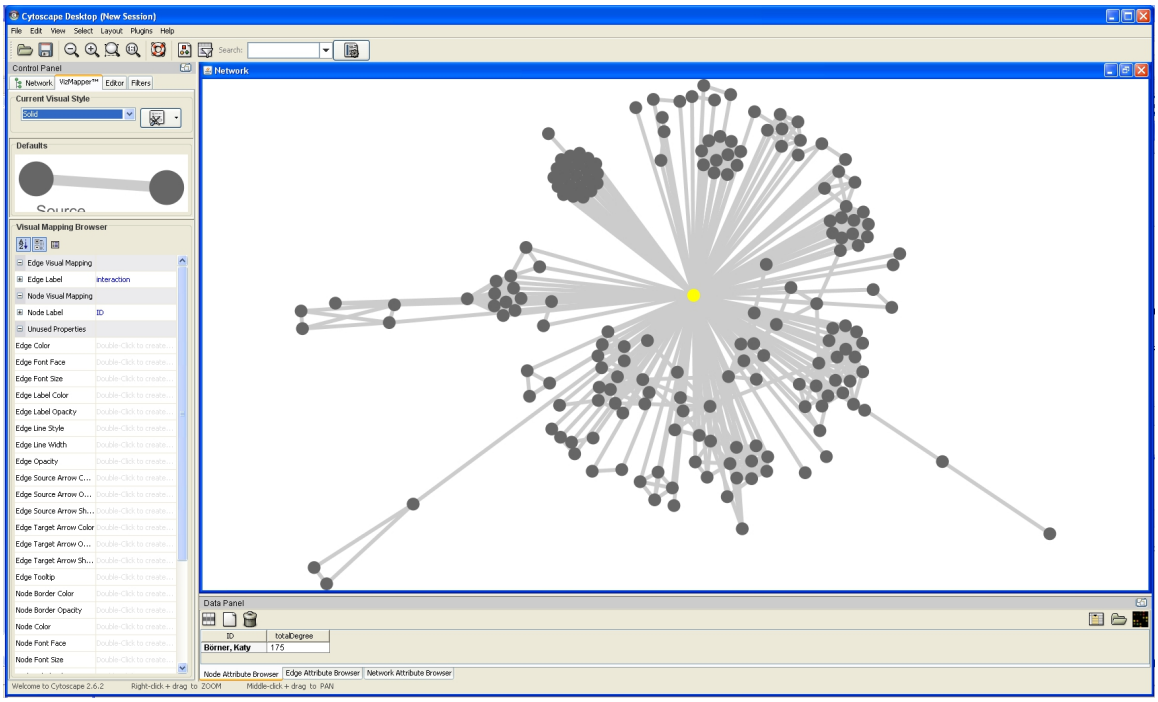
The main window of Cytoscape has several components:

- The menu bar at the top
- The toolbar, which contains icons for commonly used functions. These functions are also available via the menus. Hover the mouse pointer over an icon and wait momentarily for a description to appear as a tooltip.
- The Control Panel that allows the management of the network (top left panel). This contains an optional network overview pane (shown at the bottom left).
- The main network view window, which displays the network.
- The Data Panel (bottom panel), which displays attributes of selected nodes and edges and enables you to modify the values of attributes.

One of Cytoscape's strengths in network visualization is the ability to allow users to encode any attribute of their data (name, type, degree, weight, expression data, etc.) as a visual property (such as color, size, transparency, or font type). A set of these encoded or mapped attributes is called a *Visual Style* and can be created or edited using the Cytoscape *VizMapper*. VizMapper is located at the second tab at the Control Panel.

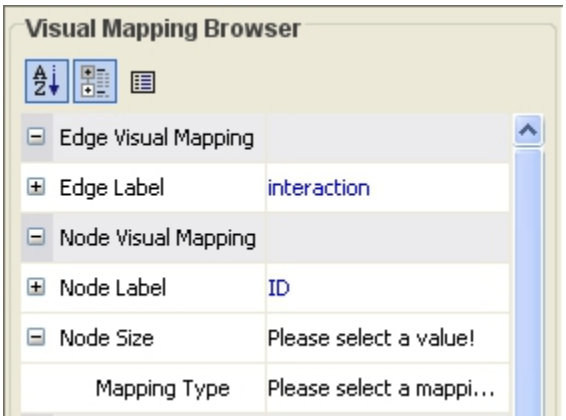


You can change visual styles by making a selection from the *Current Visual Style* dropdown list (found at the top of the *VizMapper* main panel). For example, if you select *Solid*, a new visual style will be applied to your network, and you will see
a white background and round gray nodes.

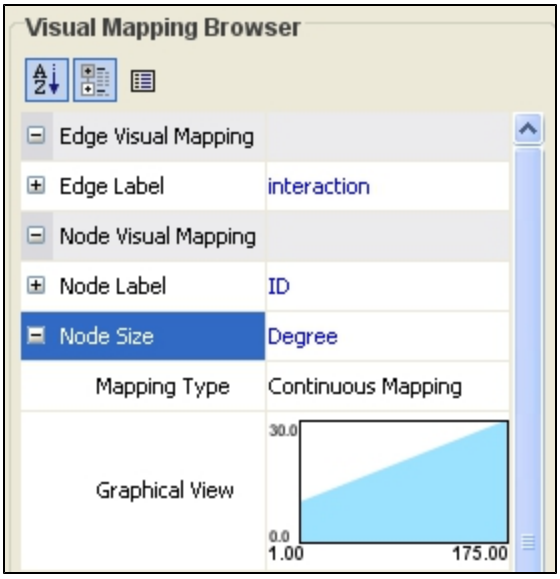


All visual attributes are listed in the *Unused Properties* category. From this panel, you can create node/edge mappings for all visual properties.

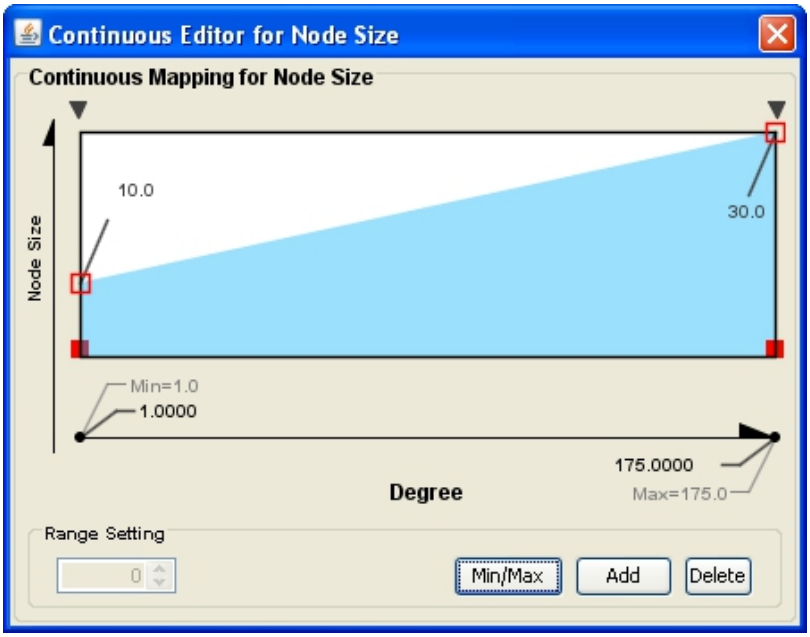
Double click the Node Size entry listed in *Unused Properties*. *Node Size* will now appear at the top of the list, under the *Node Visual Mapping* category (as shown below).



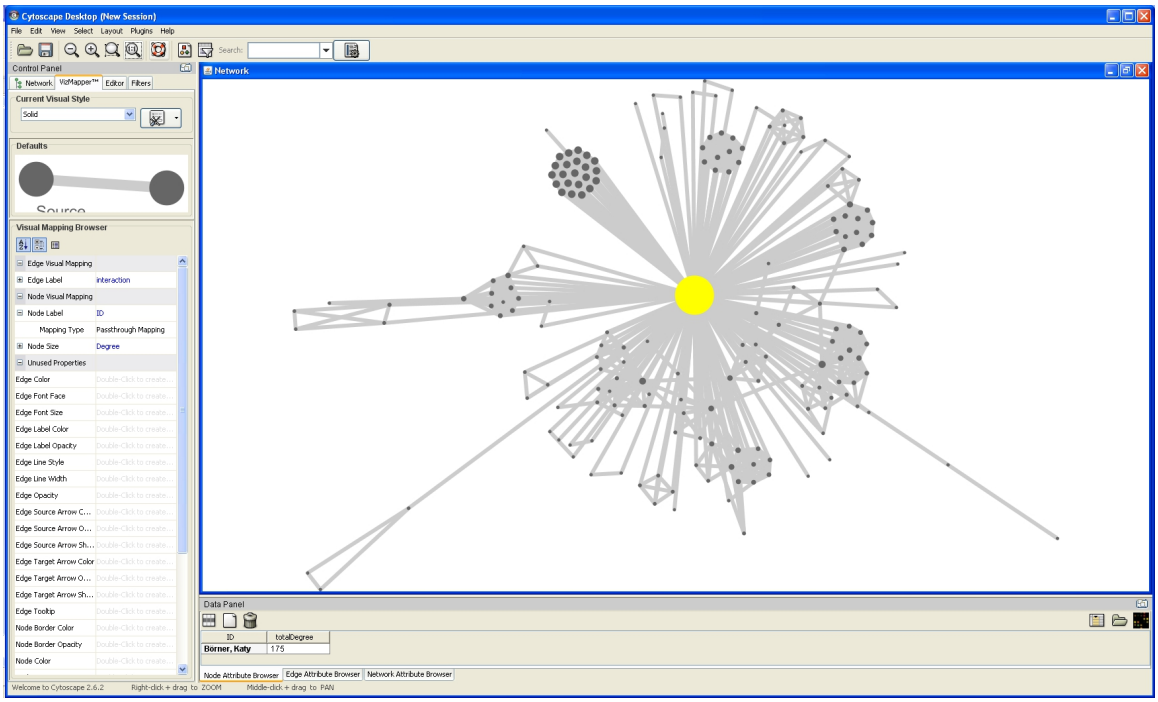
Click on the cell to the right of the *Node Size* entry and select *Degree* from the drop-down list that appears. Set the *Continuous Mapping* option as the *Mapping Type*.



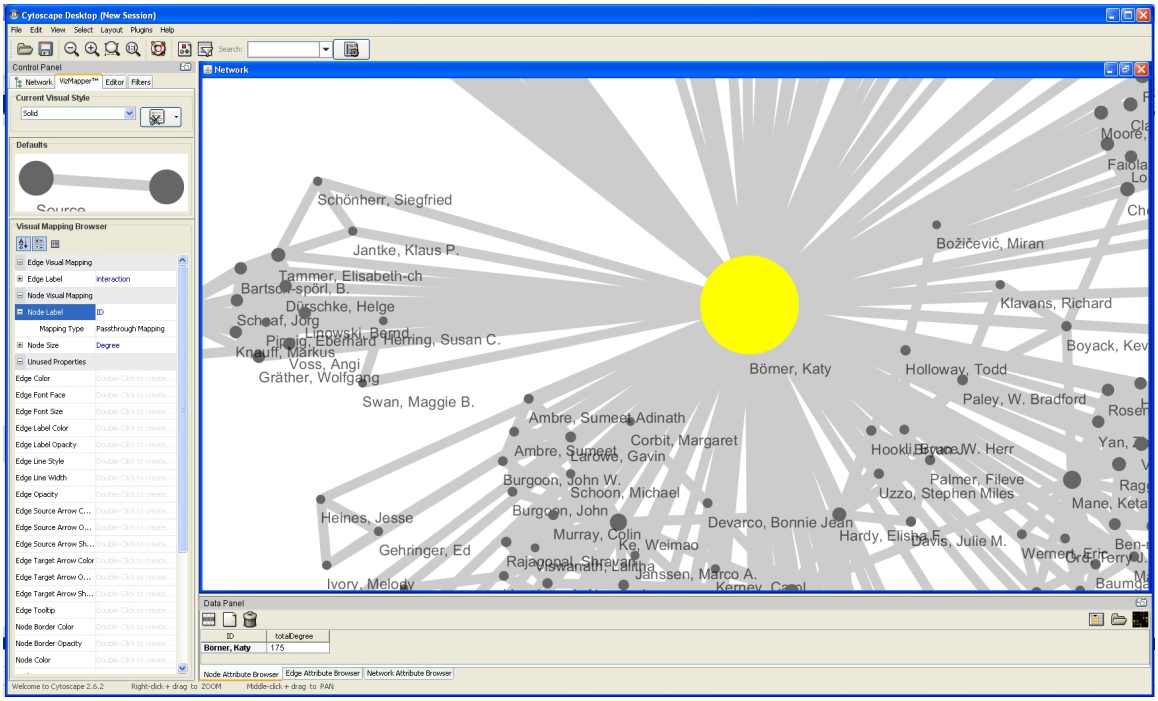
Double-click on the black-and-blue rectangle next to *Graphical View* to open the *Continuous Editor for Node Size*. Double-lick on the second point (set at 30.0 initially) and type as 120 as the new value. The nodes size will be immediately updated. Close the *Continuous Editor for Node Size*.



The network will look like this:



Nodes labels appear an zoom is applied to the network:



To see the log file from this workflow save the 5.1.1.1.1 Using Cytoscape to Visualize Netowrks log file.

5.1.1.2 NSF

KatyBorner.nsf	
Time frame:	2003-2008

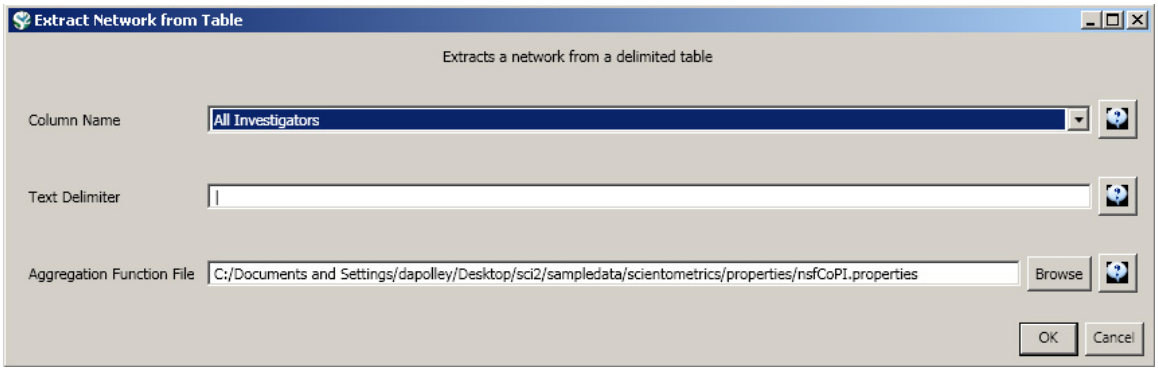
Region(s):	Indiana University
Topical Area(s):	Network Science, Library and Information Science, Informatics and Computing, Statistics, Cyberinfrastructure, Information Visualization, Cognitive Science, Biocomplexity
Analysis Type(s):	Co-PI Network, Grant Award Summary

The second part of Katy Börner's research profile will focus on her Co-PIs. The data can be downloaded for free using NSF's Award Search (See section 4.2.2.1 NSF Award Search) by searching for "Katy Borner" in the "Principal Investigator" field and keeping the "Include CO-PI" box checked.

Load the NSF data using '*File > Load*' and following this path: '***yoursci2directory****/sampledata/scientometrics/nsf/KatyBorner.nsf*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Make sure the loaded dataset in the Data Manager window is highlighted in blue, and run *'Data Preparation > Extract Co-Occurrence Network'* using these parameters:

Aggregate Function File

Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



Select the "Extracted Network on Column All Investigators" network and run *'Analysis > Networks > Network Analysis Toolkit (NAT)'* to reveal that there are 13 nodes and 28 edges without isolates in the network. Click on "Extracted Network on Column All Investigators" and select *'Visualization > Networks > GUESS'* to visualize the resulting Co-PI network. Select '*GEM*' from the layout menu.

Load the default Co-PI visualization theme via '*Script > Run Script ...*'and load '***yoursci2directory****/scripts/GUESS/co-PI-nw.py*'. Alternatively, use the "Graph Modifier" to customize the visualization. The resulting network in Figure 5.2 was modified using the following workflow:

1. Resize Linear > Nodes > totalawardmoney > From: 5 To: 35 > Do Resize Linear
2. Resize Linear > Edges > coinvestigatedawards From: 1 To: 2 > Do Resize Linear
3. Colorize > Nodes > totalawardmoney From : To: > Do Colorize
4. Colorize > Edges > coinvestigatedawards From: To: > Do Colorize
5. Object: all nodes > Show Label
6. Type in Interpreter:

```
>for n in g.nodes:
     n.strokecolor = n.color
```

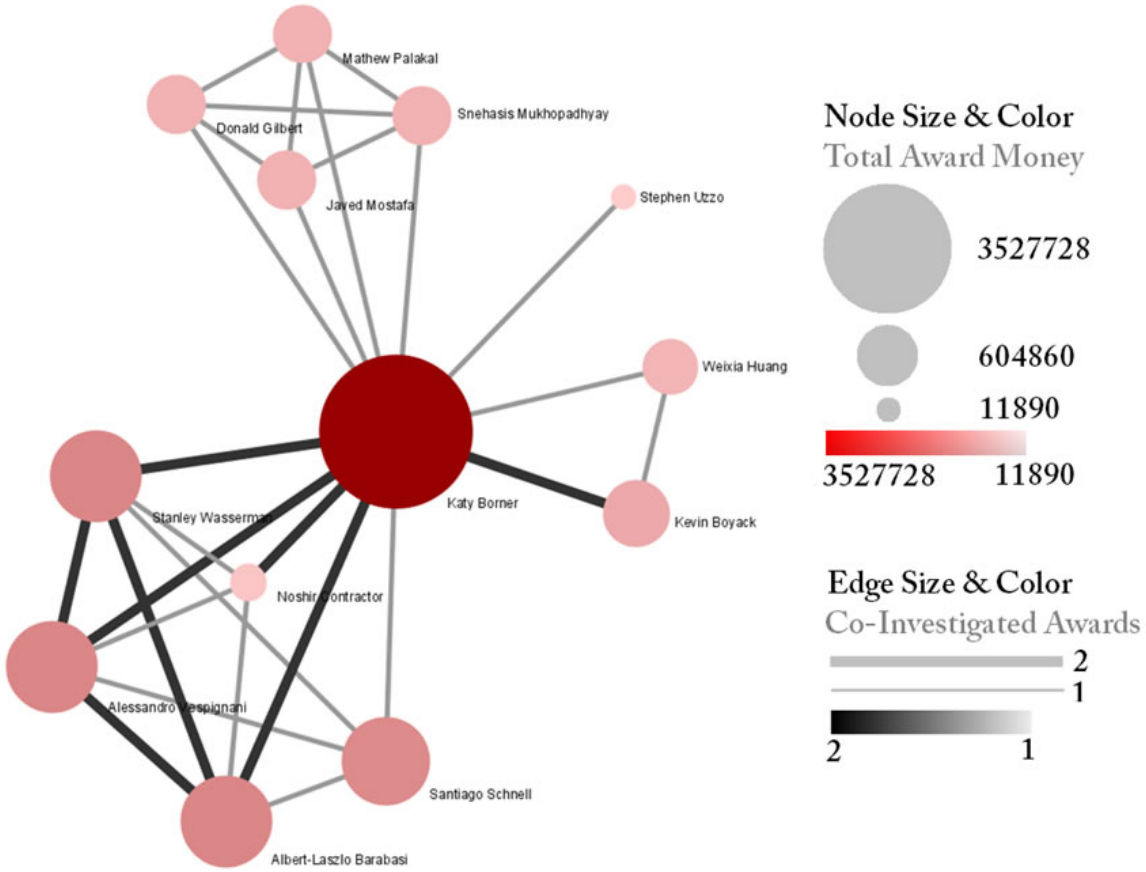
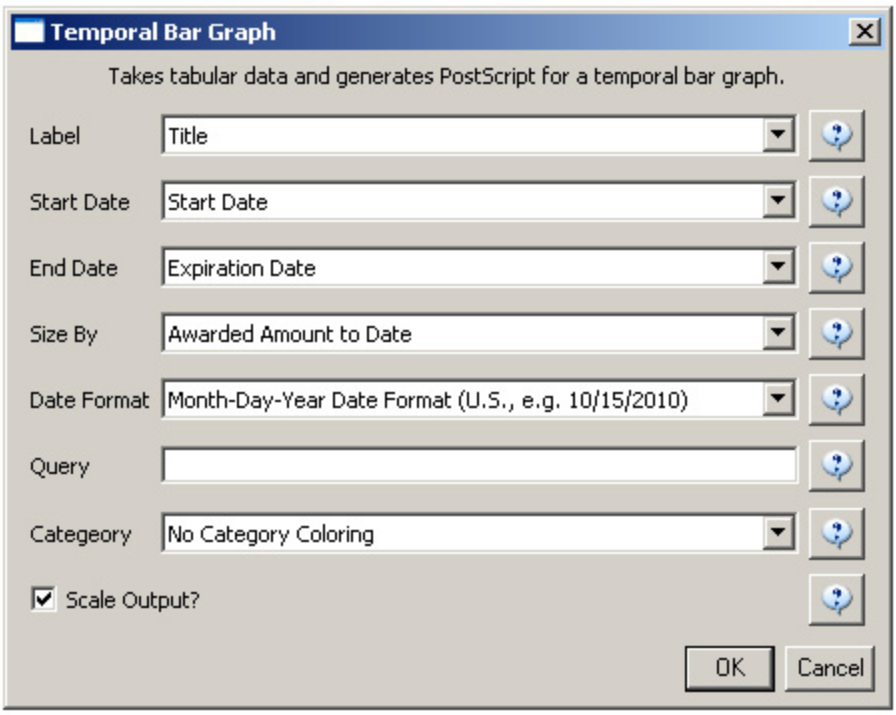


Figure 5.2: NSF Co-PI network of Katy Börner

For a summary of the grants themselves, with a visual representation of their award amount, select the original *'Katy Borner.nsf'* csv file in the Data Manager and run '*Visualization* > *Temporal* > *Temporal Bar Graph*', entering the following parameters:



The generated postscript file "HorizontalBarGraph_KatyBorner.ps" can be viewed using Adobe Distiller or GhostViewer (see section 2.4 Saving Visualizations for Publication).

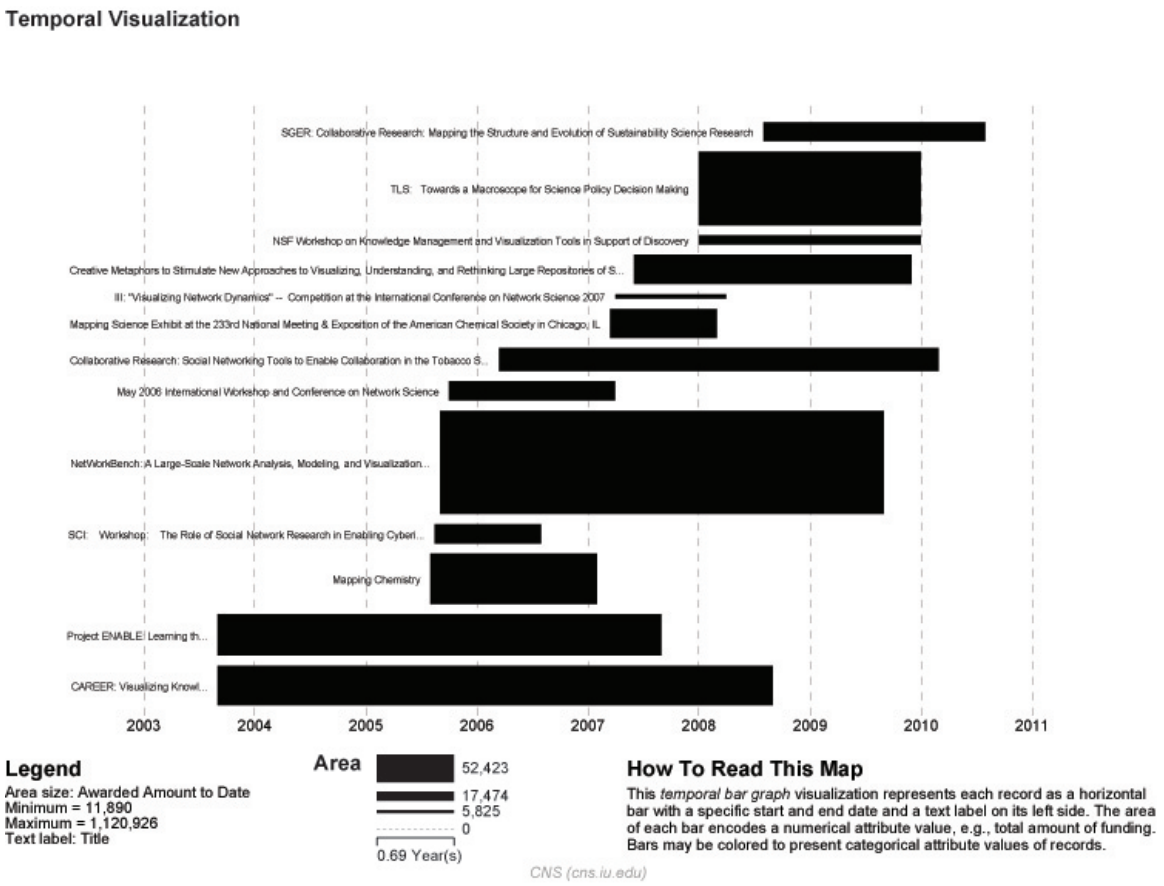


Figure 5.3: Horizontal Bar Graph of KatyBorner.nsf

To see the log file from this workflow save the 5.1.1.2 NSF log file.

5.1.2 Time Slicing of Co-Authorship Networks (ISI Data)

AlessandroVespignani.isi	
Time frame:	1990-2006
Region(s):	Indiana University, University of Rome, Yale University, Leiden University, International Center for Theoretical Physics, University of Paris-Sud
Topical Area(s):	Informatics, Complex Network Science and System Research, Physics, Statistics, Epidemics
Analysis Type(s):	Co-Authorship Network

The Sci^2 Tool supports the analysis of evolving networks. For this study, load Alessandro Vespignani's publication history from ISI, which can be downloaded from Thomson's Web of Science or loaded using '*File > Load*' and following this path: '***yoursci2directory****/sampledata/scientometrics/isi/AlessandroVespignani.isi*'using' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets).

New ISI File Format

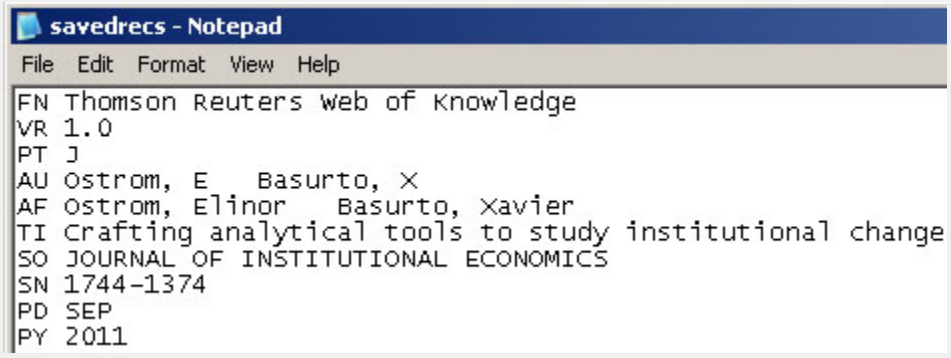
Web of Science made a change to their output format in September, 2011. Older versions of Sci2 tool (older than v0.5.2 alpha) may refuse to load these new files, with an error like "Invalid ISI format file selected."

Sci2 solution

If you are using an older version of the Sci2 tool, you can download the WOS-plugins.zip file and unzip the JAR files into your sci2/plugins/ directory. Restart Sci2 to activate the fixes. You can now load the downloaded ISI files into the Sci2 without any additional step. If you are using the old Sci2 tool you will need to follow the guidelines below before you can load the new WOS format file into the tool.

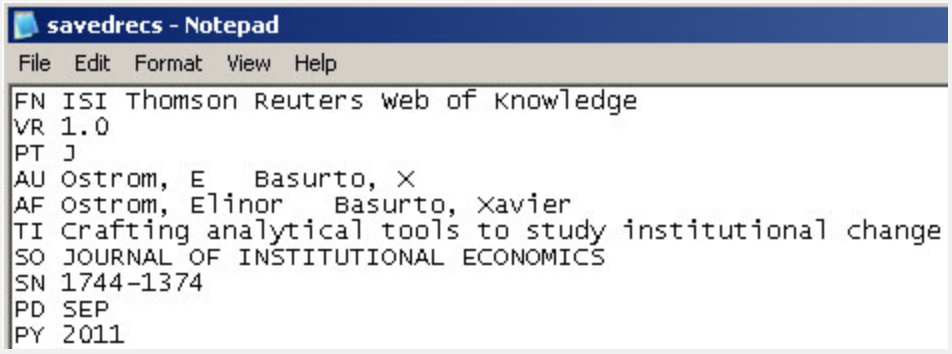
You can fix this problem for individual files by opening them in Notepad (or your favorite text editor). The file will start with the words:

Original ISI file:



Just add the word ISI.

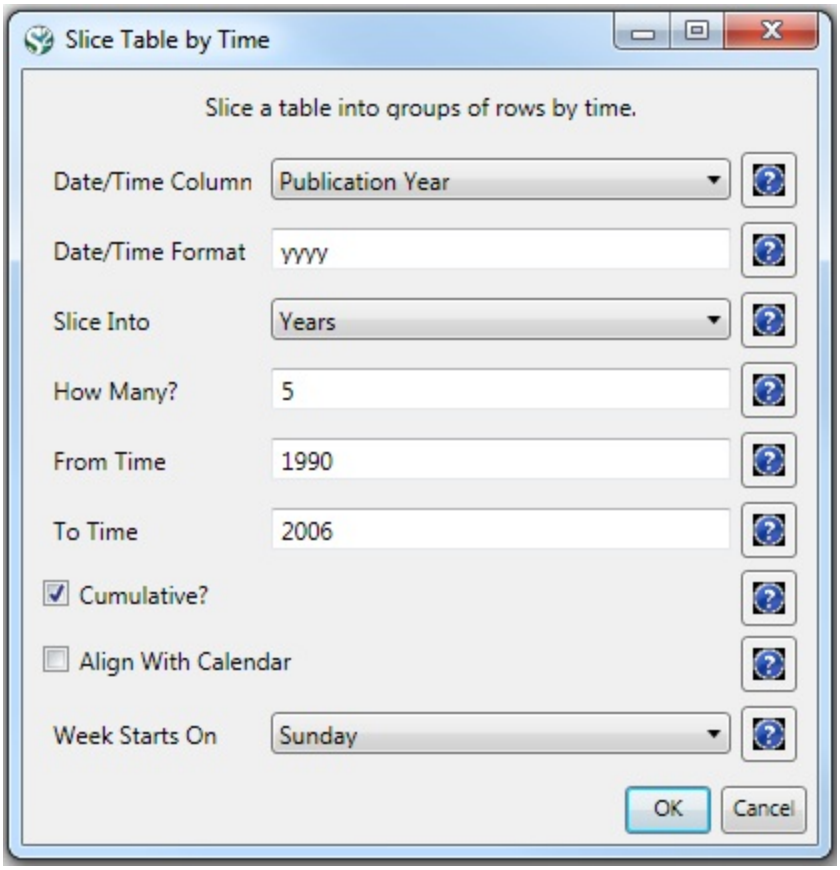
Updated ISI file:



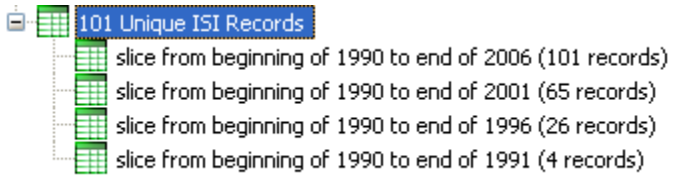
And then Save the file.

The ISI file should now load properly. More information on the ISI file format is available here (http://wiki.cns.iu.edu/display/CISHELL/ISI+%28*.isi%29).

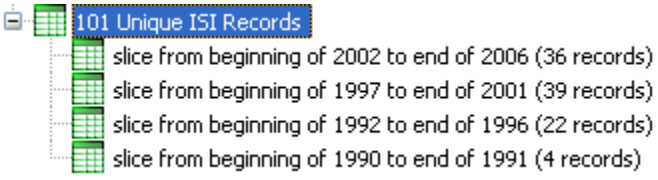
Slice the data into five year intervals from 1990-2006 using *'Preprocessing > Temporal > Slice Table by Time'* and the following parameters:



"Slice Into" allows the user to slice the table by days, weeks, months, quarters, years, decades, and centuries. There are two additional parameters for time slicing: cumulative and align with calendar. The former produces tables containing all data from the beginning to the end of each table's time interval, which can be seen in the Data Manager and below:



The latter option aligns the output tables according to calendar intervals:



Choosing "Years" under "Slice Into" creates multiple tables beginning from January 1st of the first year. If "Months" is chosen, it will start from the first day of the earliest month in the chosen time interval.

To see the evolution of Vespignani's co-authorship network over time, check "cumulative". Then, extract co-authorship networks one at a time for each sliced time table using *'Data Preparation* > *Extract Co-Author Network* ', making sure to select "ISI" from the pop-up window during the extraction. To view each of the Co-Authorship Networks over time using the same graph layout, begin by clicking on longest slice network (the 'Extracted Co-Authorship Network' under 'slice from beginning of 1990 to end of 2006 (101 records)') in the data manager.

Visualize it in GUESS using '*Visualization* > *Networks* > *GUESS*'. From here, run '*Layout* > *GEM*' followed by '*Layou t* > *Bin Pack'*. Run '*Script* > *Run Script ...*' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*'.

The resulting visualization is of the complete co-authorship network over all years. Use *Node Locking* to save the *x, y* coordinates of each node and apply them to the other time slices. In GUESS, select '*File* > *Export Node Positions*' and save the result as '***yoursci2directory****/NodePositions.csv*'. Load the remaining three networks in GUESS using the steps described above and for each network visualization, run '*File* > *Import Node Positions*' and open '***yoursci2 directory****/NodePositions.csv*'. To match the resulting networks stylistically with the original visualization, run '*Script* > *Run Script ...*' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*', followed by '*Layout* > *Bin Pack'*, for each. The evolving network is shown in Figure 5.4.

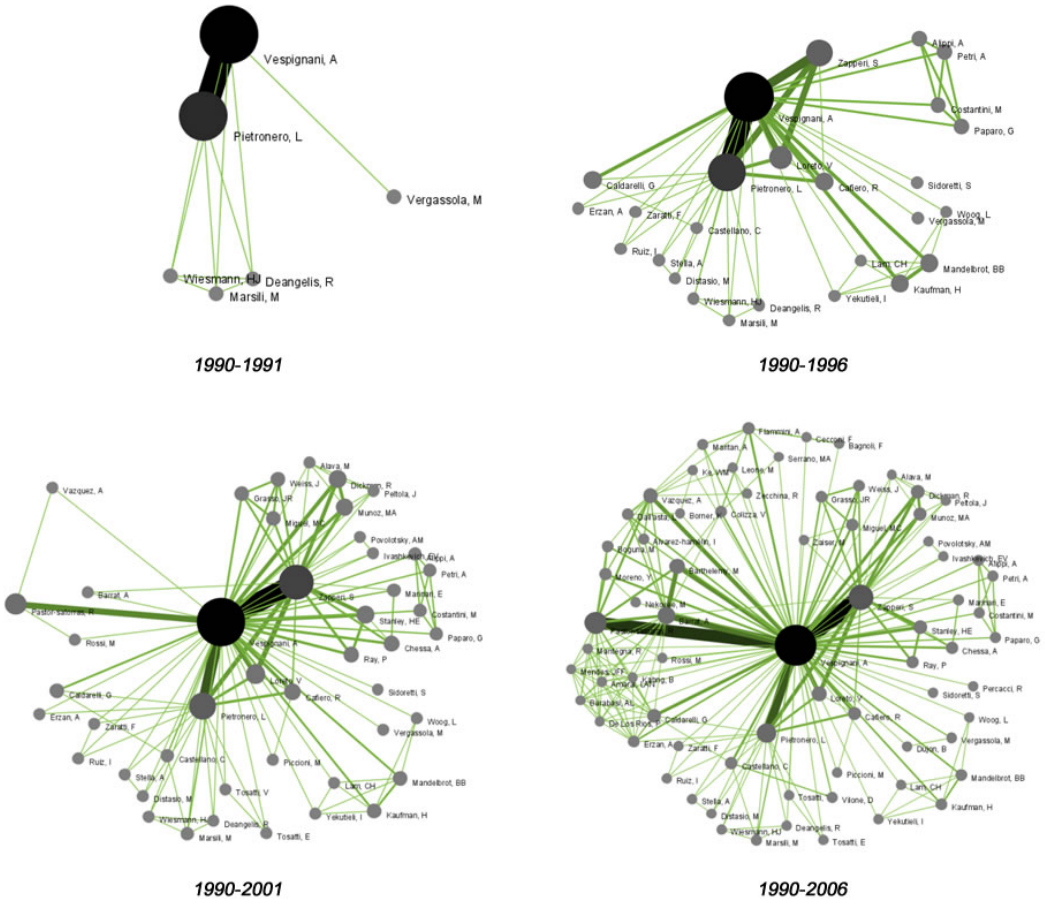


Figure 5.4: Evolving co-authorship network of Vespignani from 1990-2006
The four networks reveal that from 1988-1992, Alessandro Vespignani had one primary co-author and four secondary co-authors. His network expanded considerably to include multiple primary co-authors and more than 200 secondary co-authors in 2006.

To see the log file from this workflow save the 5.1.2 Time Slicing of Co-Authorship Networks (ISI Data) log file.

5.1.3 Funding Profiles of Three Researchers at Indiana University (NSF Data)

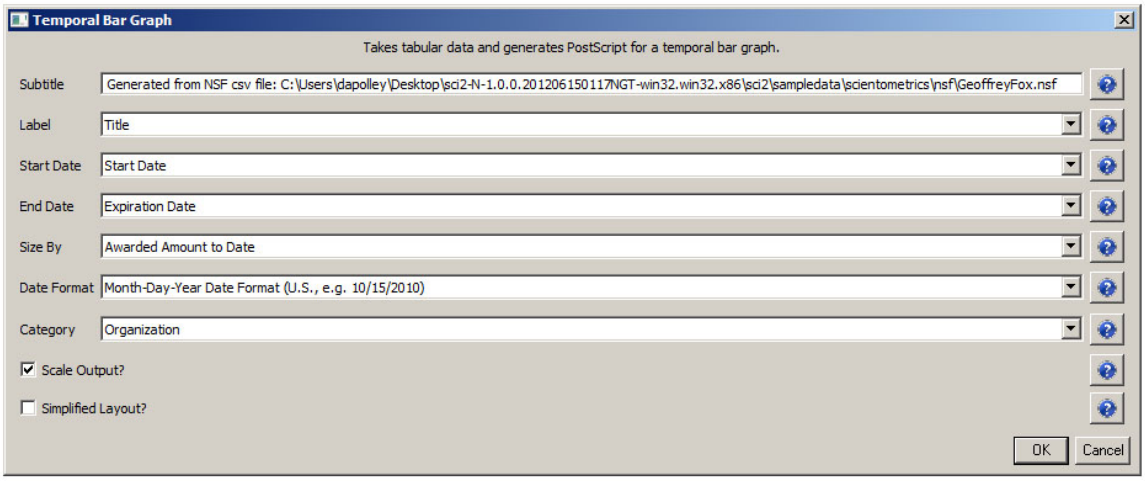
GeoffreyFox.nsf BethPlale.nsf MichaelMcRobbie.nsf	

Time frame:	1978-2010
Region(s):	Indiana University
Topical Area(s):	Informatics, Miscellaneous
Analysis Type(s):	Co-PI Network, Grant Award Summary

It is often useful to compare the profiles of multiple researchers within similar disciplinary or institutional domains. To demonstrate this comparison, load the NSF funding profiles of three Indiana University researchers into the Sci^2 Tool using '*File > Load*' and following this path: '***yoursci2directory****/sampledata/scientometrics/nsf*'. Load *'GeoffreyFox.nsf', 'MichaelMcRobbie.nsf',* and *'BethPlale.nsf'* in NSF csv format (if these files are not in the sample data directory they can be downloaded from 2.5 Sample Datasets). Then run '*Visualization > Temporal > Horizontal Bar Graph*', using the recommended parameters for each.

For instructions on how to save and view the PostScript file generated by the "Horizontal Bar Graph" algorithm, see section 2.4 Saving Visualizations for Publication.

Select *'GeoffreyFox.nsf'* in the Data Manager. Use the following parameters to generate a Horizontal Bar Graph:



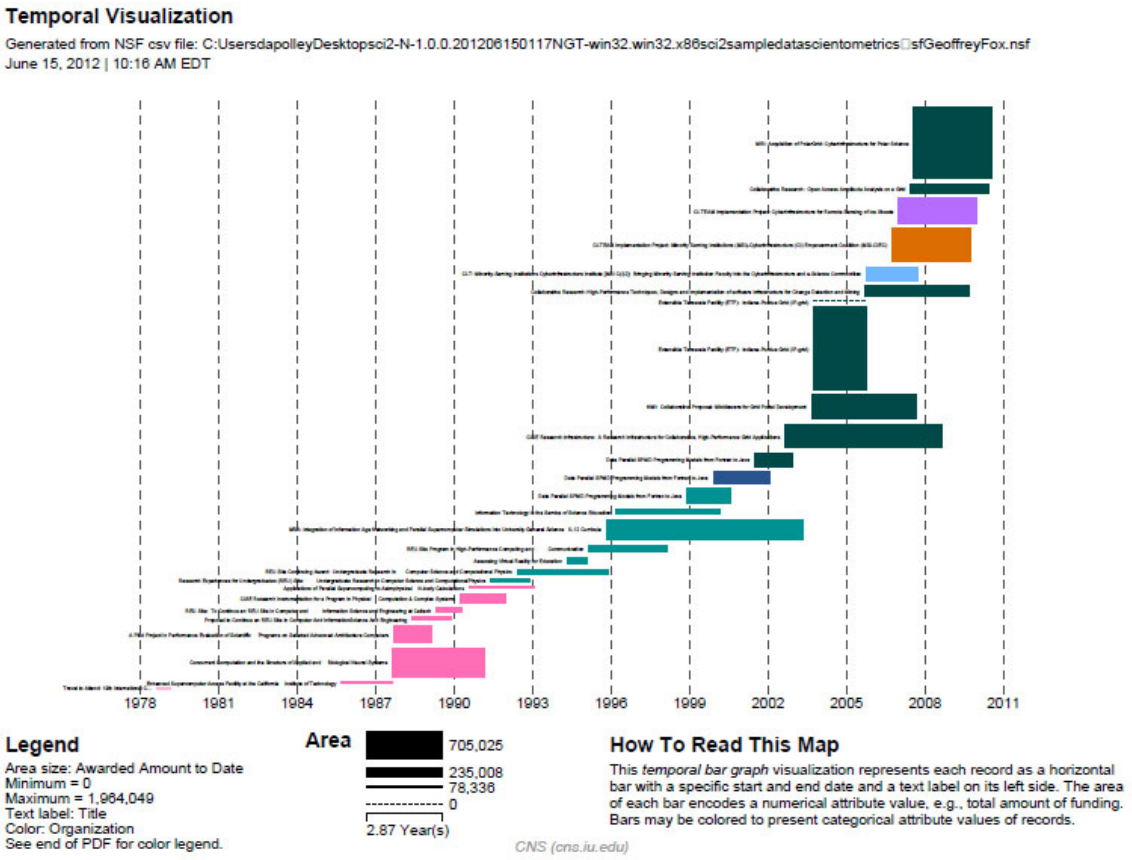


Figure 5.5: Funding profile over time of Geoffrey Fox

Now select *'BethPlale.nsf'* in the Data Manager. Use the following parameters to generate a Horizontal Bar Graph:

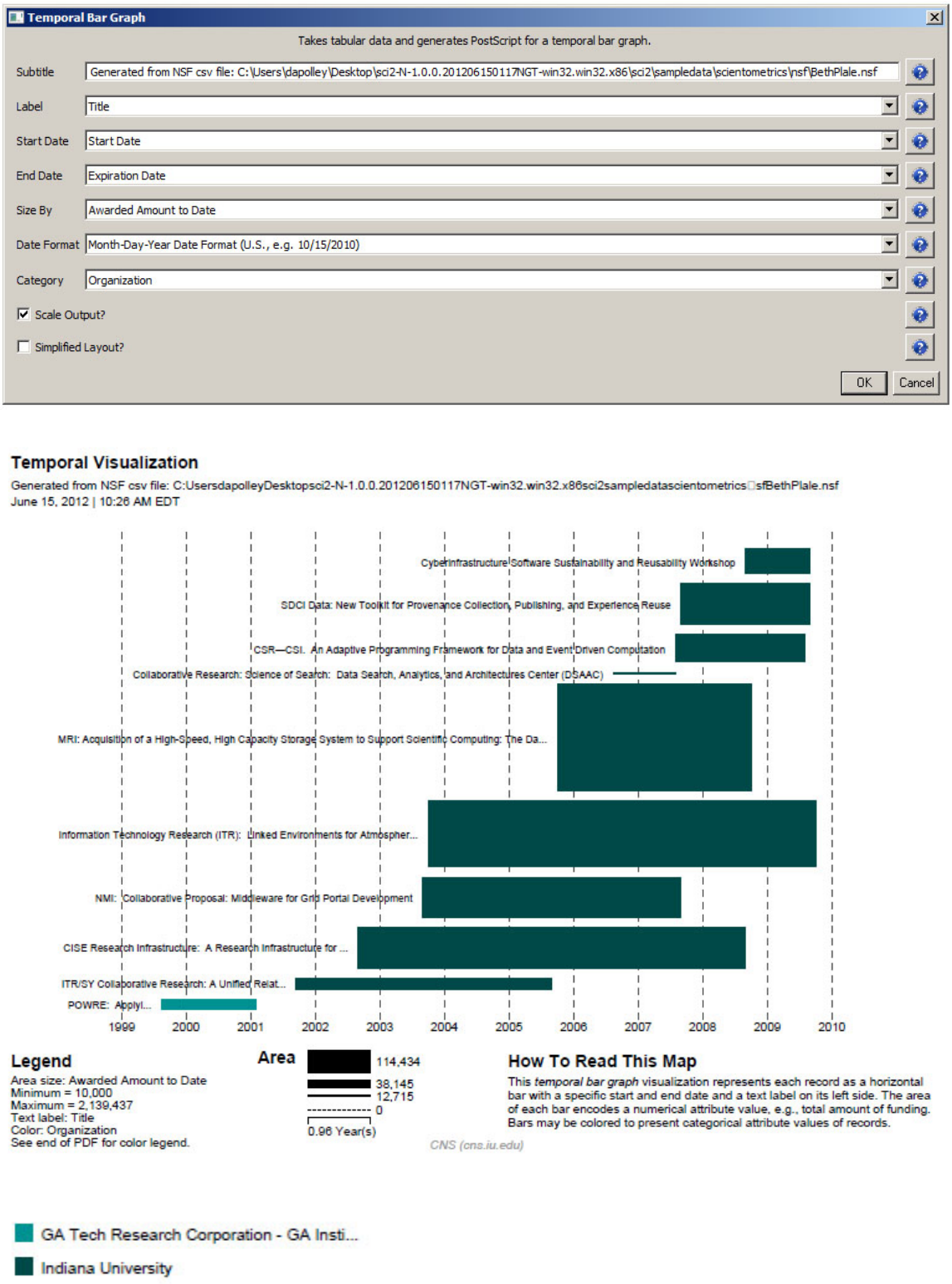
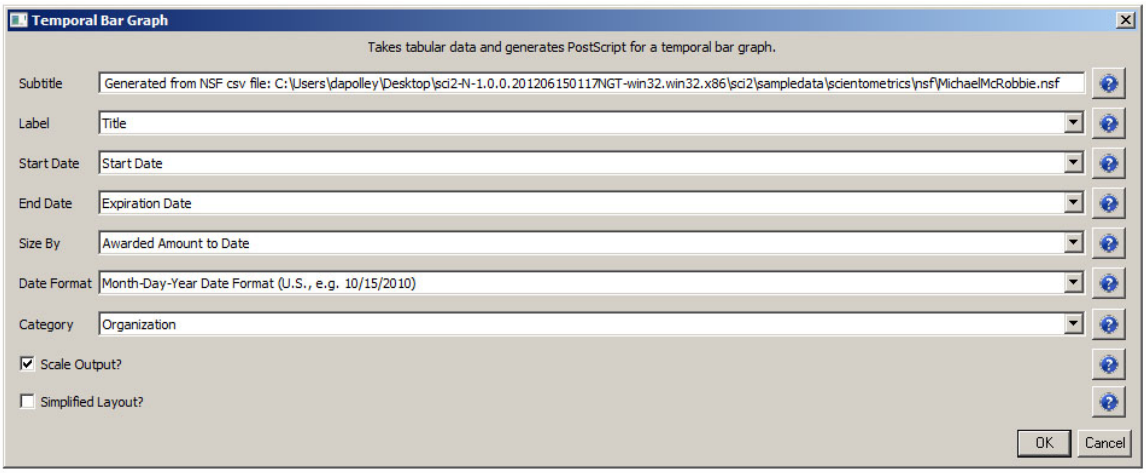


Figure 5.6: Funding profile over time of Beth Plale

Finally, select *'MichaelMcRobbie.nsf'* in the Data Manager. Use the following parameters to generate a Horizontal Bar Graph:



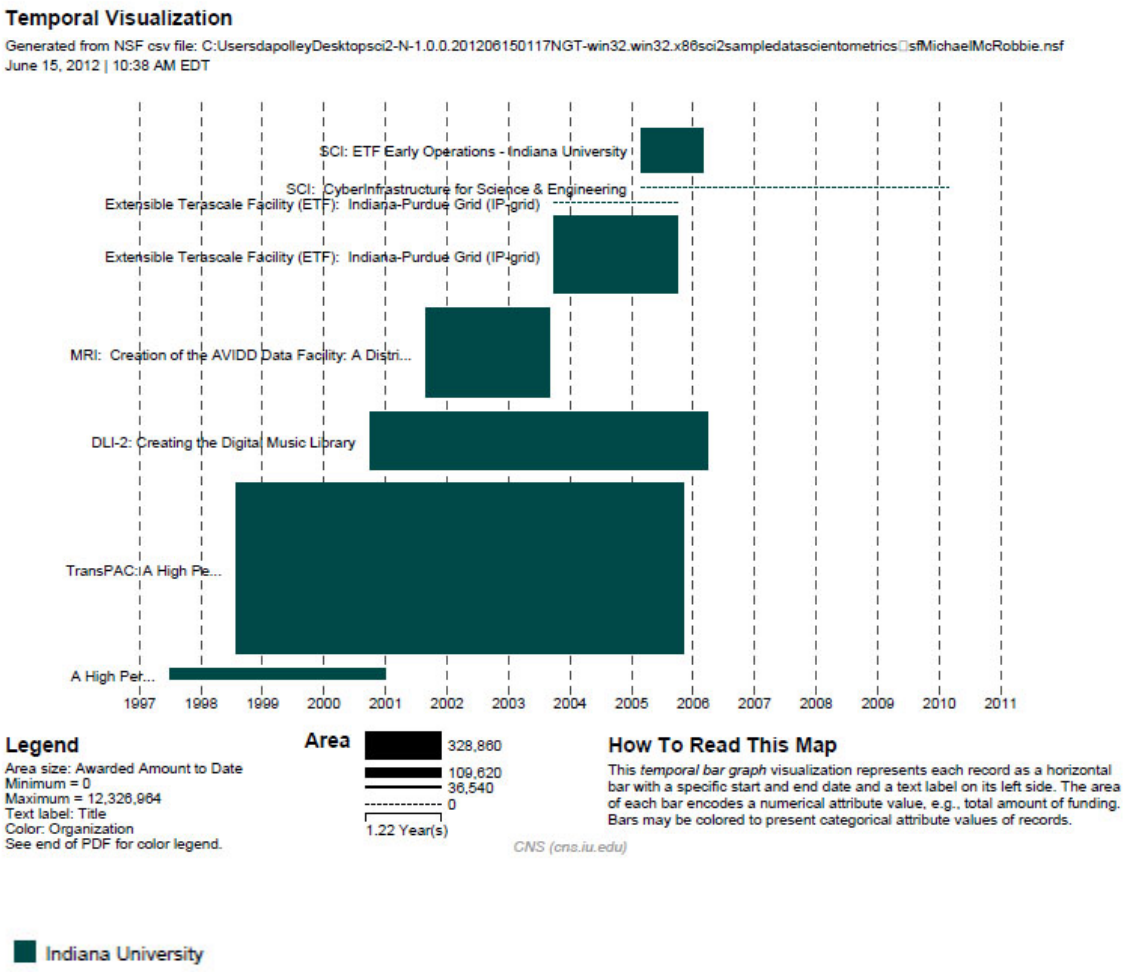
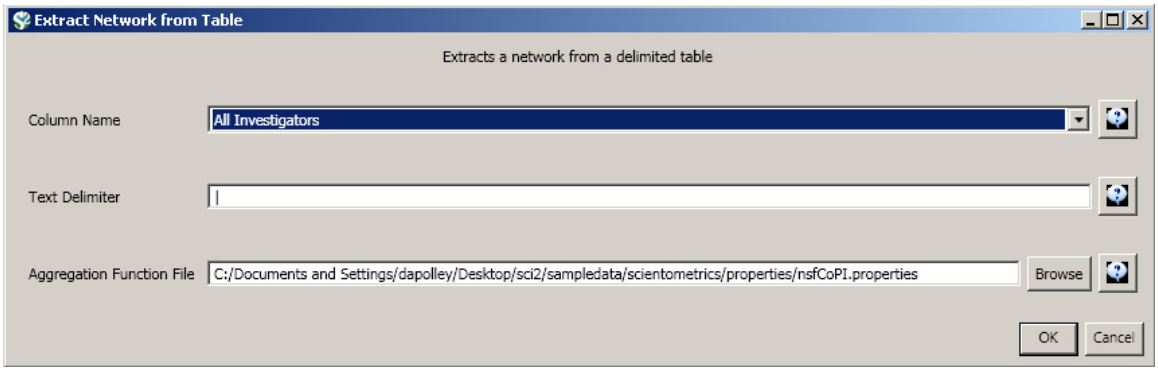


Figure 5.7: Funding profile over time of Michael McRobbie

The horizontal bar graph visualizations in Figures 5.5, 5.6, and 5.7 make it easy to see the timespan of different researchers, as well as the types and volume of grants they generally receive (e.g., many small grants or a handful of large ones). From here, it may be useful to compare their Co-PI networks and look more closely at award totals. Select each dataset in the Data Manager window and run *'Data Preparation > Extract Co-Occurrence Network'* usin g these parameters:

Aggregate Function File

Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



Run *'Visualization > Networks > GUESS'* on each generated network to visualize the resulting Co-PI relationships. Select '*GEM*' from the layout menu to organize the nodes and edges.

To color and size the nodes and edges using the default Co-PI visualization theme, run '***yoursci2directory****/scripts/GUESS/co-PI-nw.py*' from '*Script > Run Script ...*'.

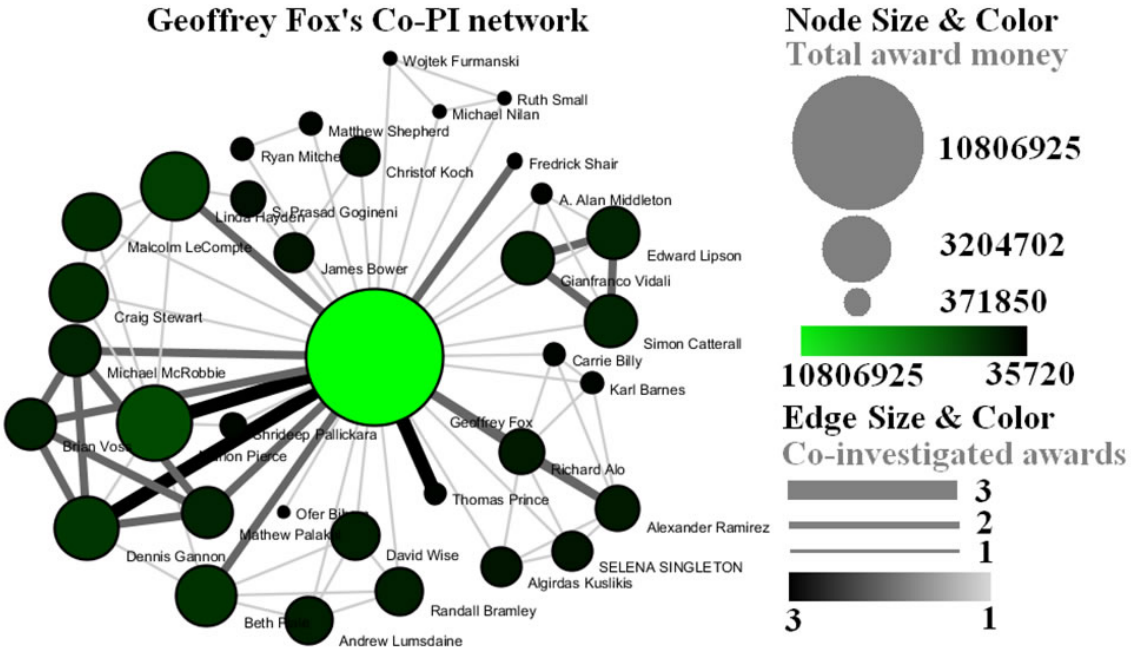


Figure 5.8: Co-PI network of Geoffrey Fox in Indiana University

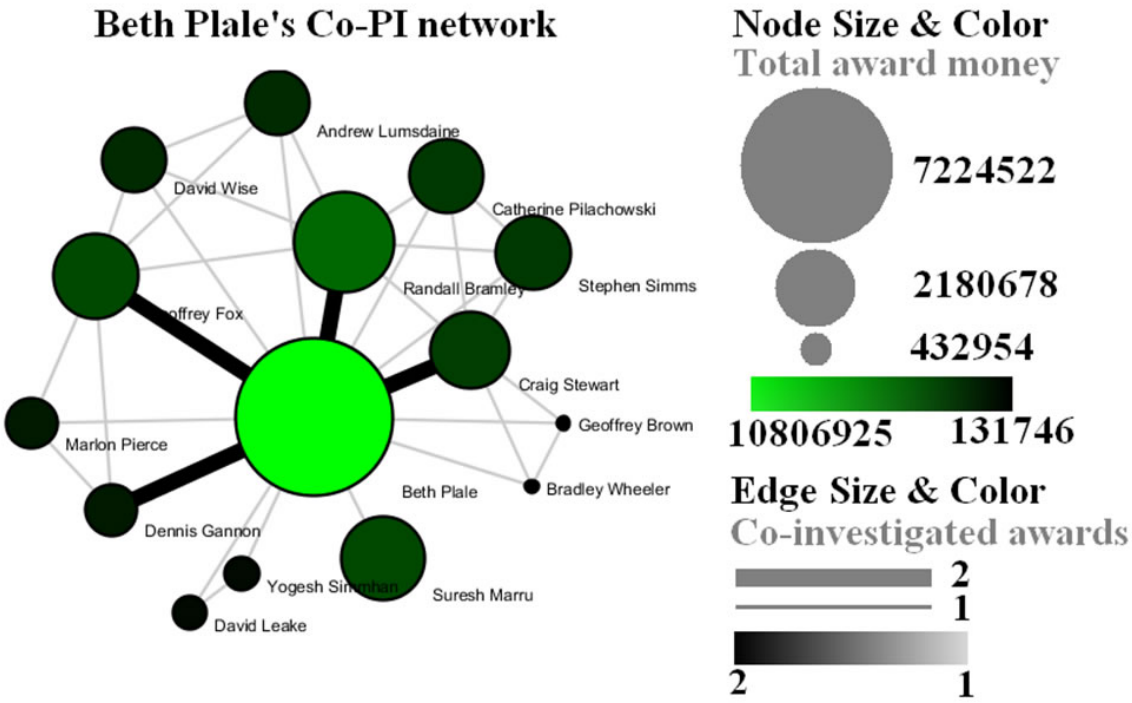


Figure 5.9: Co-PI network of Beth Plale in Indiana University

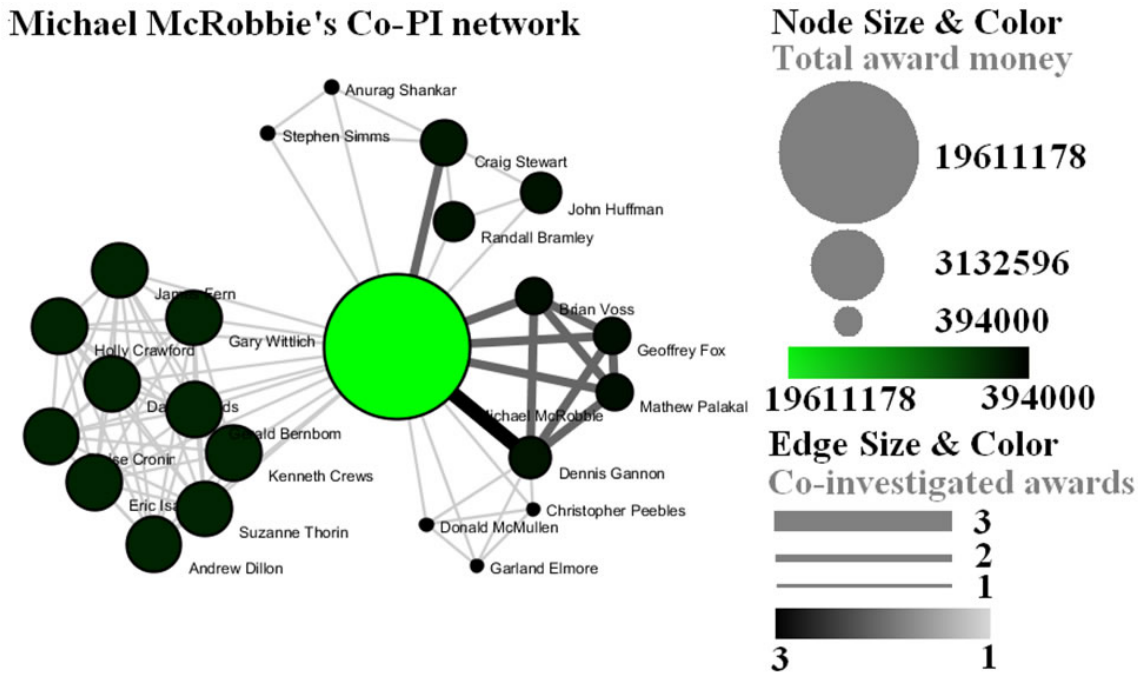


Figure 5.10: Co-PI network of Michael McRobbie in Indiana University

To see the log file from this workflow save the 5.1.3 Funding Profiles of Three Researchers at Indiana University (NSF Data) log file.

5.1.4 Studying Four Major NetSci Researchers (ISI Data)

FourNetSciResearchers.isi	
Time frame:	1955-2007
Region(s):	Miscellaneous
Topical Area(s):	Network Science
Analysis Type(s):	Paper Citation Network, Co-Author Network, Bibliographic Coupling Network, Document Co-Citation Network, Word Co-Occurrence Network

5.1.4.1 Paper-Paper (Citation) Network

Load the file using '*File* > *Load*' and following this path: *'**yoursci2directory**/sampledata/scientometrics/isi/FourNetSciResearchers.isi'* (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). A table of all records and a table of 361 records with unique ISI ids will appear in the Data Manager. In this "clean" file, each original record now has a "Cite Me As" attribute that is constructed from the first author, publication year (PY), journal abbreviation (J9), volume (VL), and beginning page (BP) fields of its ISI record. This "Cite Me As" attribute will be used when matching paper and reference records.

New ISI File Format

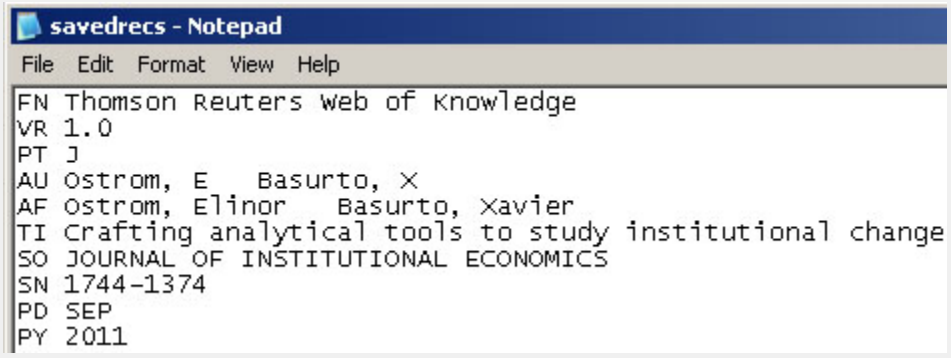
Web of Science made a change to their output format in September, 2011. Older versions of Sci2 tool (Older than v0.5.2 alpha) may refuse to load these new files, with an error like "Invalid ISI format file selected."

Sci2 solution

If you are using an older version of the Sci2 tool, you can download the WOS-plugins.zip file and unzip the JAR files into your sci2/plugins/ directory. Restart Sci2 to activate the fixes. You can now load the downloaded ISI files into the Sci2 without any additional step. If you are using the old Sci2 tool you will need to follow the guidelines below before you can load the new WOS format file into the tool.

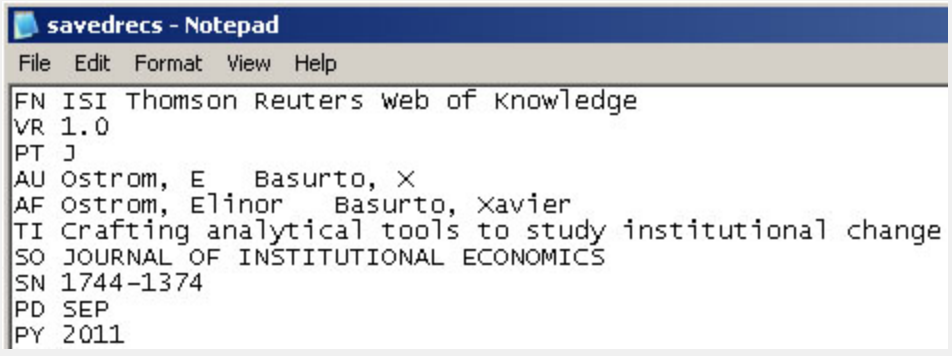
You can fix this problem for individual files by opening them in Notepad (or your favorite text editor). The file will start with the words:

Original ISI file:



Just add the word ISI.

Updated ISI file:



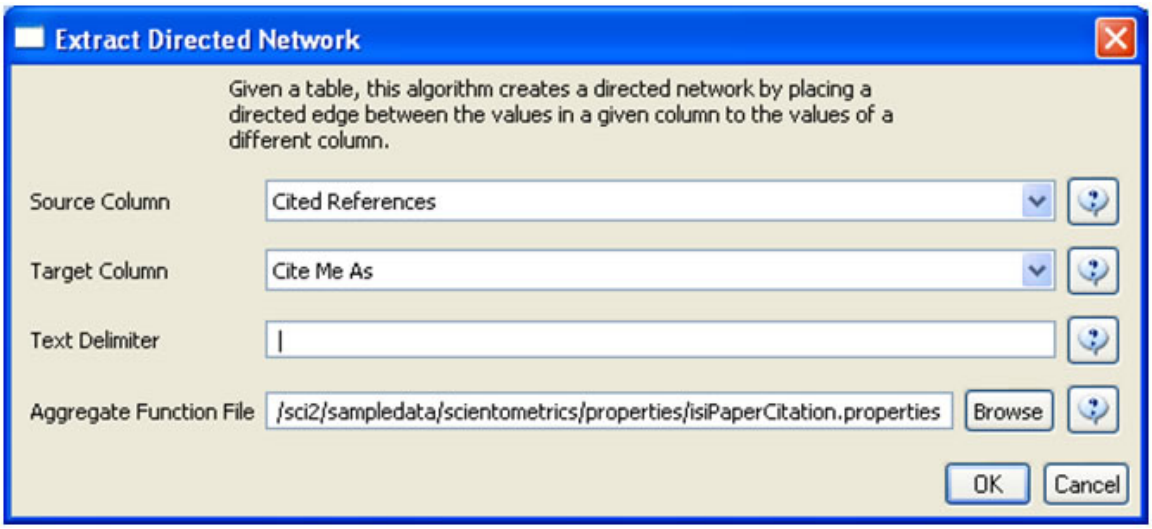
And then Save the file.

The ISI file should now load properly. More information on the ISI file format is available here (http://wiki.cns.iu.edu/display/CISHELL/ISI+%28*.isi%29).

To extract the paper citation network, select the *'361 Unique ISI Records'* table and run *'Data Preparation* > *Extract Directed Network'* using the parameters :

Aggregate Function File

Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



The result is a directed network of paper citations in the Data Manager. Each paper node has two citation counts. The local citation count (LCC) indicates how often a paper was cited by papers in the set. The global citation count (GCC) equals the times cited (TC) value in the original ISI file. Only references from other ISI records count towards an ISI paper's GCC value. Currently, the Sci^2 Tool sets the GCC of references to -1 (except for references that are not also ISI records) to prune the network to contain only the original ISI records.

To view the complete network, select the "Network with directed edges from Cited References to Cite Me As" in the Data Manager and run *'Visualization > Networks > GUESS'* and wait until the network is visible and centered. Because the FourNetSciResearchers dataset is so large, the visualization will take some time to load, even on powerful systems.

Select *' Layout > GEM'* to group related nodes in the network. Again, because of the size of the dataset, this will take some time. Use *'Layout > Bin Pack'* to "pack" the nodes and edges, bringing them closer together for easier analysis and comparison. To size and color-code nodes in the "Graph Modifier," use the following workflow:

1. Resize Linear > Nodes > globalcitationcount> From: 1 To: 50 > When the nodes have no 'globalcitationcount': 0.1 > Do Resize Linear
2. Colorize > Nodes > globalcitationcount > From:  To:  > When the nodes have no 'globalcitationcount': 0.1 >  >Do Colorize
3. Colorize > Edges > weight > From (select the "RGB" tab) 127, 193, 65 To: (select the "RGB" tab) 0, 0, 0
4. Type in Interpreter:

```
>for n in g.nodes:
     n.strokecolor = n.color
```

Or, select the 'Interpreter' tab at the bottom, left-hand corner of the GUESS window, and enter the command lines:

```
> resizeLinear(globalcitationcount,1,50)
> colorize(globalcitationcount,gray,black)
> for e in g.edges:
      e.color="127,193,65,255"
```

Note: The Interpreter tab will have '>>>' as a prompt for these commands. It is not necessary to type '>" at the beginning of the line. You should type each line individually and press "Enter" to submit the commands to the Interpreter.
This will result in nodes which are linearly sized and color coded by their GCC, connected by green directed edges, as shown in Figure 5.11 (left). Any numeric node attribute within the network can be used to code the nodes. To view the available attributes, mouse over a node. The GUESS interface supports pan and zoom, node selection, and details on demand. For more information, refer to the GUESS tutorial at http://nwb.cns.iu.edu/Docs/GettingStartedGUESSNWB.pdf

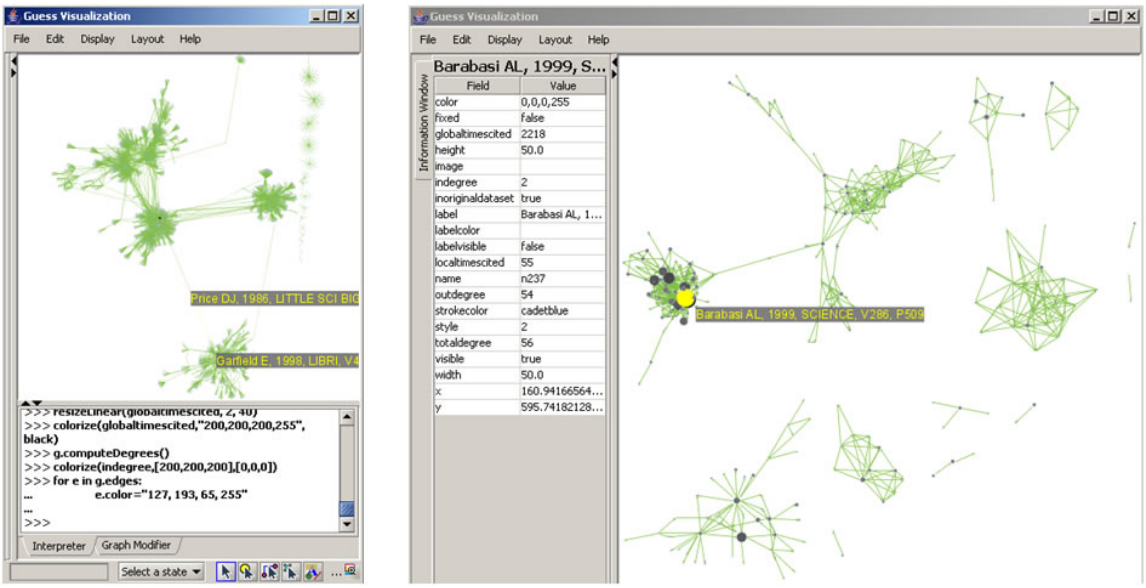
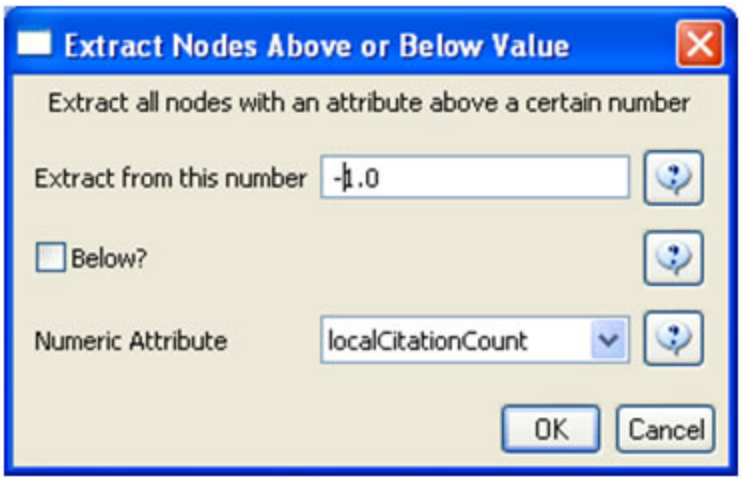


Figure 5.11: Directed, unweighted paper-paper citation network for 'FourNetSciResearchers' dataset with all papers and references in the GUESS user interface (left) and a pruned paper-paper citation network after removing all references and isolates (right)

The complete network can be reduced to papers that appeared in the original ISI file by deleting all nodes that have a GCC of -1. Simply run *'Preprocessing > Networks > Extract Nodes Above or Below Value'* with parameter values:



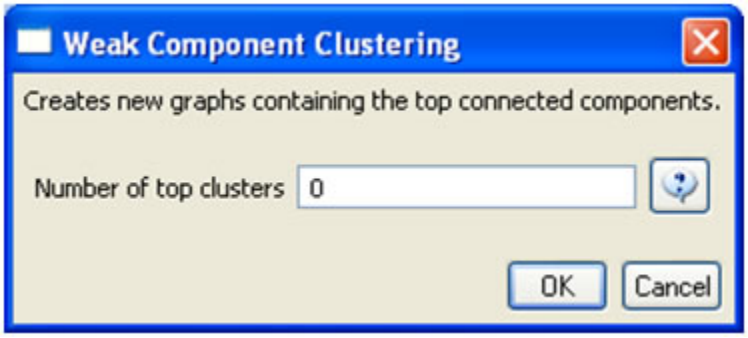
The resulting network is unconnected, i.e., it has many subnetworks many of which have only one node. These single unconnected nodes, also called isolates, can be removed using *'Preprocessing > Networks > Delete Isolates'*. Deleting isolates is a memory intensive procedure. If you experience problems at this step, refer to Section 3.4 Memory Allocation.

The *'FourNetSciResearchers'* dataset has exactly 65 isolates. Removing those leaves 12 networks shown in Figure 5.11 (right) using the same color and size coding as in Figure 5.11 (left). Using *'View > Information Window'* in GUESS reveals detailed information for any node or edge.
Alternatively, nodes could have been color and/or size coded by their degree using, e.g.:

```
> g.computeDegrees()
> colorize(outdegree,gray,black)
```

Note that the outdegree corresponds to the LCC within the given network while the indegree reflects the number of references, helping to visually identify review papers.

The complete paper-paper-citation network can be split into its subnetworks using *'Analysis > Networks > Unweighted & Directed > Weak Component Clustering'* with the default values:



The largest component has 2407 nodes; the second largest, 307; the third, 13; and the fourth has 7 nodes. The largest component is shown in Figure 5.12. The top 20 papers, by times cited in ISI, have been labeled using

```
> toptc = g.nodes[:]
> def bytc(n1, n2):
      return cmp(n1.globalcitationcount, n2.globalcitationcount)
> toptc.sort(bytc)
> toptc.reverse()
> toptc
> for i in range(0, 20):
      toptc[i].labelvisible = true
```

Alternatively, run '*Script > Run Script*' and select '***yoursci2directory****/scripts/GUESS/paper-citation-nw.py*'.

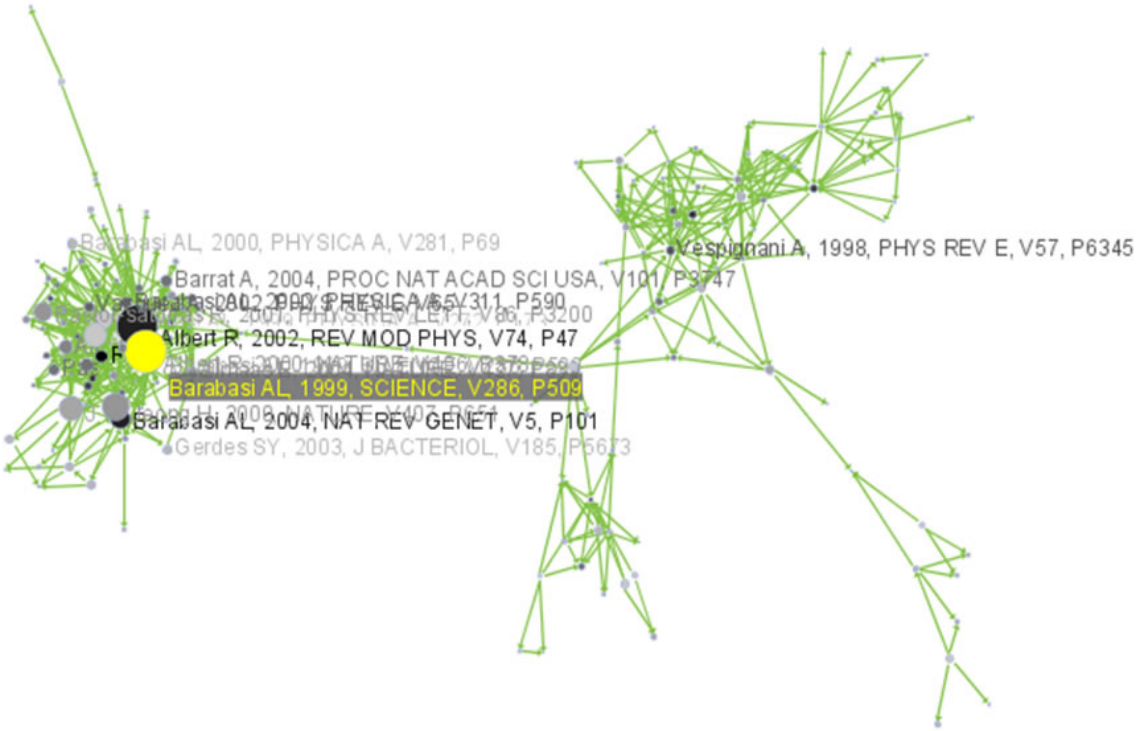


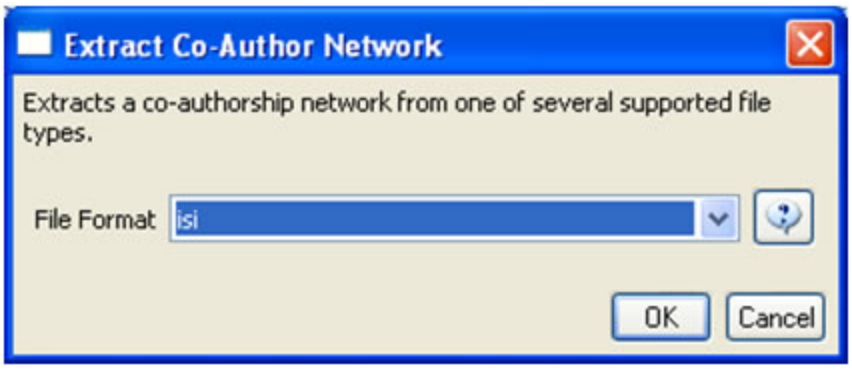
Figure 5.12: Giant components of the paper citation network

Compare the result with Figure 5.11 and note that this network layout algorithm – and most others – are non-deterministic: different runs lead to different layouts. That said, all layouts aim to group connected nodes into spatial proximity while avoiding overlaps of unconnected or sparsely connected sub-networks.

To see the log file from this workflow save the 5.1.4.1 Paper-Paper (Citation) Network log file.

5.1.4.2 Author Co-Occurrence (Co-Author) Network

To produce a co-authorship network in the Sci^2 Tool, select the table of all 361 unique ISI records from the *'FourNet SciResearchers'* dataset in the Data Manager window. Run *'Data Preparation > Extract Co-Author Network'* using the parameter:

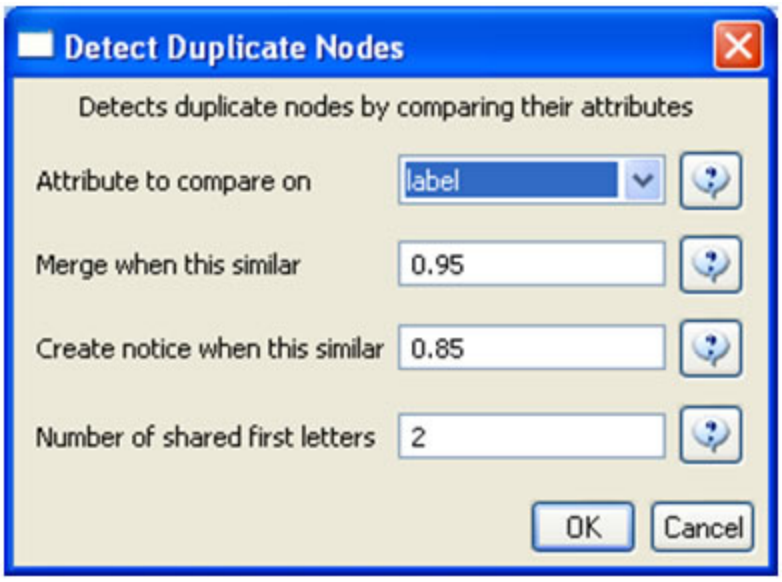


The result is two derived files in the Data Manager window: the "Extracted Co-Authorship Network" and an "Author information" table (also known as a "merge table"), which lists unique authors. In order to manually examine and edit the list of unique authors, open the merge table in your default spreadsheet program. In the spreadsheet, select all records, including "label," "timesCited," "numberOfWorks," "uniqueIndex," and "combineValues," and sort by "label." Identify names that refer to the same person. In order to merge two names, first delete the asterisk ('*') in the "combineValues" column of the duplicate node's row. Then, copy the "uniqueIndex" of the name that should be kept and paste it into the cell of the name that should be deleted. Resave the revised table as a .csv file and reload it. Select both the merge table and the network and run '*Data Preparation* > *Update Network by Merging Nodes*'. Table 5.2 shows the result of merging "Albet, R" and "Albert, R": "Albet, R" will be deleted and all of the node linkages and citation counts will be added to "Albert, R".

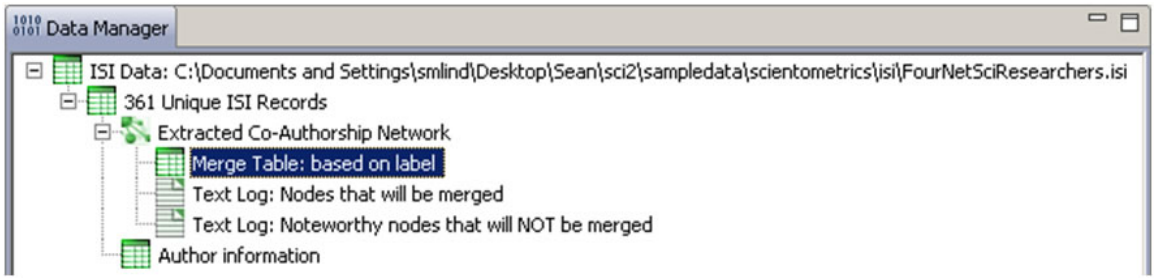
label	timesCited	numberOfWorks	uniqueIndex	combineValues
Abt, HA	3	1	142	*
Alava, M	26	1	196	*
Albert, R	7741	17	60	*
Albet, R	16	1	60	

Table 5.2: Merging of author nodes using the merge table

A merge table can be automatically generated by applying the Jaro distance metric (Jaro, 1989, 1995) available in the open source Similarity Measure Library (http://sourceforge.net/projects/simmetrics/) to identify potential duplicates. In the Sci^2 Tool, simply select the co-author network and run *'Data Preparation* > *Detect Duplicate Nodes '.* using the parameters:



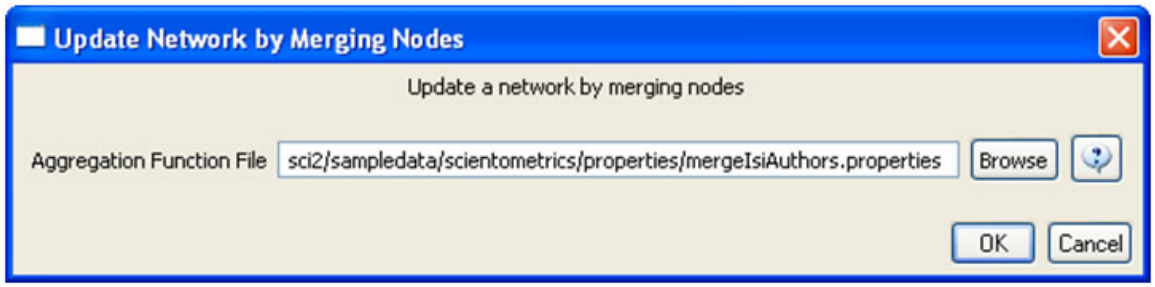
The result is a merge table that has the very same format as Table 5.2, together with two textual log files:



The log files describe, in a more human-readable form, which nodes will be merged or not merged. Specifically, the first log file provides information regarding which nodes will be merged, while the second log file lists nodes which are similar but will not be merged. The automatically generated merge table can be further modified as needed.

In sum, unification of author names can be done manually or automatically, independently or in conjunction with other data manipulation. It is recommended that users create the initial merge table automatically and fine-tune it as needed. Note that the same procedure can be used to identify duplicate references – simply select a paper-citation network and run *'Data Preparation* > *Detect Duplicate Nodes'* using the same parameters as above and a merge table for references will be created.

To merge identified duplicate nodes, select both the "Extracted Co-Authorship Network" and "Merge Table: based on label" by holding down the 'Ctrl' key. Run *'Data Preparation* > *Update Network by Merging Nodes'*. This will produce an updated network as well as a report describing which nodes were merged. To complete this workflow, an aggregation function file must also be selected from the pop-up window:



Aggregation files can be found in the Sci2 sample data. Follow this path *"sci2/sampledata/scientometrics/properties*"

and select the mergeIsiAuthors.properties.

The updated co-authorship network can be visualized using *'Visualization > Networks > GUESS'*, (See section 4.9.4.1 GUESS Visualizations for more information regarding GUESS).
Figure 5.13 shows the layout of the combined *'FourNetSciResearchers'* dataset after it was modified using the following commands in the "Interpreter":

```
> resizeLinear(numberOfWorks,1,50)
> colorize(numberOfWorks,gray,black)
> for n in g.nodes:
      n.strokecolor = n.color
> resizeLinear(numberOfCoAuthored_works, .25, 8)
> colorize(numberOfCoAuthoredworks, "127,193,65,255", black)
> nodesbynumworks = g.nodes[:]
> def bynumworks(n1, n2):
     return cmp(n1.numberofworks, n2.numberofworks)
> nodesbynumworks.sort(bynumworks)
> nodesbynumworks.reverse()
> for i in range(0, 50):
      nodesbynumworks[i].labelvisible = true
```

Alternatively, run '*Script > Run Script ...*' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*'.

For both workflows described above, the final step should be to run *'Layout > GEM'* and then *'Layout > Bin Pack'* to give a better representation of node clustering.

In the resulting visualization, author nodes are color and size coded by the number of papers per author. Edges are color and thickness coded by the number of times two authors wrote a paper together. The remaining commands identify the top 50 authors with the most papers and make their name labels visible.

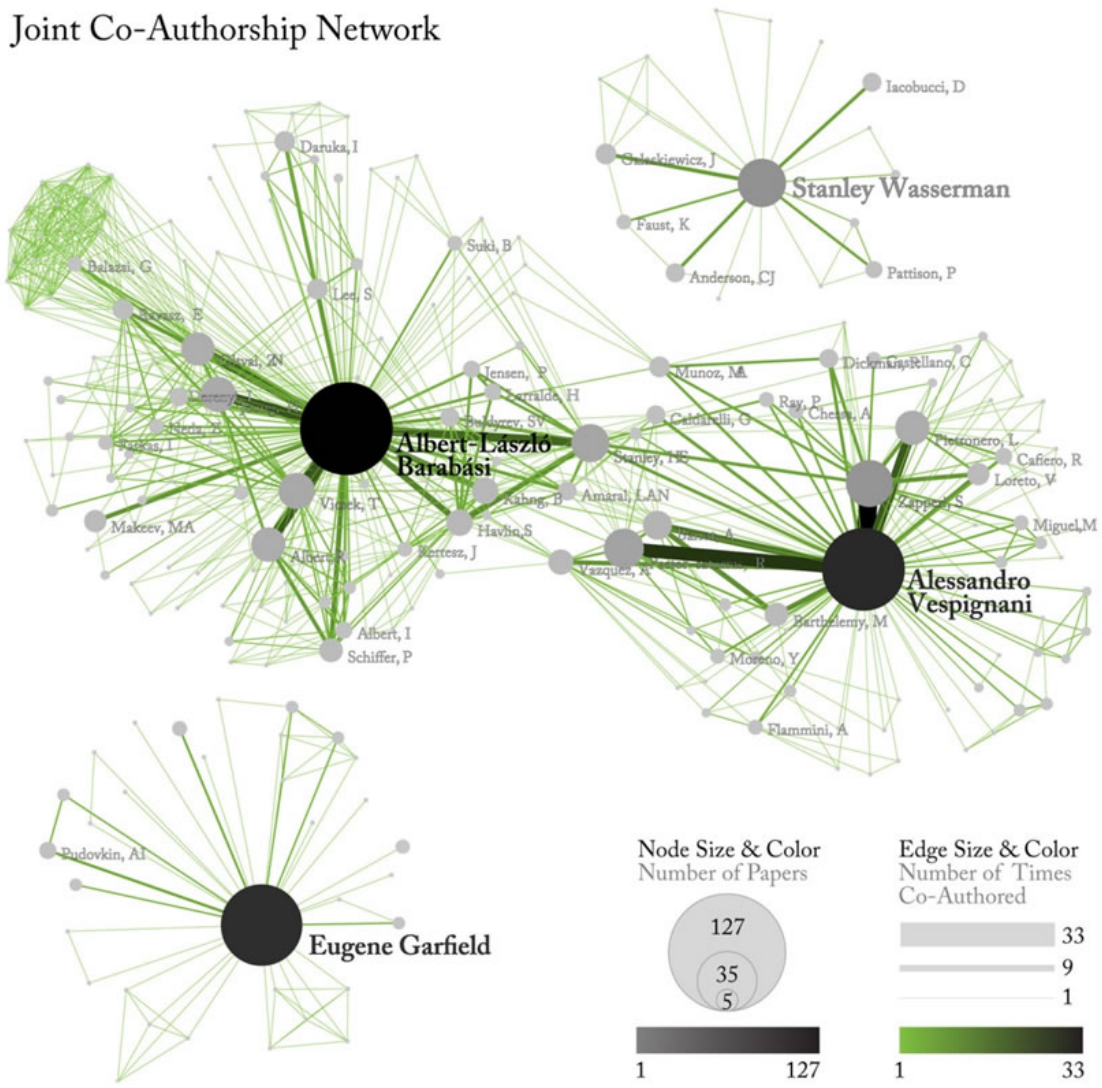


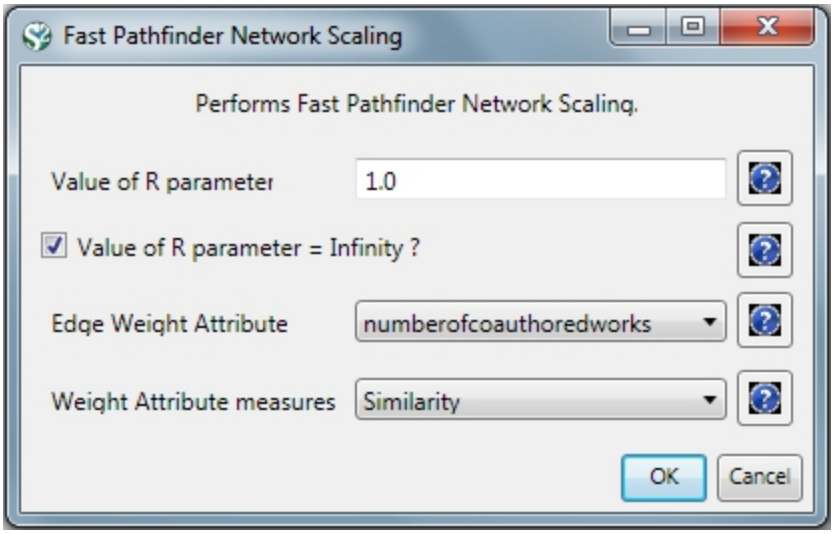
Figure 5.13: Undirected, weighted co-author network for 'FourNetSciResearchers' dataset

GUESS supports the repositioning of selected nodes. Multiple nodes can be selected by holding down the 'Shift' key and dragging a box around specific nodes. The final network can be saved via *'GUESS: File > Export Image'* and opened in a graphic design program to add a title and legend and to modify label sizes. The image above was modified using Photoshop.

Pruning the network

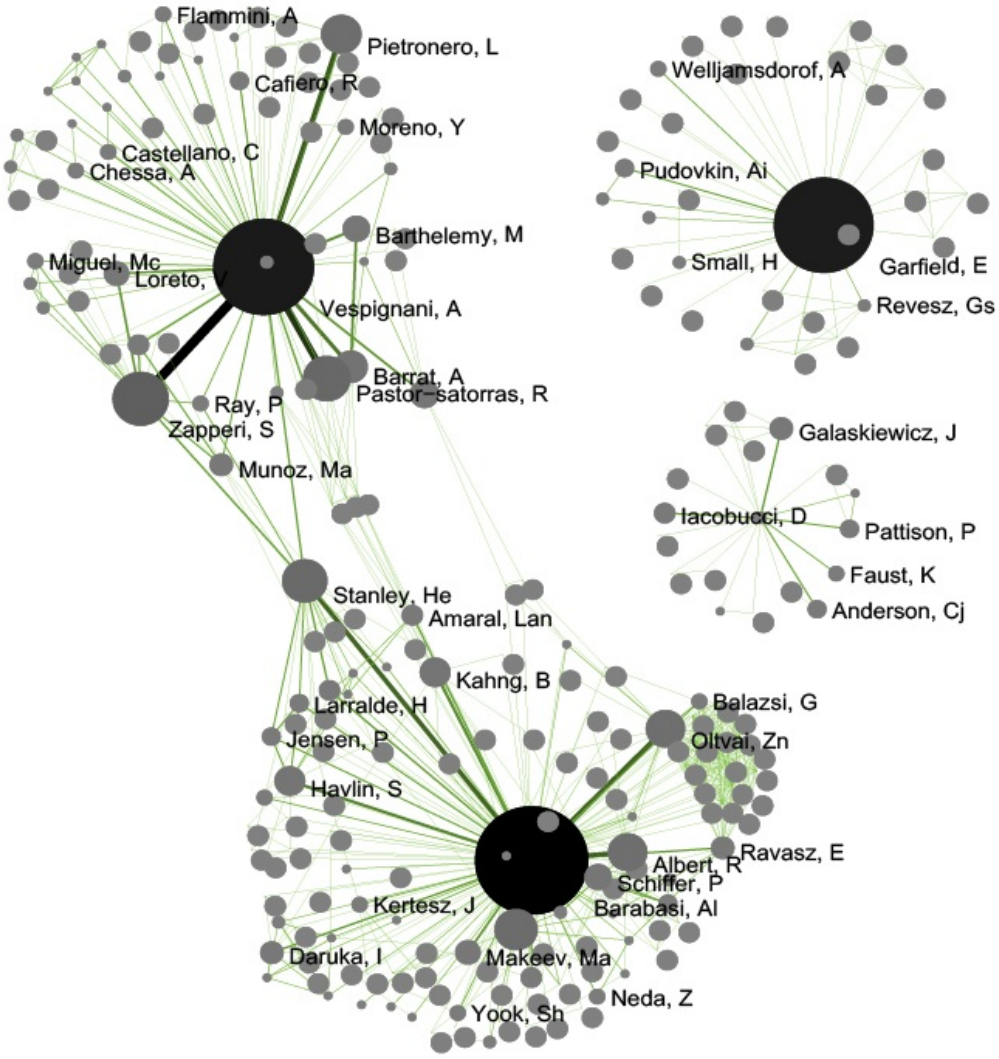
Pathfinder algorithms take estimates of the proximity's between pairs of items as input and define a network representation of the items that preserves only the most important links. The value of Pathfinder Network Scaling for networks lies in its ability to reduce the number of links in meaningful ways, often resulting in a concise representation of clarified proximity patterns.

Given the co-Authorship network extracted at workflow 5.1.4.2, run '*Preprocessing > Networks > Fast Pathfinder Network Scaling' with the following parameters:*

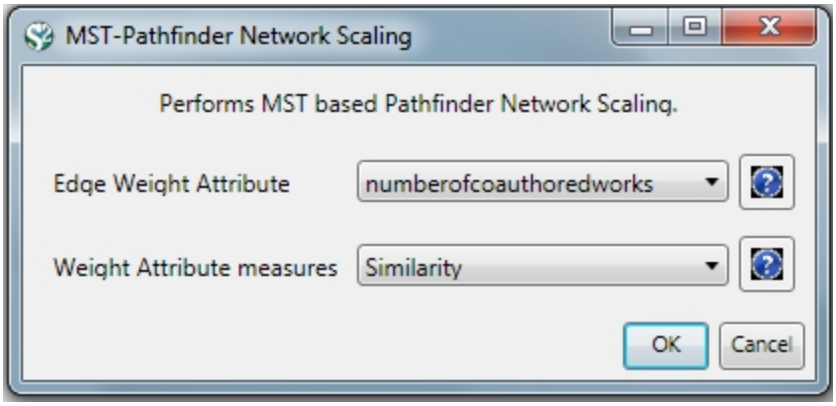


The number of edges reduced from 872 to 731. The updated co-authorship network can be visualized using *'Visualization > Networks >* GUESS*'*.

Inside GUESS, run 'Script > Run Script ...' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*'. Also run *'Layout > GEM'* and then *'Layout > Bin Pack'* to give a better representation of node clustering. The pruned network looks like this:

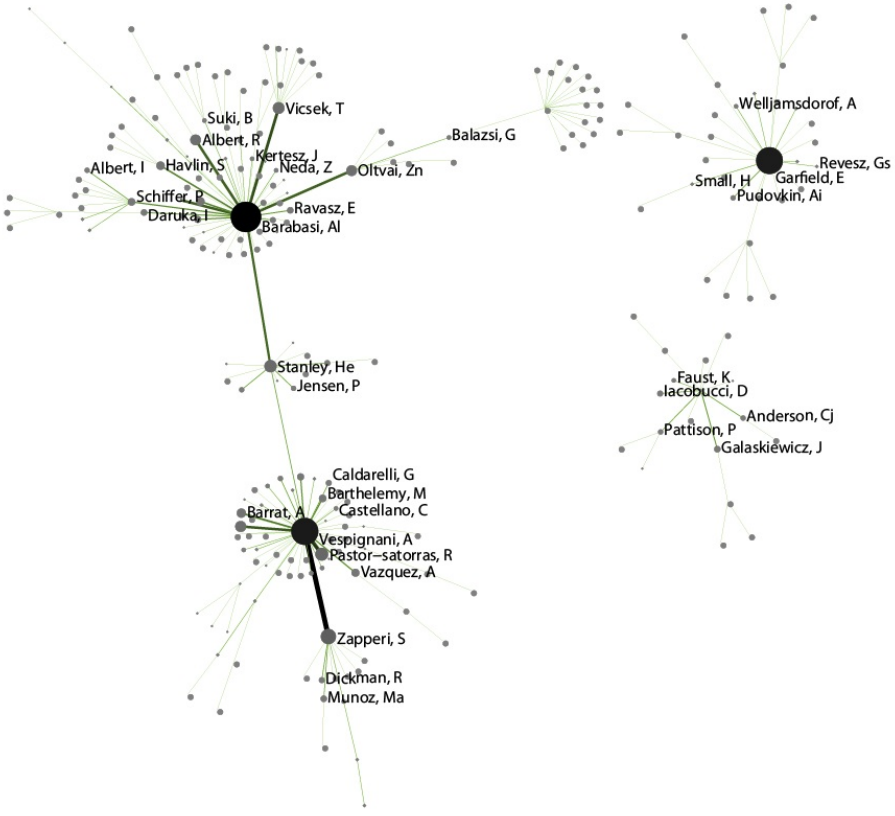


If the network being processed is undirected, which is the case, then MST-Pathfinder Network Scaling can be used to prune the networks. This will give results in 30 times faster than Fast Pathfinder Network Scaling. Also we have found that networks that have a low standard deviation for edge weights or if many of the edge weights are equal to the minimum edge weight, then the network might not be scaled as much as expected when using Fast Pathfinder Network Scaling. To see this behavior, run '*Preprocessing > Networks > MST-Pathfinder Network Scaling*' with the network named 'Updated network' selected with the following parameters:



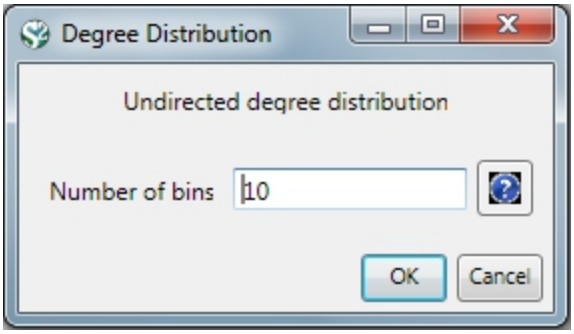
The number of edges reduced from 872 to 235 for the network pruned with the MST-Pathfinder Network Scaling alg orithm, while the reduction for the Fast Pathfinder Network Scaling algorithm was from 872 to 731 only.

Visualize this network using *'Visualization > Networks >* GUESS*'*. Inside GUESS, run 'Script > Run Script ...' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*' again. Also run *'Layout > GEM'* and then *'Layout > Bin Pack'* to give a better representation of node clustering. The pruned network with the algorithm looks like this:



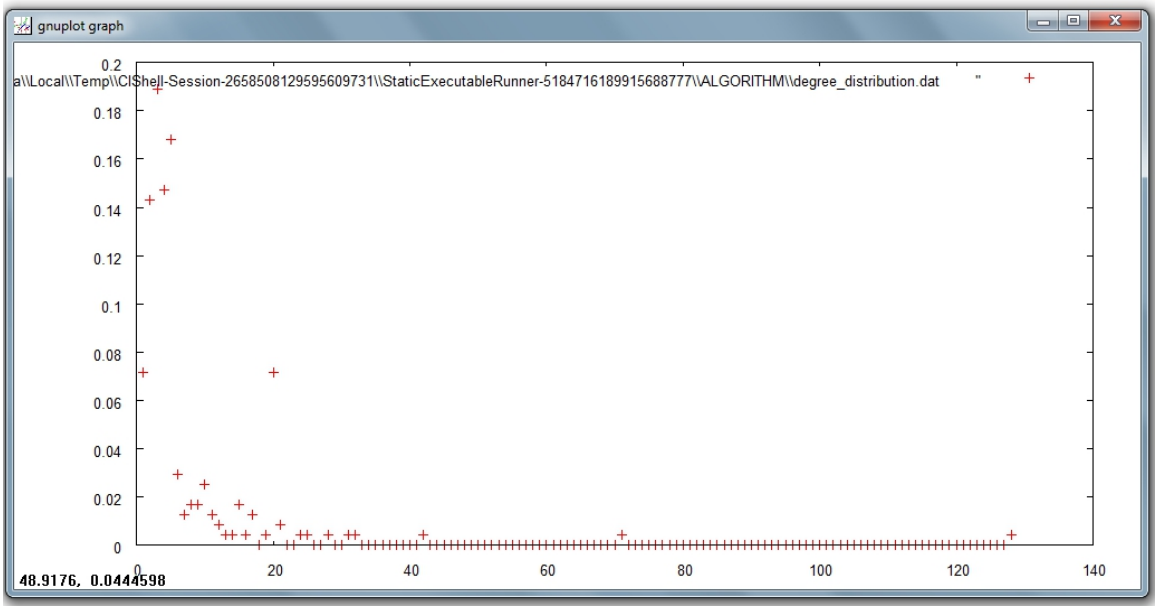
Degree Distribution

Given the co-Authorship network extracted at workflow 5.1.4.2 (name 'Updated Network' at the Data Manager), run ' *Analysis > Networks > Unweighted & Undirected >* Degree Distribution*' with the following parameter:*



This algorithm generates two output files, corresponding to two different ways of partitioning the interval spanned by the values of degree.

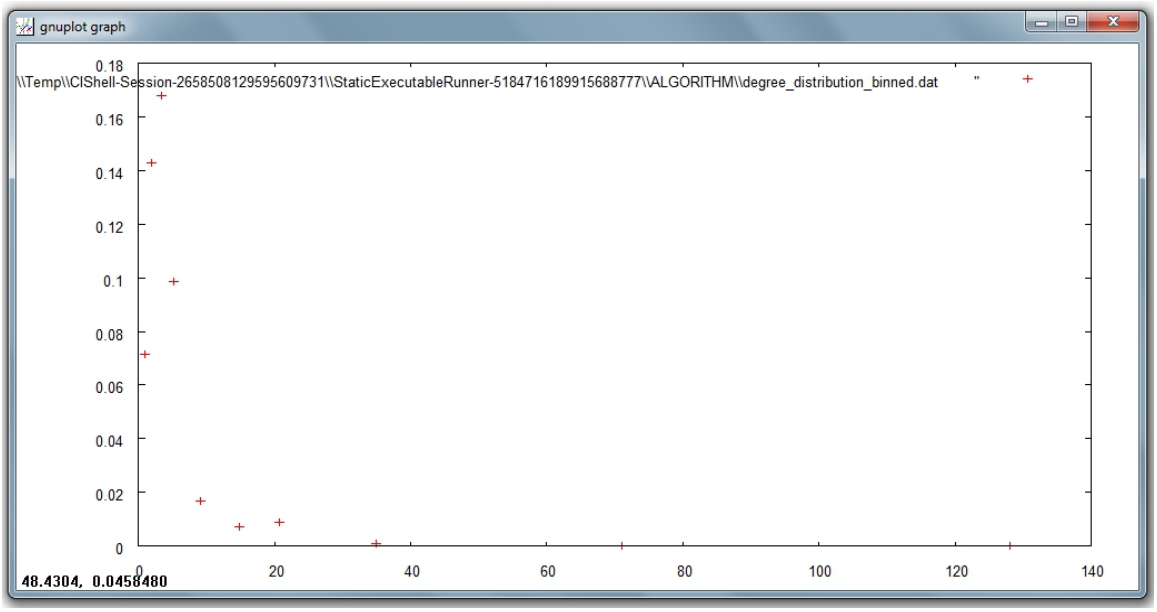
In the first output file named '*Distribution of degree for network at study (equal bins)*', the occurrence of any degree value between the minimum and the maximum is estimated and divided by the number of nodes of the network, so to obtain the probability: the output displays all degree values in the interval with their probabilities. Visualize this first output file with '*Visualization* > *General* > *GnuPlot*':



The second output file named '*Distribution of degree for network at study (logarithmic bins)*' gives the *binned* distribu tion, i.e. the interval spanned by the values of degree is divided into bins whose size grows while going to higher values of the variable. The size of each bin is obtained by multiplying by a fixed number the size of the previous bin. The program calculates the fraction of nodes whose degree falls within each bin. Because of the different sizes of the bins, these fractions must be divided by the respective bin size, to have meaningful averages.

This second type of output file is particularly suitable to study skewed distributions: the fact that the size of the bins grows large for large degree values compensates for the fact that not many nodes have high degree values, so it suppresses the fluctuations that one would observe by using bins of equal size. On a double logarithmic scale, which is very useful to determine the possible power law behavior of the distribution, the points of the latter will appear equally spaced on the x-axis.

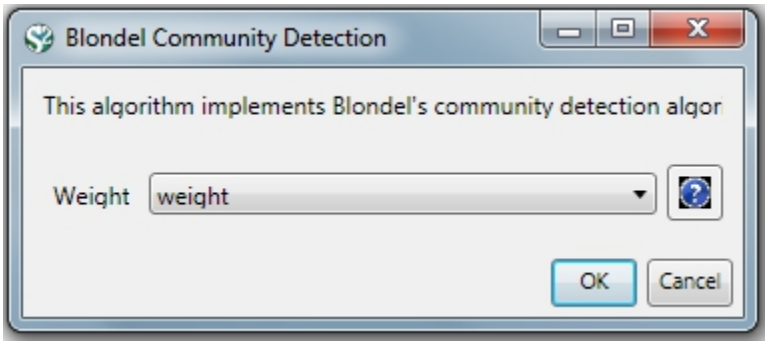
Visualize also this second output file with '*Visualization* > *General* > *GnuPlot*':



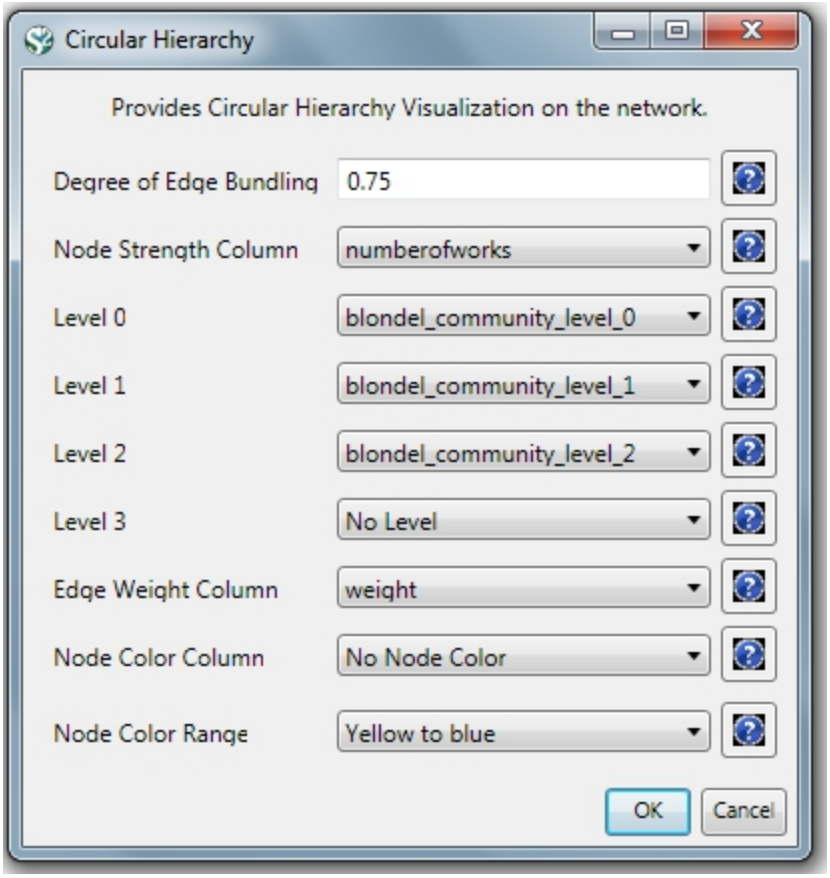
Community Detection

Community Detection algorithms look for subgraphs where nodes are highly interconnected among themselves and poorly connected with nodes outside the subgraph. Many community detection algorithms are based on the optimization of the modularity - a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities. The Blondel Community Detection finds high modularity partitions of large networks in short time and that unfolds a complete hierarchical community structure for the network, thereby giving access to different resolutions of community detection.

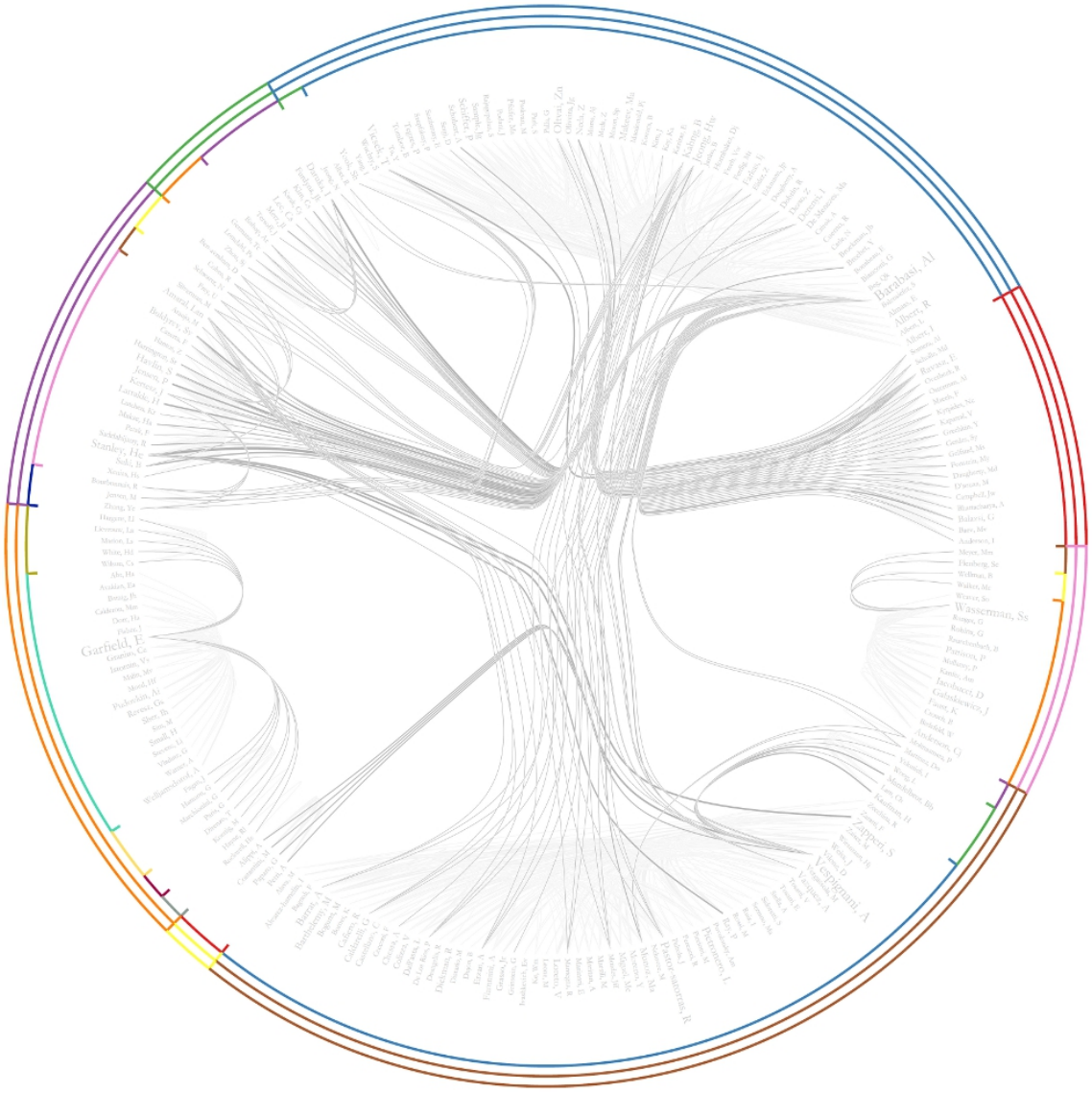
Run 'Analysis > Networks > Weighted & Undirected > Blondel Community Detection' with the following parameter:



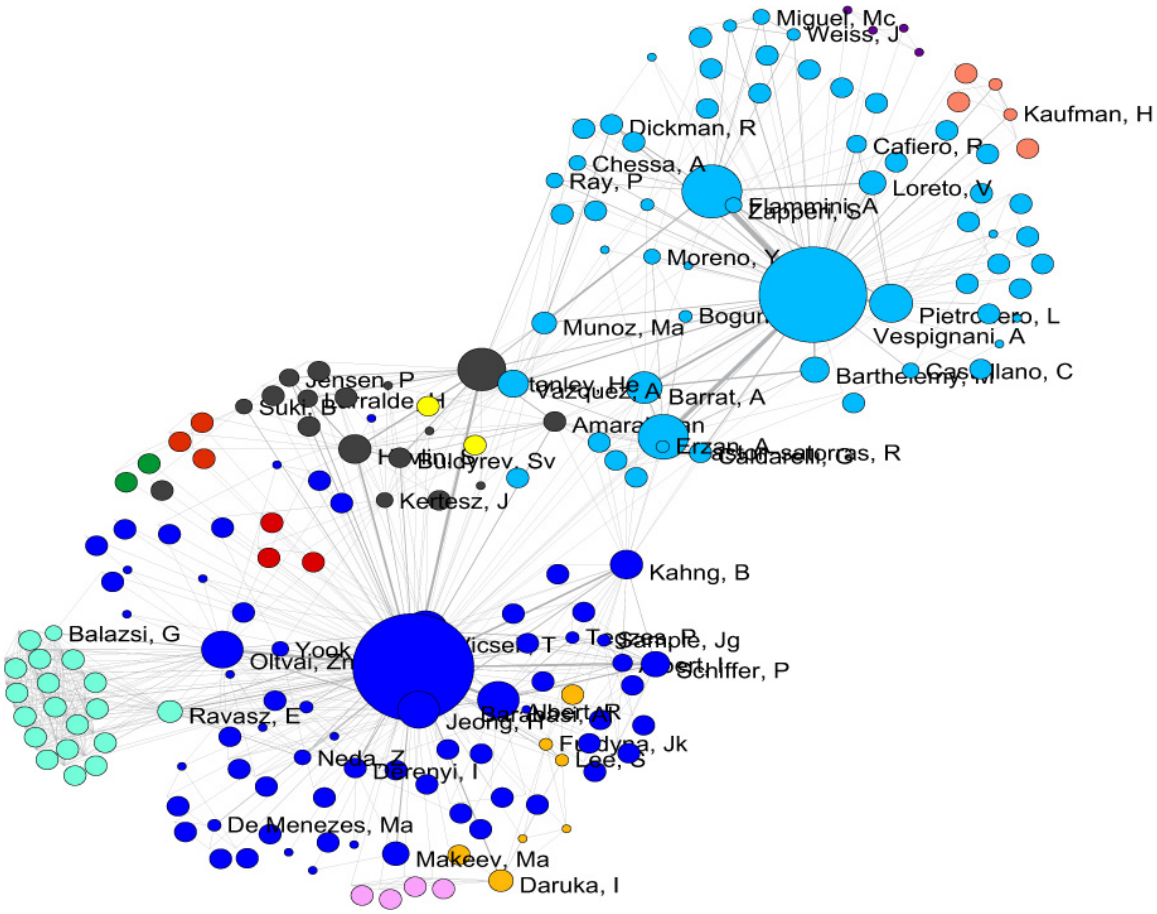
This will generate a network named 'With community attributes' in the Data Manager. To visualize this network run 'Visualization > Networks > Circular Hierarchy' with the following parameters:



The generated postscript file "CircularHierarchy_With community attributes.ps" can be viewed using Adobe Distiller or GhostViewer (see section 2.4 Saving Visualizations for Publication).



To view to the network with community attributes with in GUESS select the network used to generate the image above and run *'Visualization > Networks >* GUESS.'



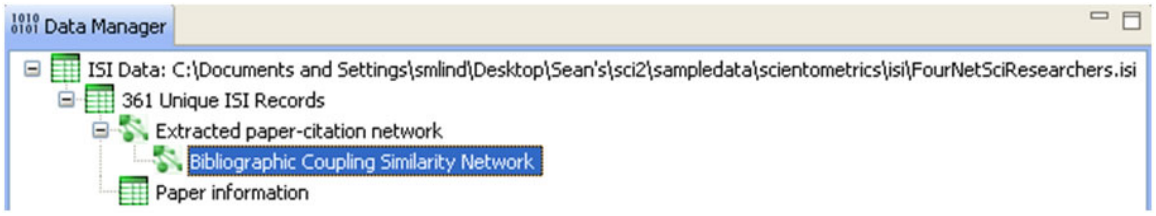
To see the log file from this workflow save the 5.1.4.2 Author Co-Occurrence (Co-Author) Network log file.

5.1.4.3 Cited Reference Co-Occurrence (Bibliographic Coupling) Network

In Sci^2, a bibliographic coupling network is derived from a directed paper citation network (see section 4.9.1.1.1 Document-Document (Citation) Network).

Load the file using '*File > Load*' and following this path: *'yoursci2directory/sampledata/scientometrics/isi/FourNetSci Researchers.isi*'. A table of all records and a table of 361 records with unique ISI ids will appear in the Data Manager.

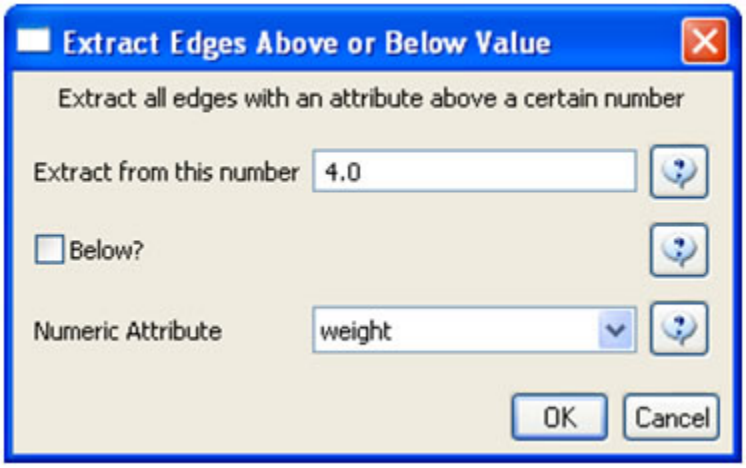
Select the "361 Unique ISI Records" in the Data Manager and run *'Data Preparation > Extract Paper Citation Network.'* Select "Extracted Paper Citation Network" and run *'Data Preparation > Extract Reference Co-Occurrence (Bibliographic Coupling) Network.'*



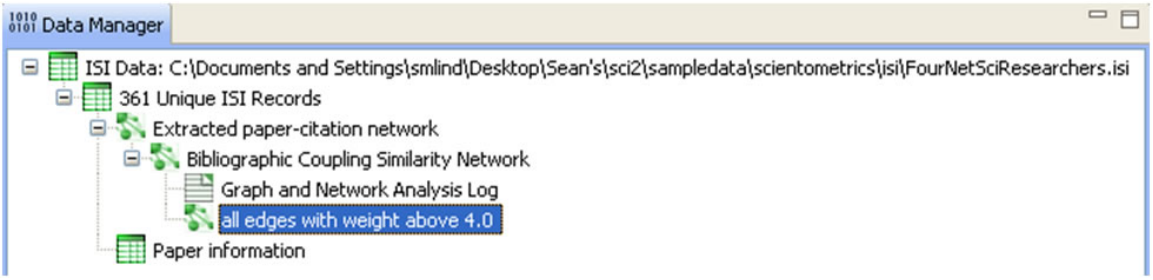
Running *'Analysis > Networks > Network Analysis Toolkit (NAT)'* reveals that the network has 5,342 nodes (5,013 of which are isolate nodes) and 6,277 edges.

In the "Bibliographic Coupling Similarity Network," edges with low weights can be eliminated by running *'Preprocessi*

ng > Networks > Extract Edges Above or Below Value' with the following parameter values:



to produce the following network:



Isolate nodes can be removed running '*Preprocessing > Networks > Delete Isolates*'. The resulting network has 242 nodes and 1,534 edges in 12 weakly connected components.

This network can be visualized in GUESS; see Figure 5.14. Nodes and edges can be color and size coded, and the top 20 most-cited papers can be labeled by entering the following lines in the GUESS "Interpreter":

```
> resizeLinear(globalcitationcount,2,40)
> colorize(globalcitationcount,gray,black)
> resizeLinear(weight,.25,8)
> colorize(weight, "127,193,65,255", black)
> for n in g.nodes:
      n.strokecolor=n.color
> toptc = g.nodes[:]
> def bytc(n1, n2):
      return cmp(n1.globalcitationcount, n2.globalcitationcount)
> toptc.sort(bytc)
> toptc.reverse()
> toptc
> for i in range(0, 20):
      toptc[i].labelvisible = true
```

Alternatively, run 'GUESS: File > Run Script ...' and select '***yoursci2directory****/scripts/GUESS/reference-co-occurence-nw.py*'.

For both workflows described above, the final step should be to run 'Layout > GEM' and then 'Layout > Bin Pack' to give a better representation of node clustering.

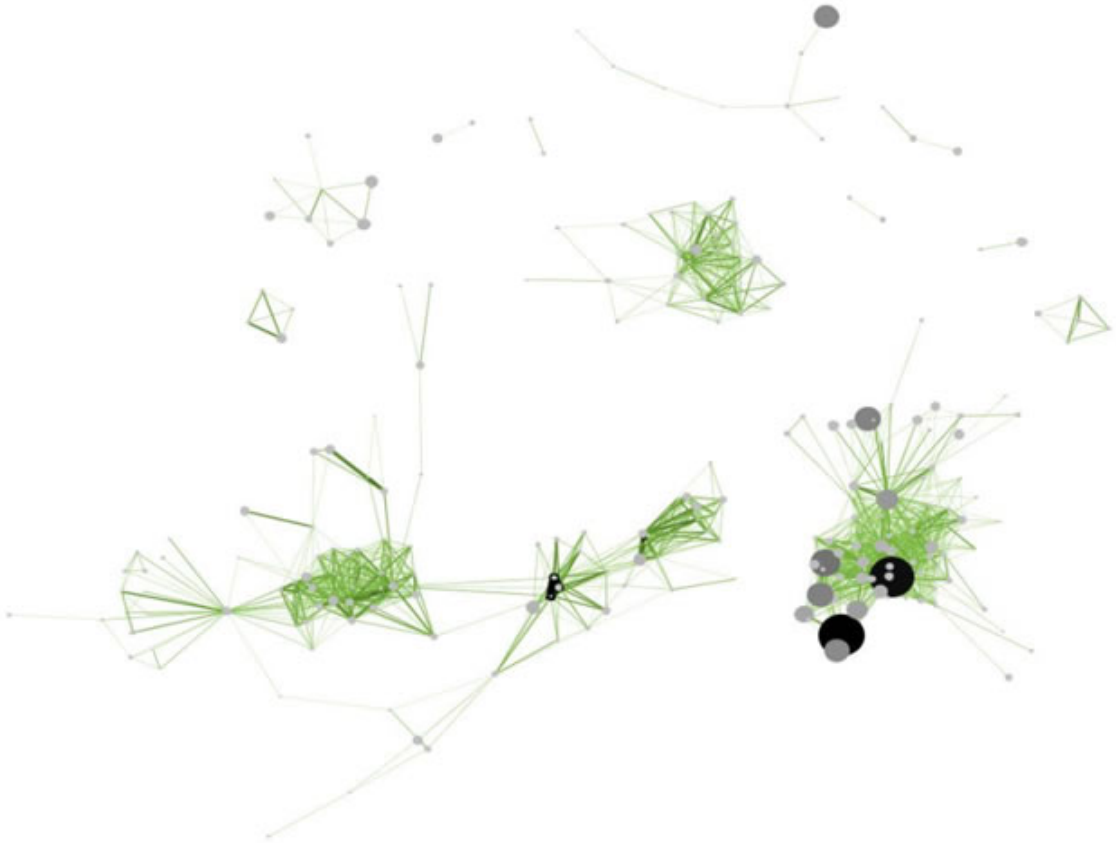


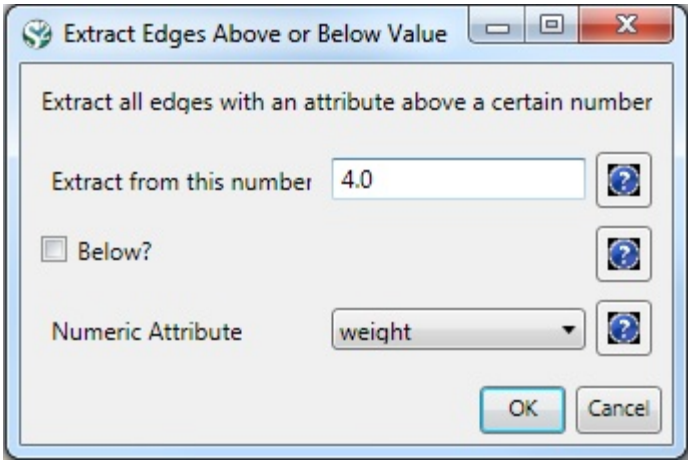
Figure 5.14: Reference co-occurrence network layout for 'FourNetSciResearchers' dataset

To see the log file from this workflow save the 5.1.4.3 Cited Reference Co-Occurrence (Bibliographic Coupling) Network log file.

5.1.4.4 Document Co-Citation Network (DCA)

Load the file using '*Load > File*" and following this path: *'yoursci2directory/sampledata/scientometrics/isi/FourNetSciResearchers.isi*'. A table of all records and a table of 361 records with unique ISI ids will appear in the Data Manager.

Select the "361 Unique ISI Records" and run '*Data Preparation > Extract Paper Citation Network'*. Select the 'Extracted Paper-Citation Network' and run '*Data Preparation > Extract Document Co-Citation Network*.' The co-citation network will have 5,335 nodes (213 of which are isolates) and 193,039 edges. Isolates can be removed by running *'Preprocessing > Networks > Delete Isolates.'* The resulting network has 5122 nodes and 193,039 edges – and is too dense for display in GUESS. Edges with low weights can be eliminated by running '*Preprocessing > Networks > Extract Edges Above or Below Value*' with parameter values:



Here, only edges with a local co-citation count of five or higher are kept. The giant component in the resulting network has 265 nodes and 1,607 edges. All other components have only one or two nodes.

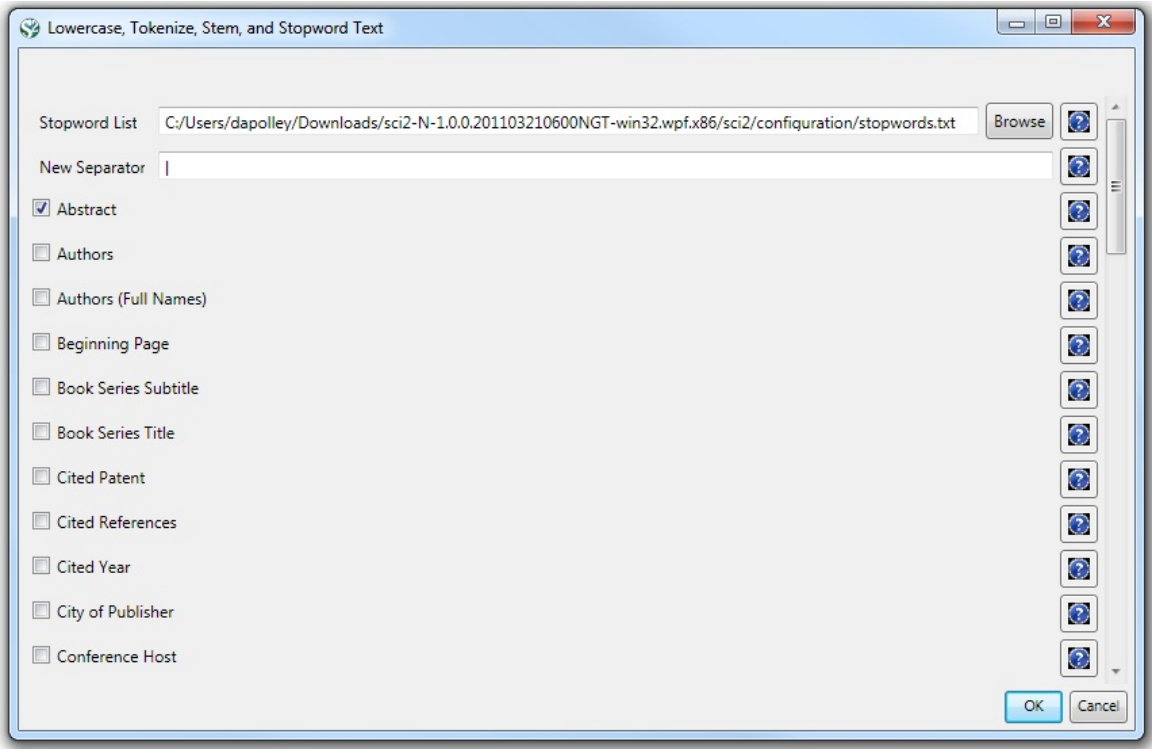
The giant component can be visualized in GUESS, see Figure 5.15 (right); see the above explanation, and use the same size and color coding and labeling as the bibliographic coupling network. Simply run *'GUESS: File > Run Script ...'* and select '***yoursci2directory****/scripts/GUESS/reference-co-occurence-nw.py*'

Figure 5.15: Undirected, weighted bibliographic coupling network (left) and undirected, weighted co-citation network (right) of 'FourNetSciResearchers' dataset, with isolate nodes removed

To see the log file from this workflow save the 5.1.4.4 Document Co-Citation Network (DCA) log file.

5.1.4.5 Word Co-Occurrence Network

In the Sci^2 Tool, select "361 unique ISI Records" from the *'FourNetSciResearchers'* dataset in the Data Manager. Run '*Preprocessing > Topical >* *Lowercase, Tokenize, Stem, and Stopword Text*' using the following parameters:

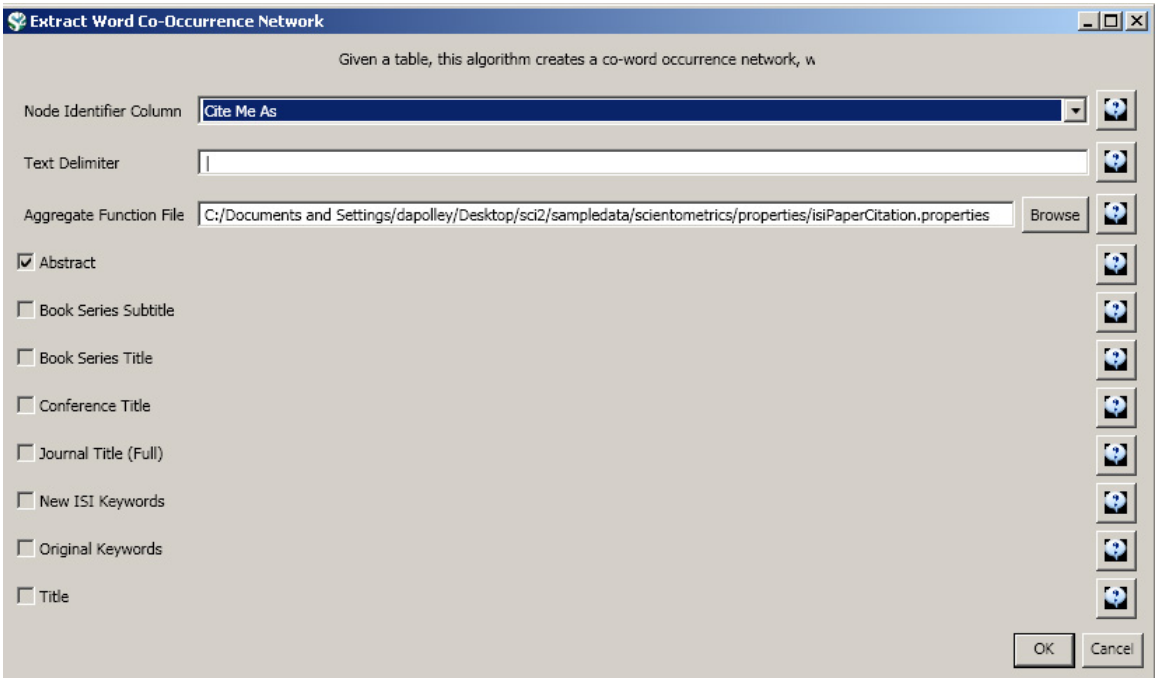


Text normalization utilizes the Standard Analyzer provided by Lucene (http://lucene.apache.org|). It separates text into word tokens, normalizes word tokens to lower case, removes "s" from the end of words, removes dots from acronyms, deletes stop words, and applies the English Snowball stemmer (http://snowball.tartarus.org/algorithms/english/stemmer.html), which is a version of the Porter2 stemmer designed for the English language..

The result is a derived table – "with normalized Abstract" – in which the text in the abstract column is normalized. Select this table and run '*Data Preparation* > *Extract Word Co-Occurrence Network*' using parameters:

Aggregate Function File

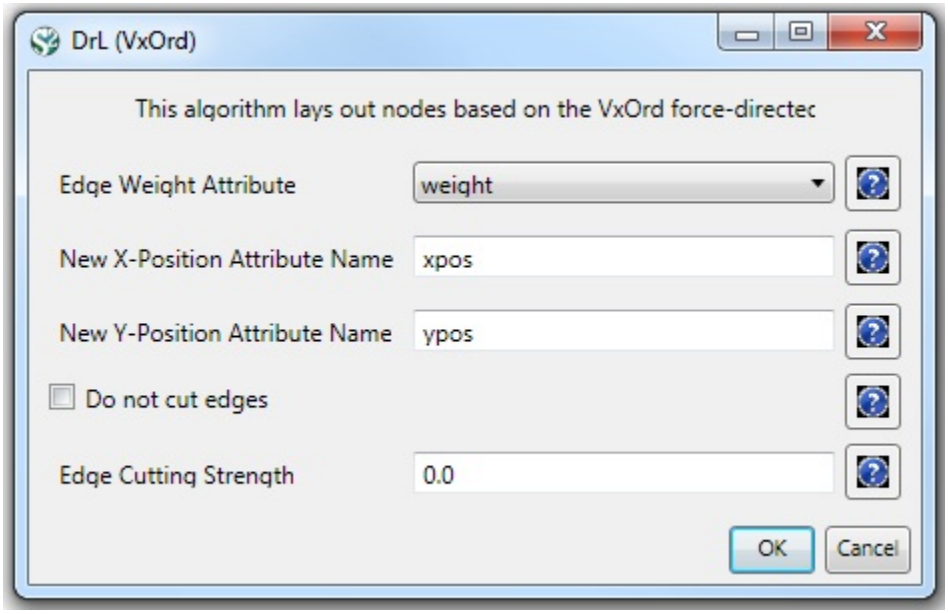
Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



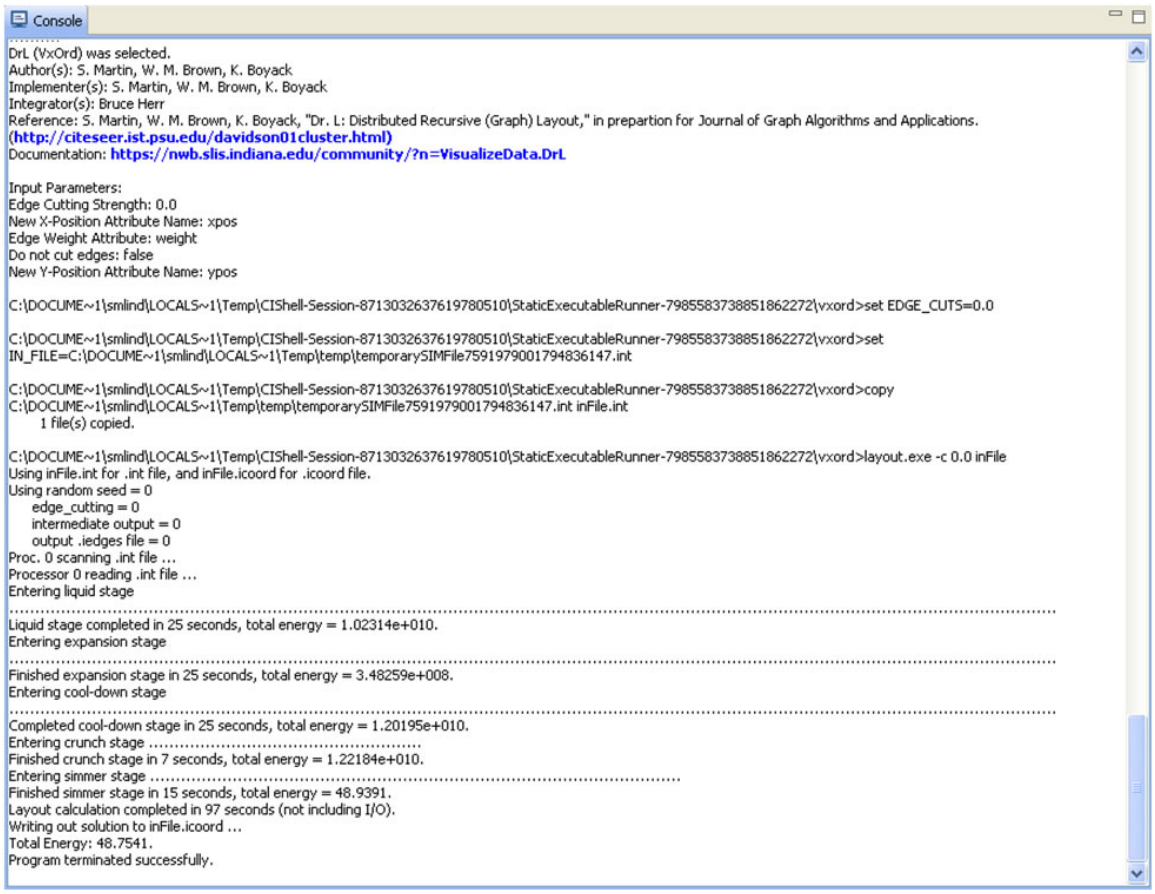
The outcome is a network in which nodes represent words and edges and denote their joint appearance in a paper. Word co-occurrence networks are rather large and dense. Running the *'Analysis > Networks > Network Analysis Toolkit (NAT)'* reveals that the network has 2,821 word nodes and 242,385 co-occurrence edges.

There are 354 isolated nodes that can be removed by running '*Preprocessing > Networks > Delete Isolates'* on the Co-Word Occurrence network. Note that when isolates are removed, papers without abstracts are removed along with the keywords.

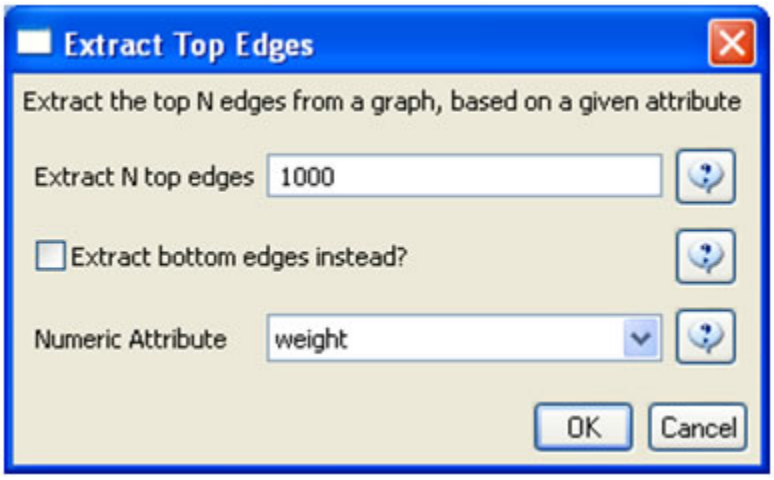
The result is one giant component with 2,467 nodes and 242,385 edges. To visualize this rather large network, begin by running *'Visualization > Networks > DrL (VxOrd)'* with default values:



Note that the DrL algorithm requires extensive data processing and requires a bit of time to run, even on powerful systems. See the console window for details:



To keep only the strongest edges in the "Laid out with DrL" network, run '*Preprocessing > Networks >* [*Extract Top Edges*]*'* on the new network using the following parameters:



Once edges have been removed, the network "top 1000 edges by weight" can be visualized by running *'Visualizatio n > Networks >* [*GUESS*]*'*. In GUESS, run the following commands in the Interpreter:

```
> for node in g.nodes:
      node.x = node.xpos * 40
      node.y = node.ypos * 40
> resizeLinear(references, 2, 40)
> colorize(references,[200,200,200],[0,0,0])
> resizeLinear(weight, .1, 2)
> g.edges.color = "127,193,65,255"
```

The result should look something like Figure 5.16.

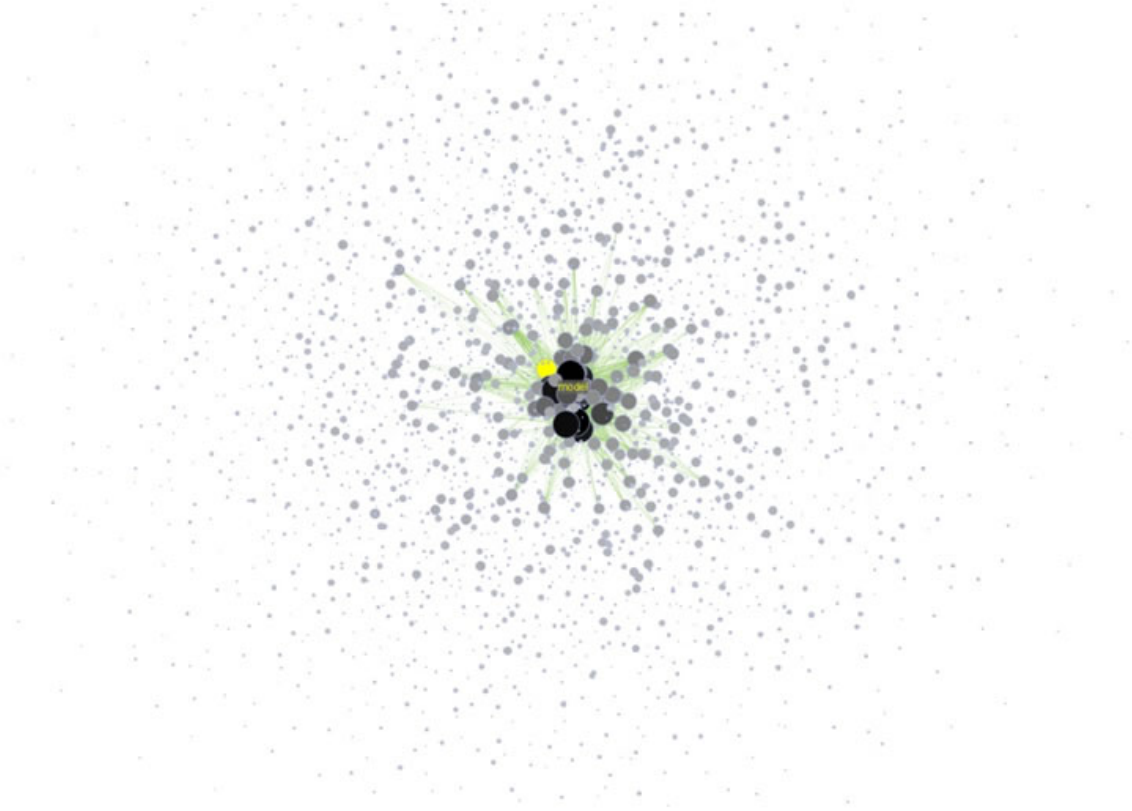


Figure 5.16: Undirected, weighted word co-occurrence network visualization for the DrL-processed 'FourNetSciResearchers' dataset

Currently, when you resize large networks in the GUESS visualization tool, the network visualizations can become "uncentered" in the display window. Running '*View* > *Center*' does not solve this problem. Users should zoom out to find the visualization, center it, and then zoom back in.

Note that only the top 1000 edges (by weight) in this large network appear in the above visualization, creating the impression of isolate nodes. To remove nodes that are not connected by the top 1000 edges (by weight), run '*Prepr ocessing* > *Networks* > *Delete Isolates*' on the "top 1000 edges by weight" network and visualize the result using the workflow described above.

To see the log file from this workflow save the 5.1.4.5 Word Co-Occurrence Network log file.

Database Extractions

Extended Version

This workflow uses the extended version of the Sci^2 Tool. To know how to extend Sci^2 view Section 3.2 Additional Plugins.

The Sci^2 Tool supports the creation of databases from ISI files. Database loading improves the speed and functionality of data preparation and preprocessing. While the initial loading can take quite some time for larger datasets (see sections 3.4 Memory Allocation and 3.5 Memory Limits) it results in vastly faster and more powerful data processing and extraction.

Once again load *'**yoursci2directory**/sampledata/scientometrics/isi/FourNetSciResearchers.isi'*, this time using '*File > Load*' instead of '*File > Load.*' Choose 'ISI database' from the load window. Right-click to view the database schema:

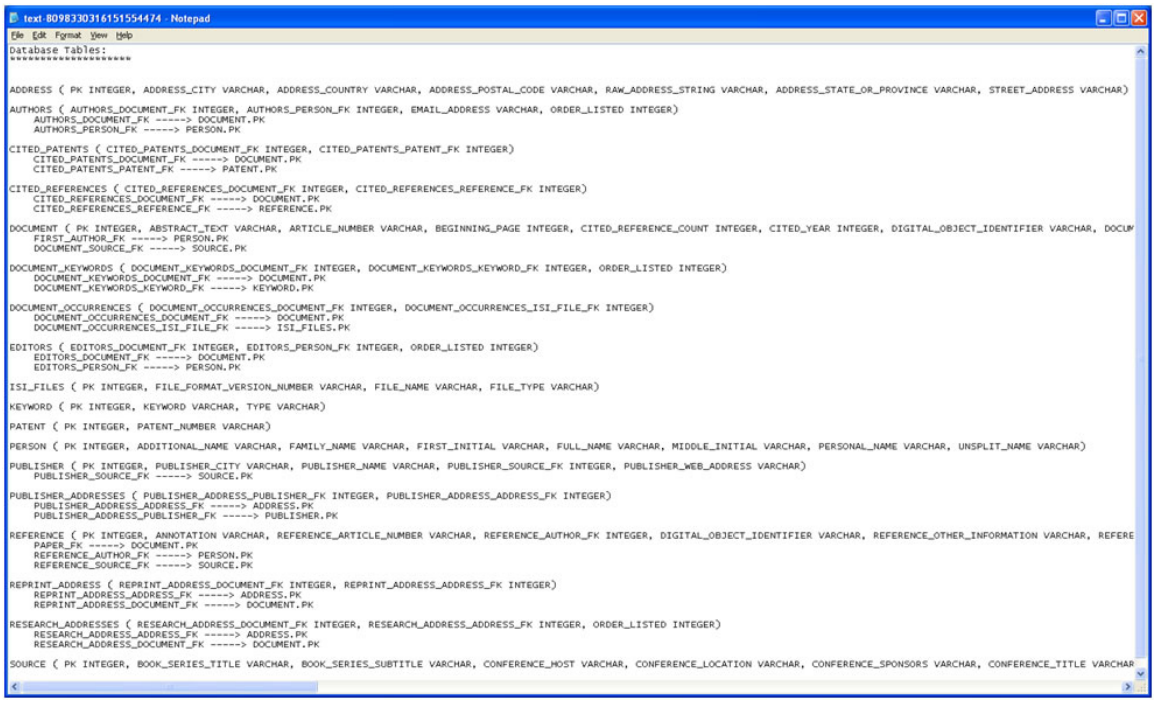


Figure 5.17: The database schema as viewed in Notepad

As before, it is important to clean the database before running any extractions by merging and matching authors, journals, and references. Run '*Data Preparation > Database > ISI > Merge Identical ISI People*', followed by '*Data Preparation > Database > ISI > Merge Document Sources''* and '*Data Preparation > Database > ISI > Match References to Papers*'. Make sure to wait until each cleaning step is complete before beginning the next one.

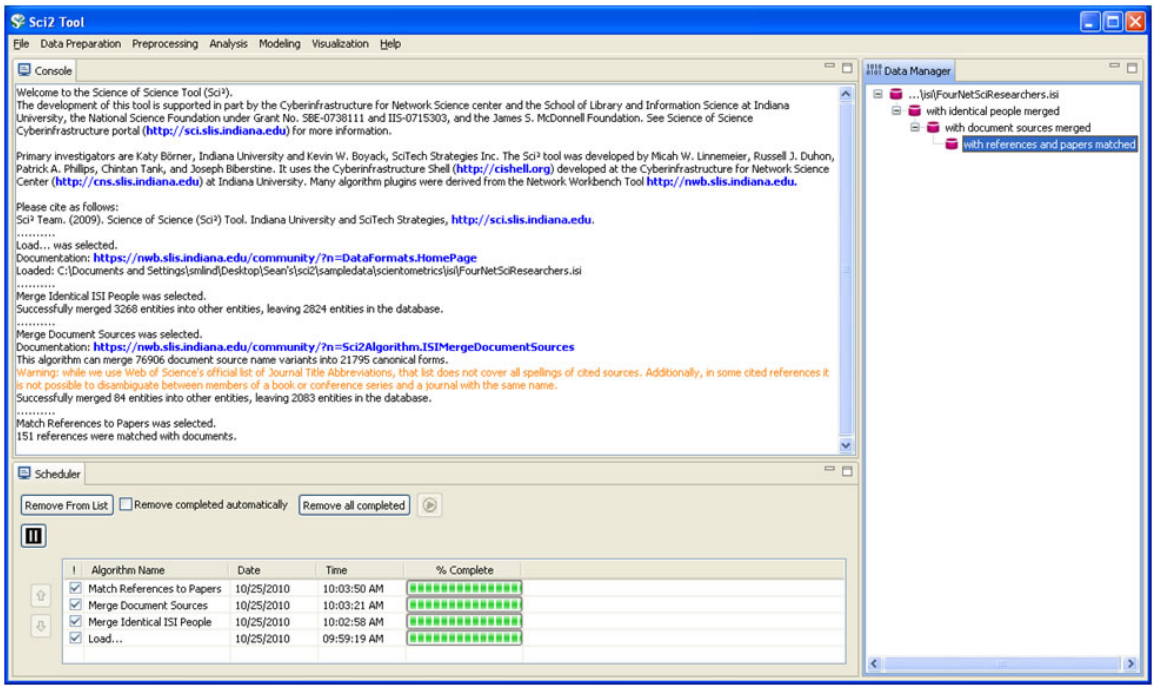
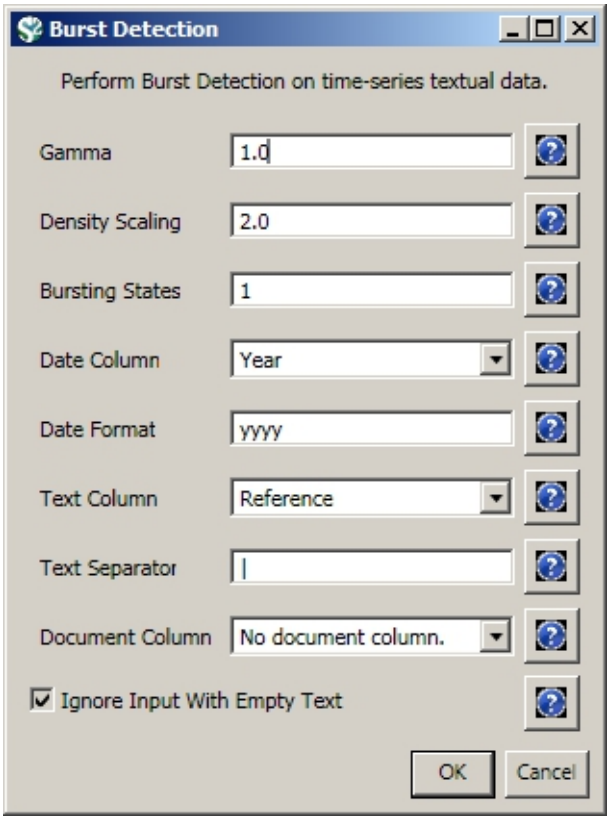


Figure 5.18: Cleaned database of 'FourNetSciResearchers'

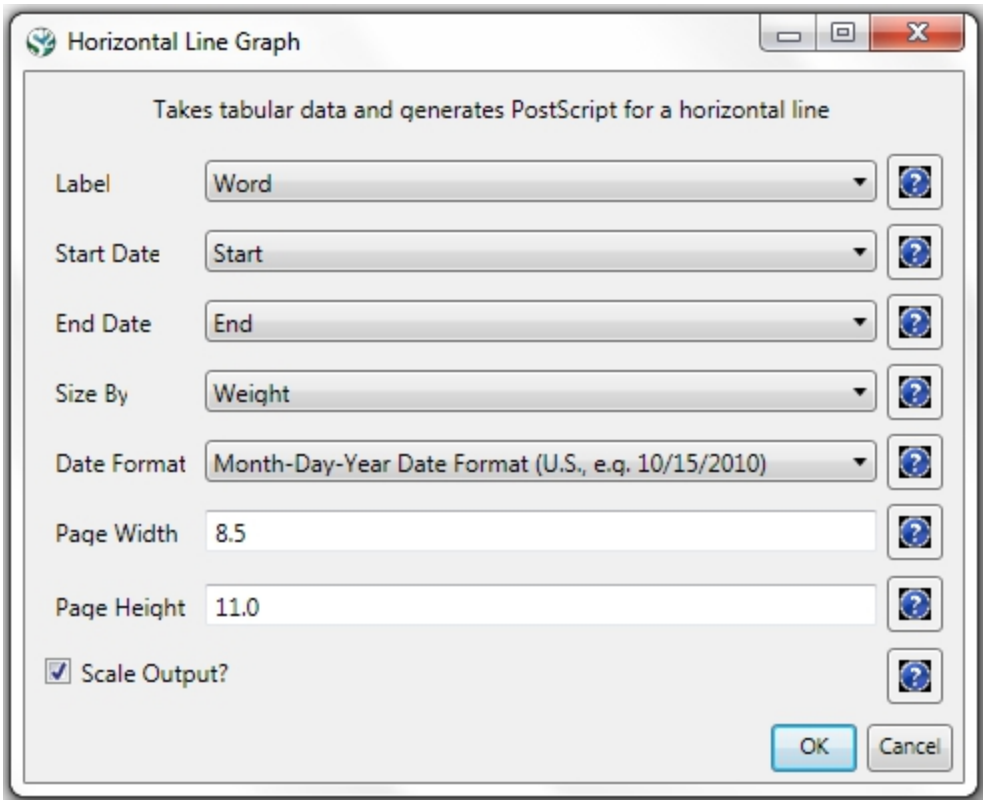
Extracting different tables will provide different views of the data. Run '*Data Preparation* > *Database* > *ISI* > *Extract Authors*' to view all the authors from FourNetSciResearchers.isi. The table includes the number of papers each person in the dataset authored, their Global Citation Count (how many times they have been cited according to ISI), and their Local Citation Count (how many times they were cited in the current dataset.)
The queries can also output data specifically tailored for the burst detection algorithm (see section 4.6.1 Burst Detection).
Run '*Data Preparation* > *Database* > *ISI* > *Extract References by Year for Burst Detection*' on the cleaned "with references and papers matched" database, followed by '*Analysis* > *Topical* > Burst Detection' with the following parameters:



View the file "Burst detection analysis (Publication Year, Reference): maximum burst level 1". On a PC running Windows, right click on this table and select view to see the data in Excel. On a Mac or a Linux system, right click and save the file, then open using the spreadsheet program of your choice. See Burst Detection for the meaning of each field in the output.

A empty value in the "End" field indicates that the burst lasted until the last date present in the dataset. Where the "End" field is empty, put manually the last year present in the dataset. In this case, 2007.

After you add manually this information, save this .csv file somewhere in your computer. Load back this .csv file into Sci2 using '*File* > *Load*'. Select 'Standart csv format' int the pop-up window. A new table will appear in the Data Manager. To visualize these table that contains the results of the Burst Detection algorithm, select the table you just loaded in the Data Manager and run '*Visualization* > *Temporal* > Horizontal Bar Graph' with the following parameters:



See section [2.4 Saving Visualizations for Publication] to save and view the graph.

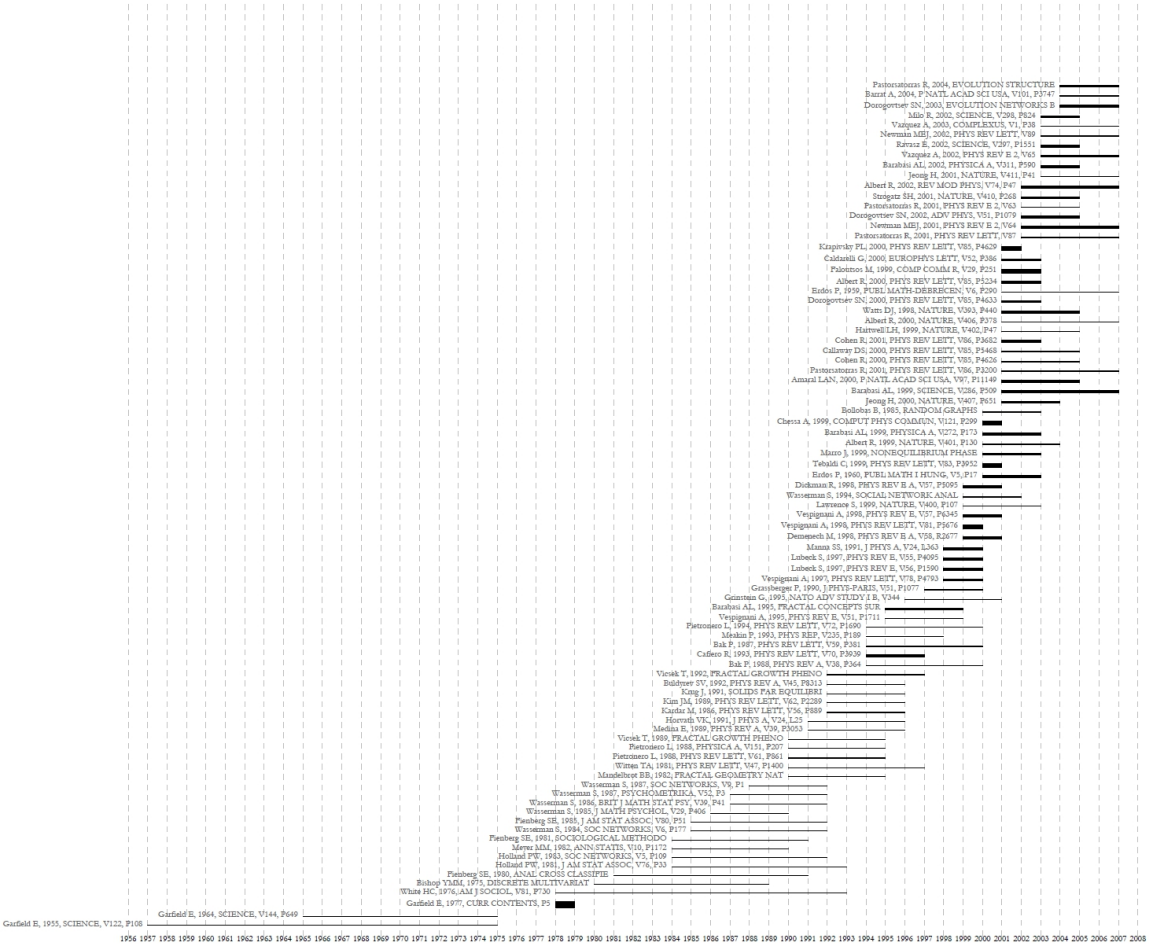


Figure 5.19: Top reference bursts in the 'FourNetSciResearchers' dataset

For temporal studies, it can be useful to aggregate data by year rather than by author, reference, etc. Running '*Data Preparation > Database > ISI > Extract Longitudinal Summary*' will output a table which lists metrics for every year mentioned in the dataset. The longitudinal study table contains the volume of documents and references published per year, as well as the total number of references, the number of distinct references, distinct authors, distinct sources, and distinct keywords per year. The results are graphed in Figure 5.20 (the graph was created using Excel).

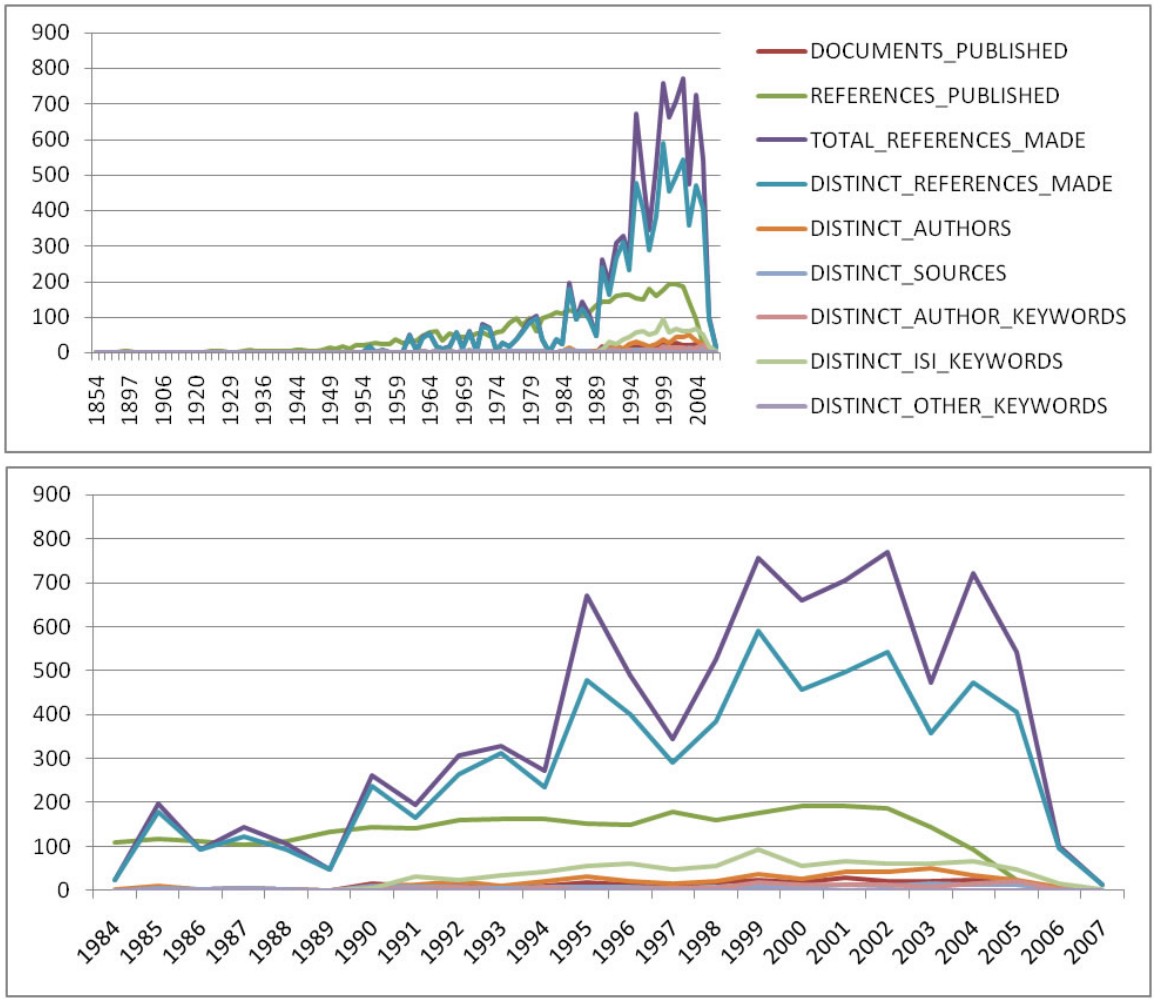


Figure 5.20: Longitudinal study of 'FourNetSciResearchers'

The largest speed increases from the database functionality can be found in the extraction of networks. First, compare the results of a co-authorship extraction with those from section 5.1.4.2 Author Co-Occurrence (Co-Author) Network. Run '*Data Preparation > Database > ISI > Extract Co-Author Network*' followed by '*Analysis > Networks > Network Analysis Toolkit (NAT)*'. Notice that both networks have 247 nodes and 891 edges. Visualize the extracted co-author network in GUESS using '*Visualization > Networks > GUESS*' and reformat the visualization using '*Layout*

> *GEM'* and *'Layout > Bin Pack*.' To apply the default co-authorship theme, go to 'Script > Run Script' and find *'**your sci2directory**/scripts/GUESS/co-author-nw_database.py'*. The resulting network will look like Figure 5.21.

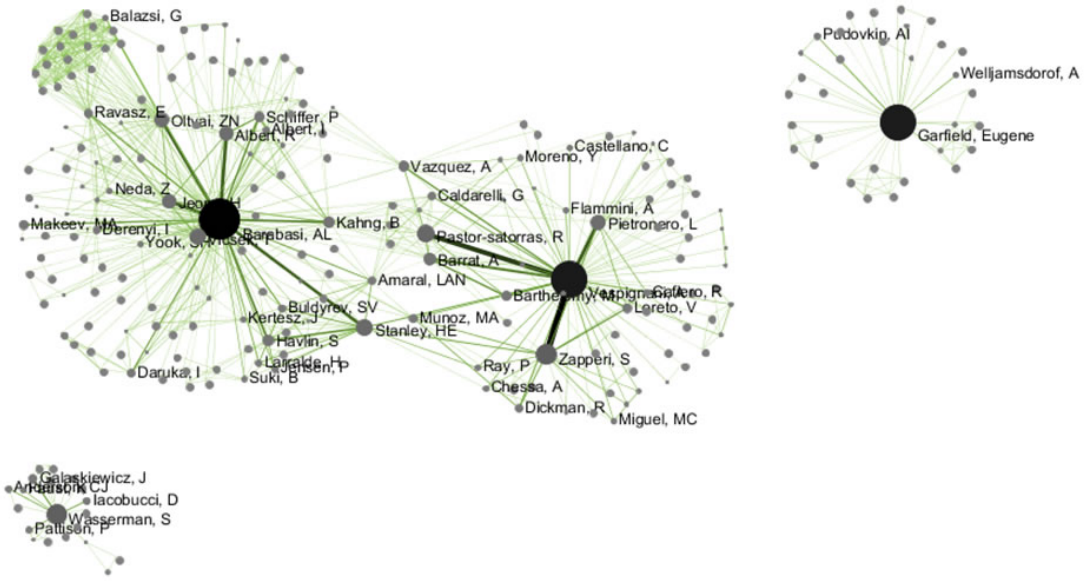
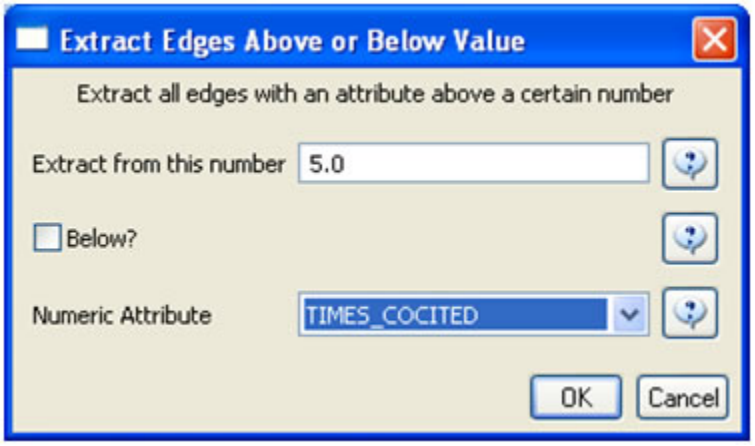


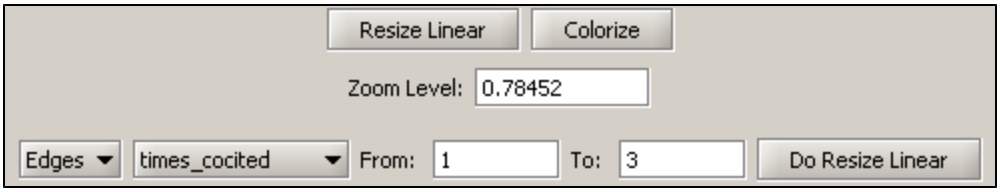
Figure 5.21: Longitudinal study of 'FourNetSciResearchers,' visualized in GUESS

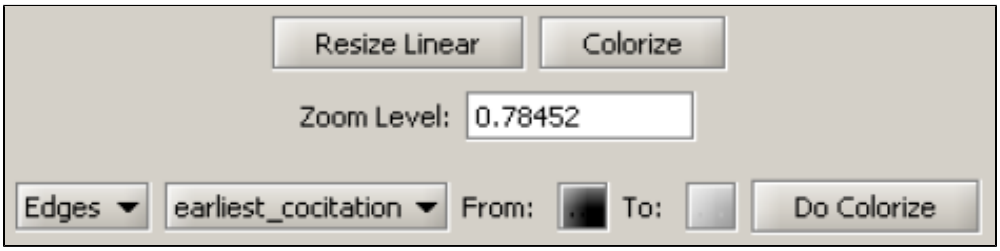
Using Sci^2's database functionality allows for several network extractions that cannot be achieved with the text-based algorithms. For example, extracting journal co-citation networks reveals which journals are cited together most frequently. Run '*Data Preparation > Database > ISI >* *Extract Document Co-Citation Network (Core and References)*' on the database to create a network of co-cited journals, and then prune it using '*Preprocessing > Networks >* *Extract Edges Above or Below Value*' with the parameters:



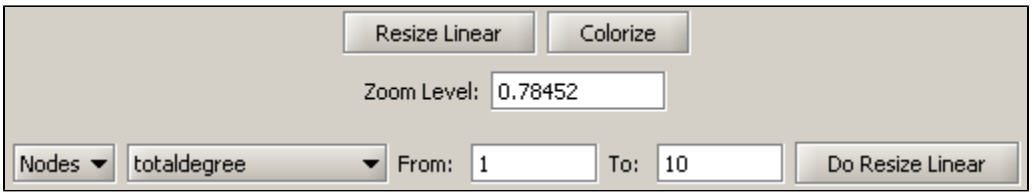
Now remove isolates ('*Preprocessing > Networks >* *Delete Isolates*') and append node degree attributes to the network ('*Analysis > Networks > Unweighted & Undirected >* *Node Degree*'). The workflow in your Data Manager should look like this:

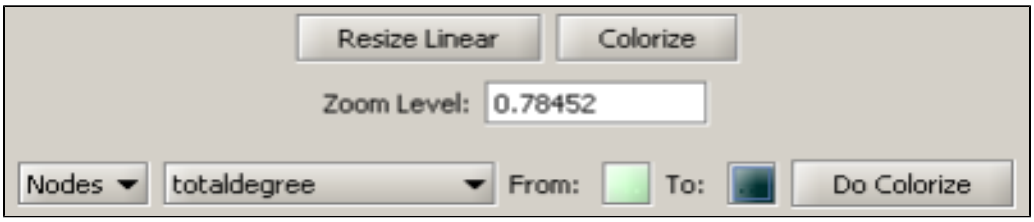
View the network in GUESS using '*Visualization > Networks >* *GUESS*.' Use '*Layout > GEM'* and *'Layout > Bin Pack'* to reformat the visualization. Resize and color the edges to display the strongest and earliest co-citation links using the following parameters:





Resize, color, and label the nodes to display their degree using the following parameters:





The resulting Journal Co-Citation Analysis (JCA) appears in Figure 5.22.

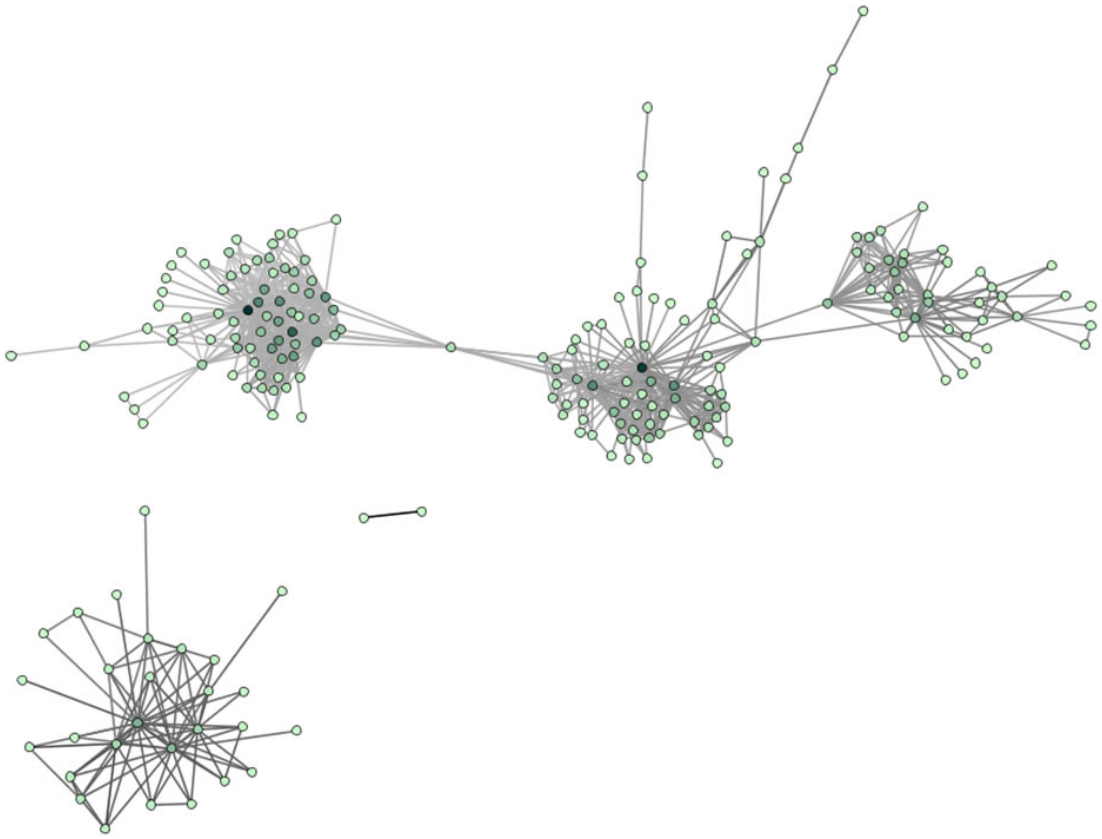


Figure 5.22: Journal co-citation analysis of 'FourNetSciResearchers,' edges with 5 or more cocitations and nodes with node degree added, visualized in GUESS

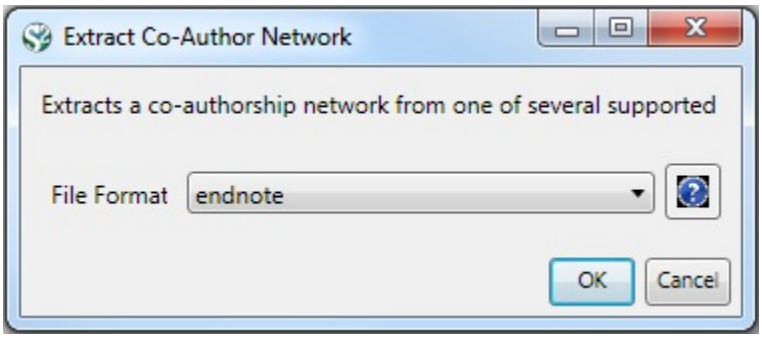
5.1.1 Mapping Collaboration, Publication, and Funding Profiles of One Researcher (EndNote and NSF Data)

5.1.1.1 Endnote

KatyBorner.enw	
Time frame:	1992-2010
Region(s):	Indiana University, University of Technology in Leipzig, University of Freiburg, University of Bielefeld
Topical Area(s):	Network Science, Library and Information Science, Informatics and Computing, Statistics, Cyberinfrastructure, Information Visualization, Cognitive Science, Biocomplexity
Analysis Type(s):	Co-Authorship Network

Many researchers, tools, and online services use EndNote to organize their bibliographies. To analyze an individual researcher's collaboration and publication profile, load an EndNote file which includes the researcher's entire CV into the Sci^2 Tool. To generate a research profile for Katy Börner, load Katy Börner's EndNote CV at '***yoursci2direc***

***tory**/sampledata/scientometrics/endnote/KatyBorner.enw*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Then run *'Data Preparation > Extract Co-Author Network'* using the parameter:



After generating Dr. Börner's co-authorship network, run *'Analysis > Networks > Unweighted & Undirected > Node Degree'* to append degree information to each node. To visualize the network, run *'Visualization > Networks > GUESS'* and select '*GEM*' in the Layout menu once the graph is fully loaded.. The resulting network in Figure 5.1 was modified using the following workflow:

1. Resize Linear > Nodes > totaldegree > From: 5 To: 30 > Do Resize Linear (Note: total degree is the number of papers)
2. Resize Linear > Edges > weight From: 1 To: 10 > Do Resize Linear (Note: weight is the number of co-authored papers)
3. Colorize > Nodes > totaldegree From :  To:  > Do Colorize
4. Colorize > Edges > weight From:  To:  > Do Colorize
5. Object: nodes based on -> > Property: totaldegree > Operator: >= > Value: 10 > Show Label
6. Type in Interpreter:

```
>for n in g.nodes:
     n.strokecolor = n.color
```

GUESS Graph Modifier

Make sure to maximize the window displaying the GUESS visualization tool in order to see all the options in the Graph Modifier.

Note: The Interpreter tab will have '>>>' as a prompt for these commands. It is not necessary to type '>" at the beginning of the line. You should type each line individually and press "Enter" to submit the commands to the Interpreter. For more information, refer to the GUESS tutorial at http://nwb.cns.iu.edu/Docs/GettingStartedGUESSNWB.pdf. The largest cluster in the network is outlined in black, and represents one single paper with many authors.

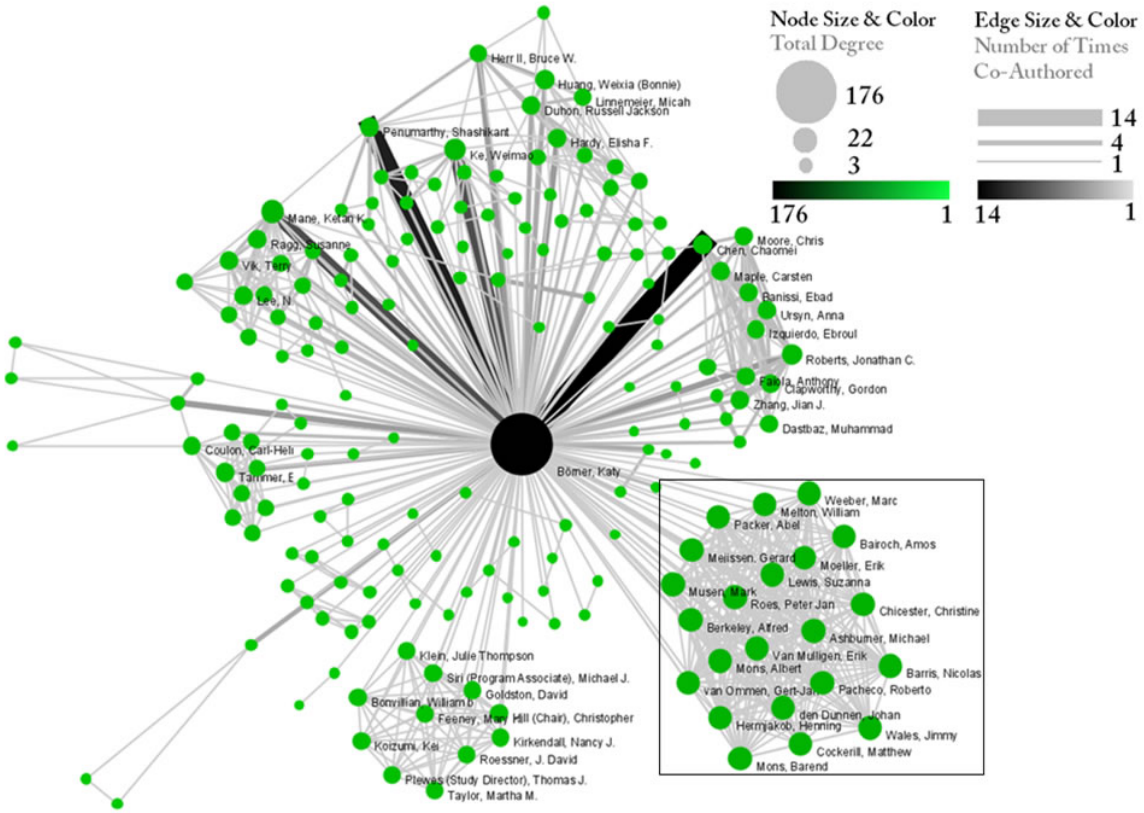


Figure 5.1: Co-authorship network of Katy Börner
GUESS supports the repositioning of selected nodes. Multiple nodes can be selected by holding down the 'Shift' key and dragging a box around specific nodes. The final network can be saved via *'GUESS: File > Export Image'* and opened in a graphic design program to add a title and legend. The image above was created using Adobe Photoshop. Node clusters were highlighted and increased in size, the label font size was increased for emphasis, and a legend was added to clarify the significance of node and edge size and color.

To see the log file from this workflow save the 5.1.1.1 Endnote log file.

5.1.1.1.1 Using Cytoscape to Visualize Networks

> **Extended Version**
>
> This workflow uses the extended version of the Sci2 Tool. To know how to extend Sci2 view Section 3.2 Additional Plugins.

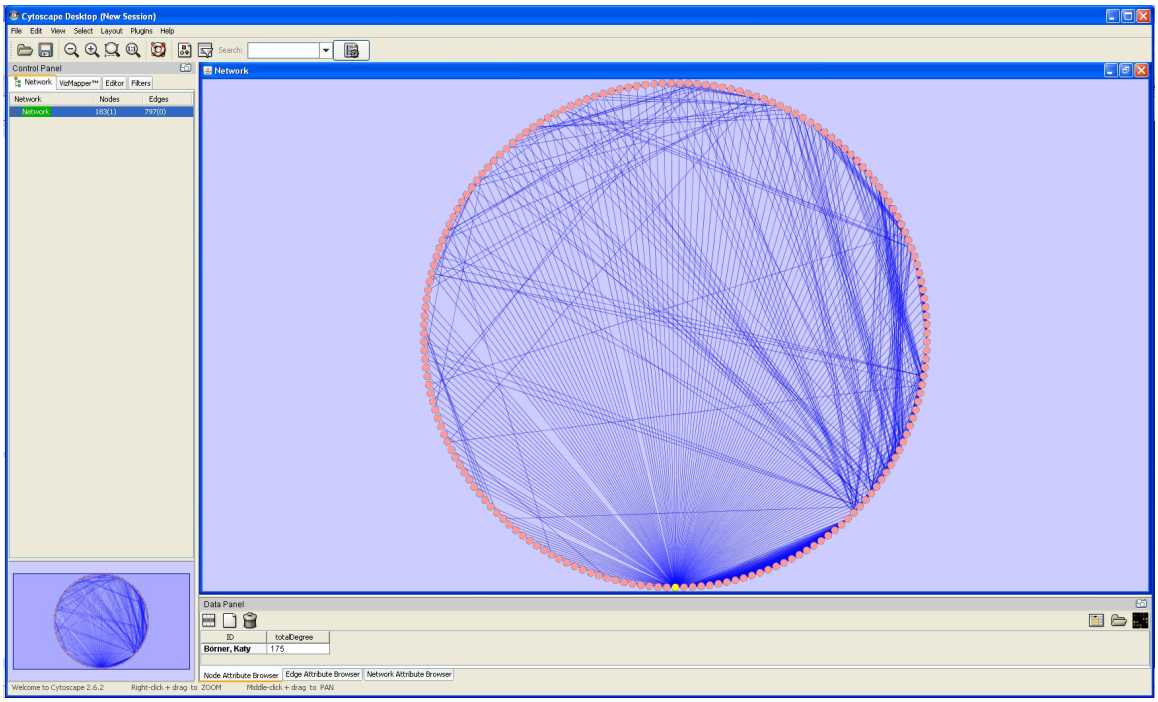
Cytoscape (http://www.cytoscape.org) is an open source software platform for visualizing networks and integrating these with any type of attribute data. Although Cytoscape was originally designed for biological research, now it is a general platform for network analysis and visualization. A variety of layout algorithms are available, including cyclic, tree, force-directed, edge-weight, and yFiles Organic layouts.

Cytoscape is available as a plugin in the extended version of the Sci^2 Tool. To visualize networks in Sci^2 using Cytoscape, first you have to extend the Sci^2 Tool according to the directions at Section 3.2 Additional Plugins.

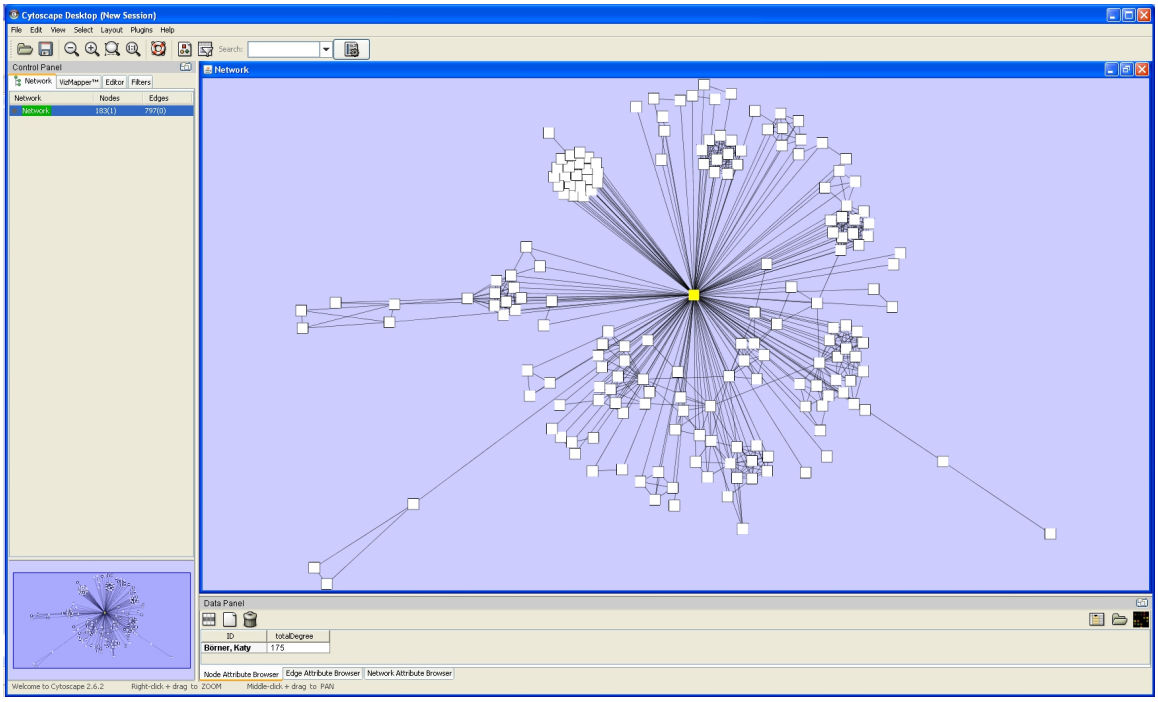
To exemplify how to use Cytoscape to visualize Networks, let's use Dr. Börner's co-authorship network generated above. To visualize this network, run '*Visualization > Networks > Cytoscape*'.

In Cytoscape, the Layout menu has an array of features for organizing the network visually according to one of several algorithms, aligning and rotating groups of nodes, and adjusting the size of the network.

For example, run 'Layout > Degree Sorted Circle Layout > All Nodes' once the network is fully loaded. This layout algorithm sort nodes in a circle by degree of the nodes. When the layout for the network finishes, Cytoscape will look similar to the image below:



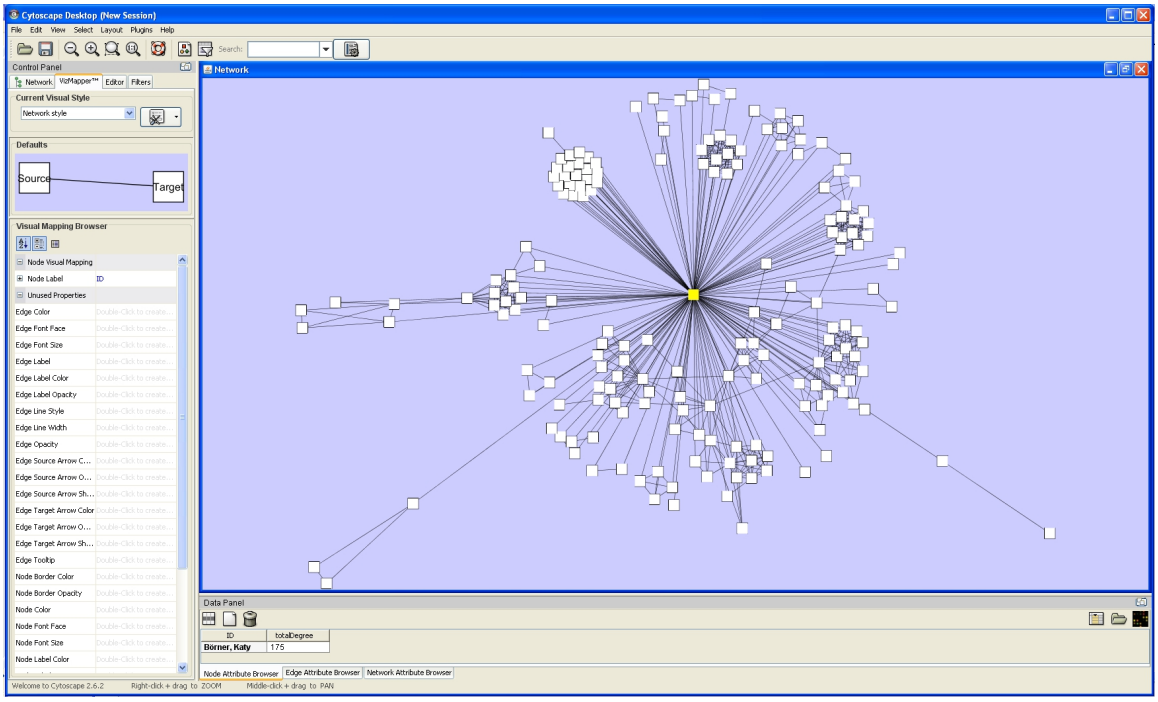
Now, run 'Layout > Cytoscape Layouts > Spring Embedded > All nodes'. This spring-embedded layout is based on a "force-directed" paradigm as implemented by Kamada and Kawai (1988). Network nodes are treated like physical objects that repel each other, such as electrons. The connections between nodes are treated like metal springs attached to the pair of nodes. These springs repel or attract their end points according to a force function. The layout algorithm sets the positions of the nodes in a way that minimizes the sum of forces in the network. To see a description of all layouts and functionalities available in Cytoscape see [Cytoscape User Manual.]



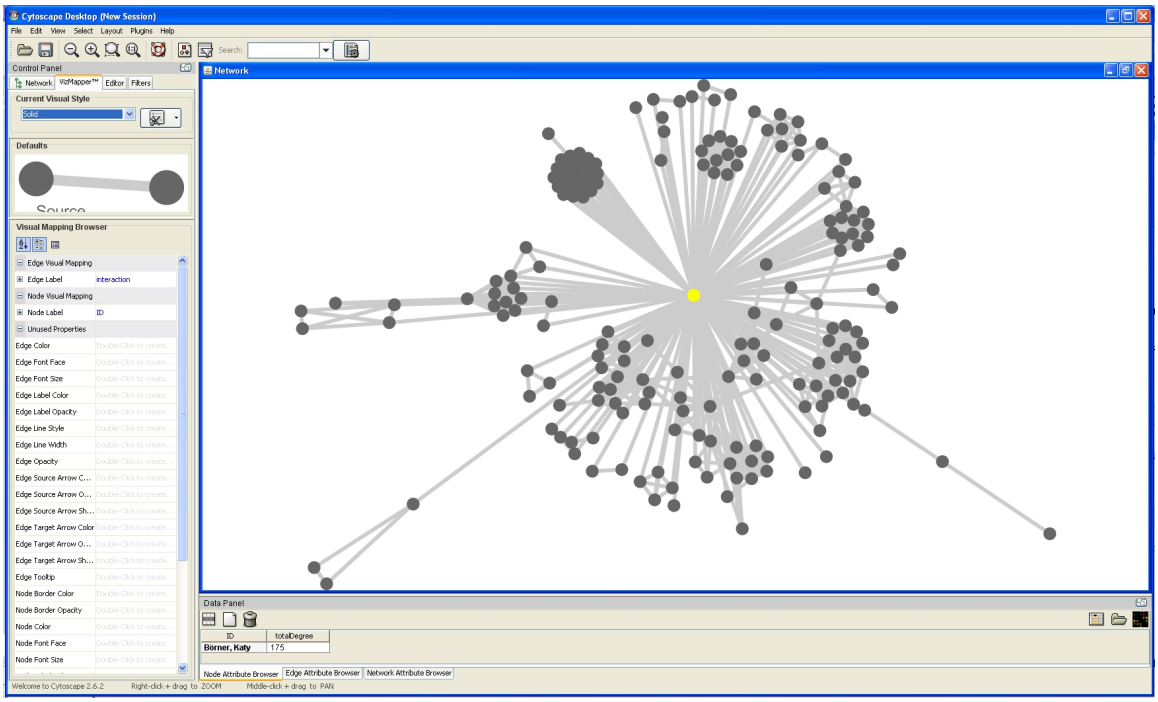
The main window of Cytoscape has several components:

- The menu bar at the top
- The toolbar, which contains icons for commonly used functions. These functions are also available via the menus. Hover the mouse pointer over an icon and wait momentarily for a description to appear as a tooltip.
- The Control Panel that allows the management of the network (top left panel). This contains an optional network overview pane (shown at the bottom left).
- The main network view window, which displays the network.
- The Data Panel (bottom panel), which displays attributes of selected nodes and edges and enables you to modify the values of attributes.

One of Cytoscape's strengths in network visualization is the ability to allow users to encode any attribute of their data (name, type, degree, weight, expression data, etc.) as a visual property (such as color, size, transparency, or font type). A set of these encoded or mapped attributes is called a *Visual Style* and can be created or edited using the Cytoscape *VizMapper*. VizMapper is located at the second tab at the Control Panel.

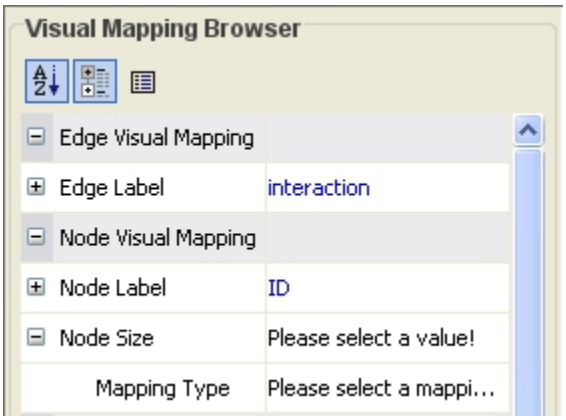


You can change visual styles by making a selection from the *Current Visual Style* dropdown list (found at the top of the *VizMapper* main panel). For example, if you select *Solid*, a new visual style will be applied to your network, and you will see
a white background and round gray nodes.

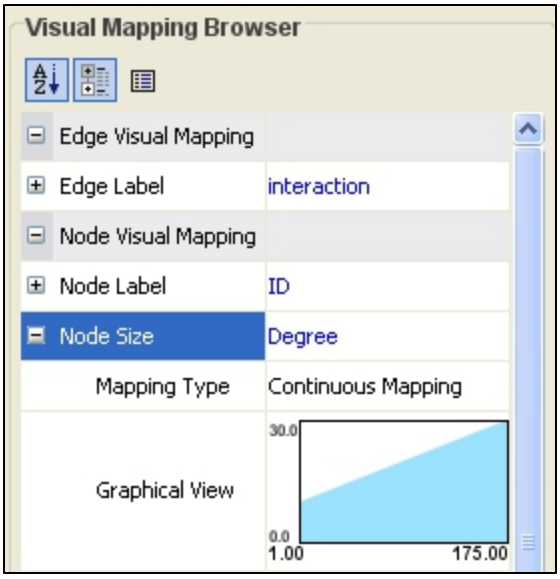


All visual attributes are listed in the *Unused Properties* category. From this panel, you can create node/edge mappings for all visual properties.

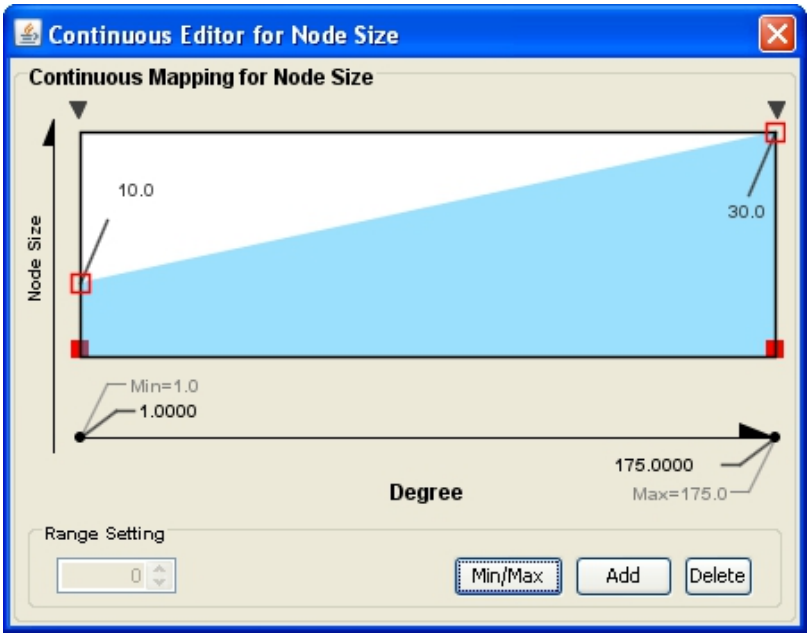
Double click the Node Size entry listed in *Unused Properties*. *Node Size* will now appear at the top of the list, under the *Node Visual Mapping* category (as shown below).



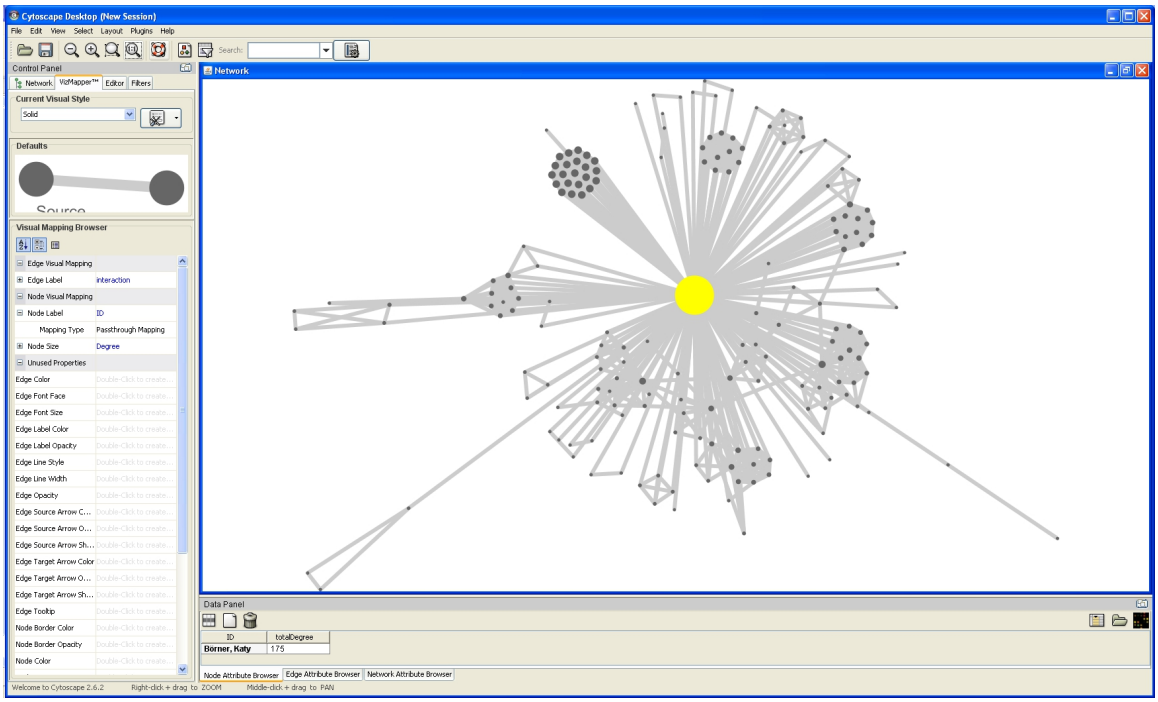
Click on the cell to the right of the *Node Size* entry and select *Degree* from the drop-down list that appears. Set the *Continuous Mapping* option as the *Mapping Type*.



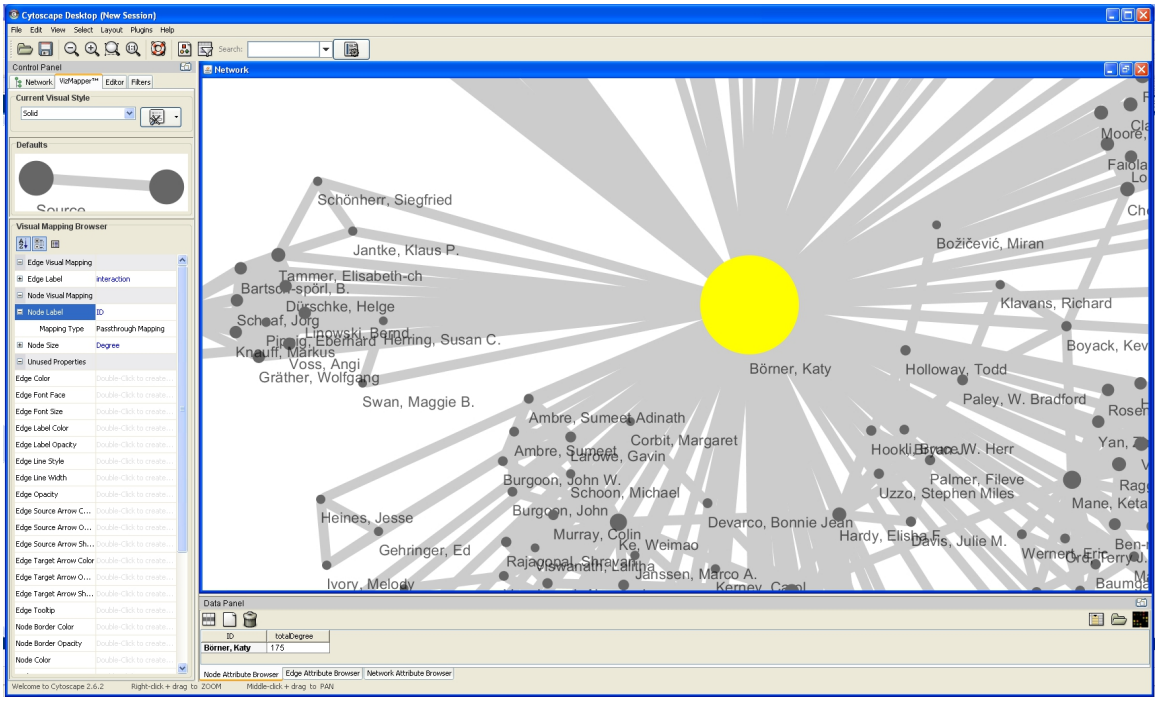
Double-click on the black-and-blue rectangle next to *Graphical View* to open the *Continuous Editor for Node Size*. Double-lick on the second point (set at 30.0 initially) and type as 120 as the new value. The nodes size will be immediately updated. Close the *Continuous Editor for Node Size*.



The network will look like this:



Nodes labels appear an zoom is applied to the network:



To see the log file from this workflow save the 5.1.1.1.1 Using Cytoscape to Visualize Netowrks log file.

5.1.1.2 NSF

KatyBorner.nsf	
Time frame:	2003-2008

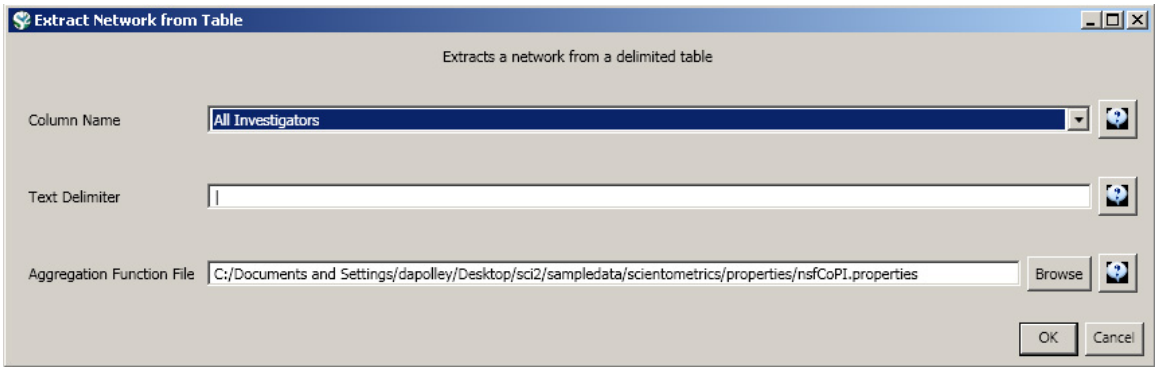
Region(s):	Indiana University
Topical Area(s):	Network Science, Library and Information Science, Informatics and Computing, Statistics, Cyberinfrastructure, Information Visualization, Cognitive Science, Biocomplexity
Analysis Type(s):	Co-PI Network, Grant Award Summary

The second part of Katy Börner's research profile will focus on her Co-PIs. The data can be downloaded for free using NSF's Award Search (See section 4.2.2.1 NSF Award Search) by searching for "Katy Borner" in the "Principal Investigator" field and keeping the "Include CO-PI" box checked.

Load the NSF data using '*File > Load*' and following this path: '***yoursci2directory****/sampledata/scientometrics/nsf/KatyBorner.nsf*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Make sure the loaded dataset in the Data Manager window is highlighted in blue, and run *'Data Preparation > Extract Co-Occurrence Network'* using these parameters:

Aggregate Function File

Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



Select the "Extracted Network on Column All Investigators" network and run *'Analysis > Networks > Network Analysis Toolkit (NAT)'* to reveal that there are 13 nodes and 28 edges without isolates in the network. Click on "Extracted Network on Column All Investigators" and select *'Visualization > Networks > GUESS'* to visualize the resulting Co-PI network. Select '*GEM*' from the layout menu.

Load the default Co-PI visualization theme via '*Script > Run Script ...*'and load '***yoursci2directory****/scripts/GUESS/co-PI-nw.py*'. Alternatively, use the "Graph Modifier" to customize the visualization. The resulting network in Figure 5.2 was modified using the following workflow:

1. Resize Linear > Nodes > totalawardmoney > From: 5 To: 35 > Do Resize Linear
2. Resize Linear > Edges > coinvestigatedawards From: 1 To: 2 > Do Resize Linear
3. Colorize > Nodes > totalawardmoney From : To: > Do Colorize
4. Colorize > Edges > coinvestigatedawards From: To: > Do Colorize
5. Object: all nodes > Show Label
6. Type in Interpreter:

```
>for n in g.nodes:
     n.strokecolor = n.color
```

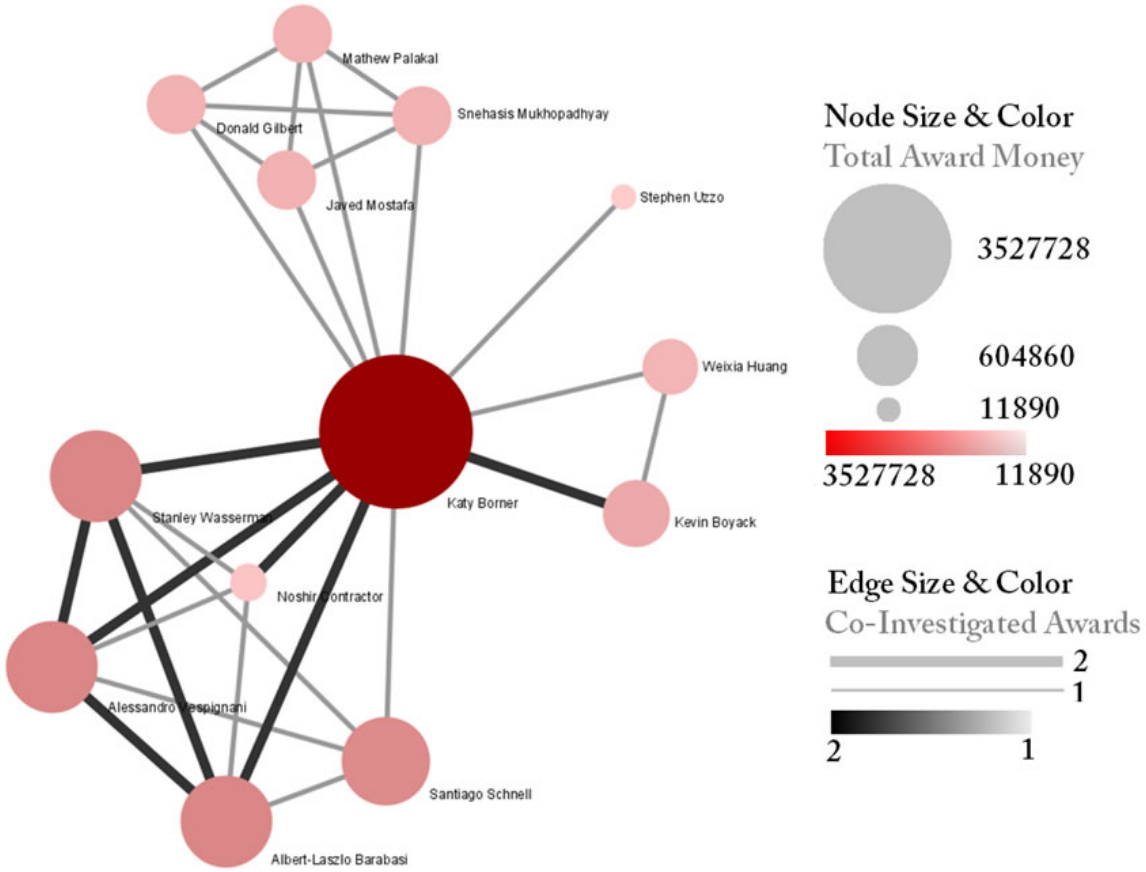
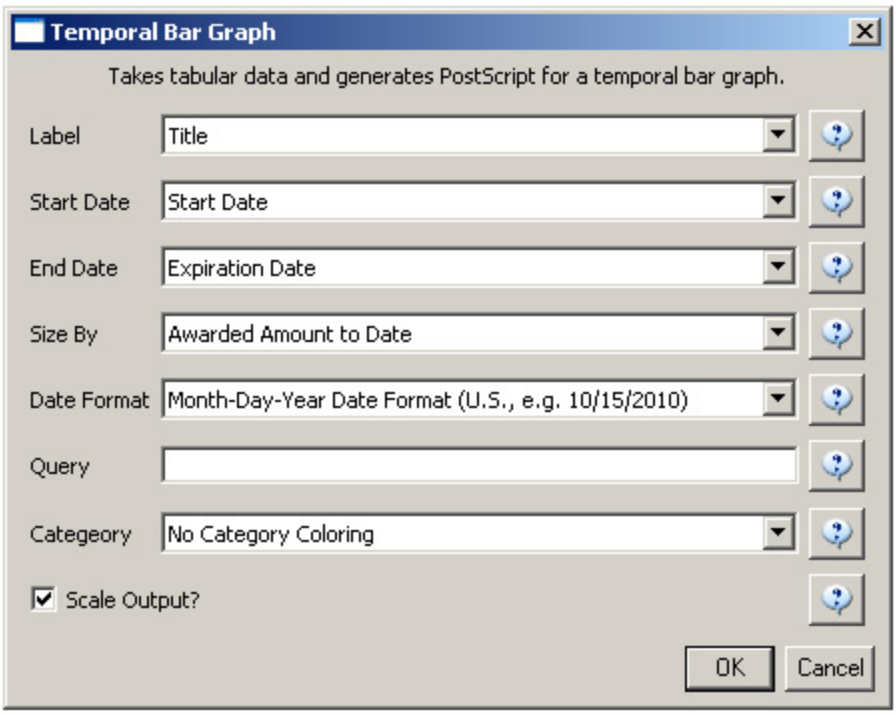


Figure 5.2: NSF Co-PI network of Katy Börner

For a summary of the grants themselves, with a visual representation of their award amount, select the original *'Katy Borner.nsf'* csv file in the Data Manager and run '*Visualization* > *Temporal* > *Temporal Bar Graph*', entering the following parameters:



The generated postscript file "HorizontalBarGraph_KatyBorner.ps" can be viewed using Adobe Distiller or GhostViewer (see section 2.4 Saving Visualizations for Publication).

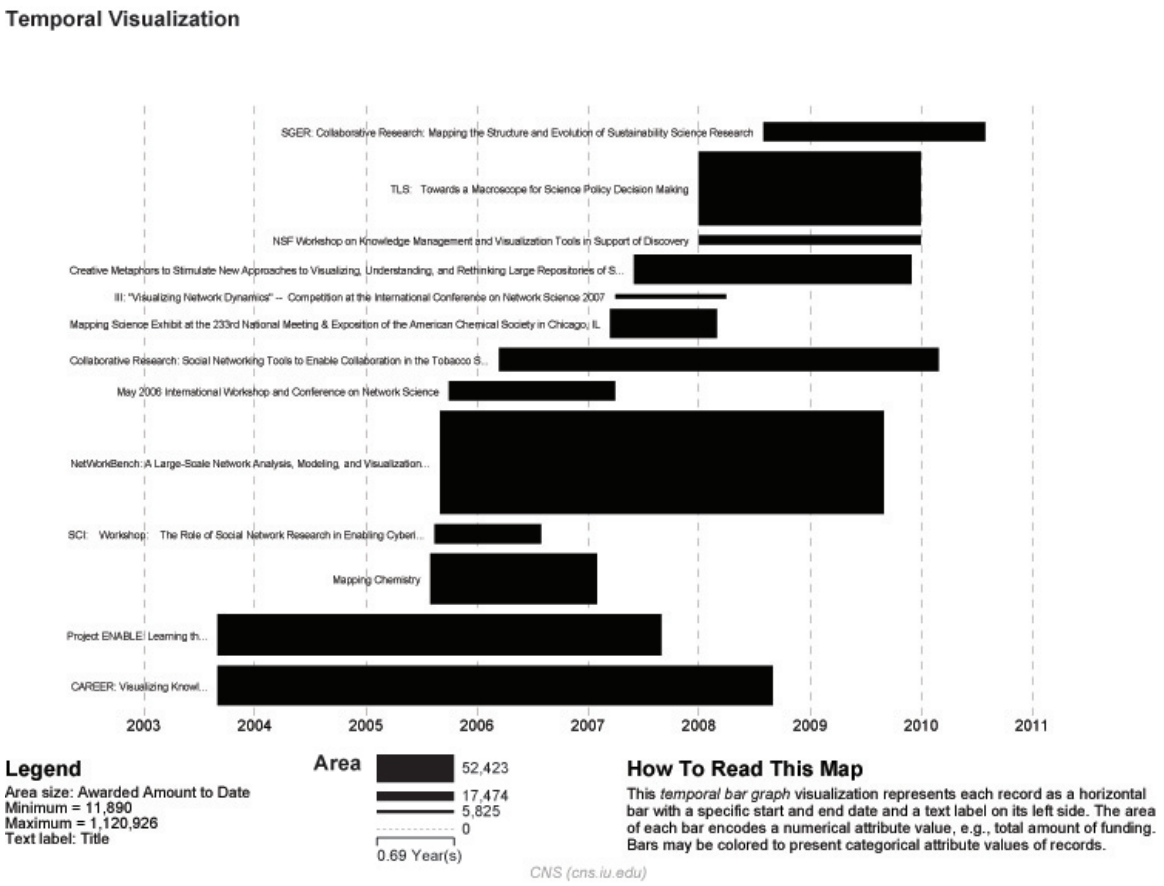


Figure 5.3: Horizontal Bar Graph of KatyBorner.nsf

To see the log file from this workflow save the 5.1.1.2 NSF log file.

5.1.2 Time Slicing of Co-Authorship Networks (ISI Data)

AlessandroVespignani.isi	
Time frame:	1990-2006
Region(s):	Indiana University, University of Rome, Yale University, Leiden University, International Center for Theoretical Physics, University of Paris-Sud
Topical Area(s):	Informatics, Complex Network Science and System Research, Physics, Statistics, Epidemics
Analysis Type(s):	Co-Authorship Network

The Sci^2 Tool supports the analysis of evolving networks. For this study, load Alessandro Vespignani's publication history from ISI, which can be downloaded from Thomson's Web of Science or loaded using '*File > Load*' and following this path: '***yoursci2directory****/sampledata/scientometrics/isi/AlessandroVespignani.isi*'using' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets).

New ISI File Format

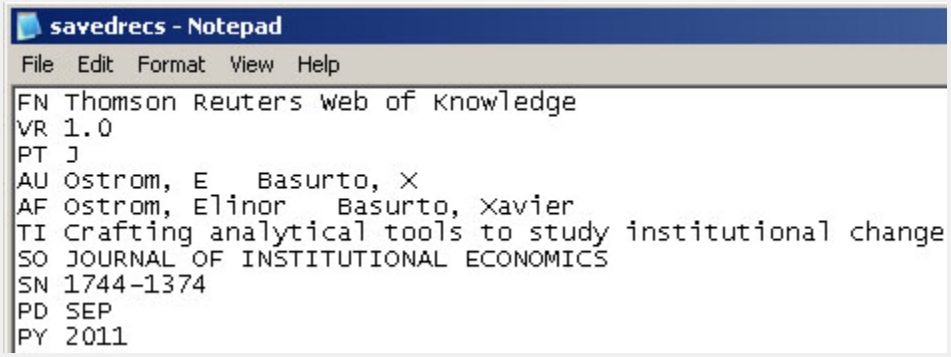
Web of Science made a change to their output format in September, 2011. Older versions of Sci2 tool (older than v0.5.2 alpha) may refuse to load these new files, with an error like "Invalid ISI format file selected."

Sci2 solution

If you are using an older version of the Sci2 tool, you can download the WOS-plugins.zip file and unzip the JAR files into your sci2/plugins/ directory. Restart Sci2 to activate the fixes. You can now load the downloaded ISI files into the Sci2 without any additional step. If you are using the old Sci2 tool you will need to follow the guidelines below before you can load the new WOS format file into the tool.

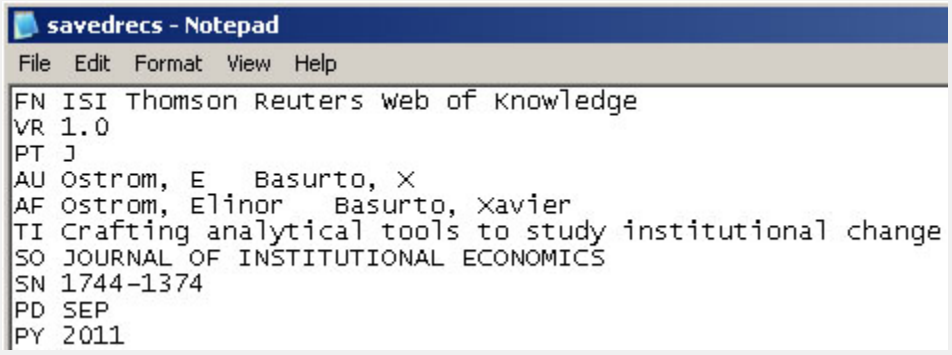
You can fix this problem for individual files by opening them in Notepad (or your favorite text editor). The file will start with the words:

Original ISI file:



Just add the word ISI.

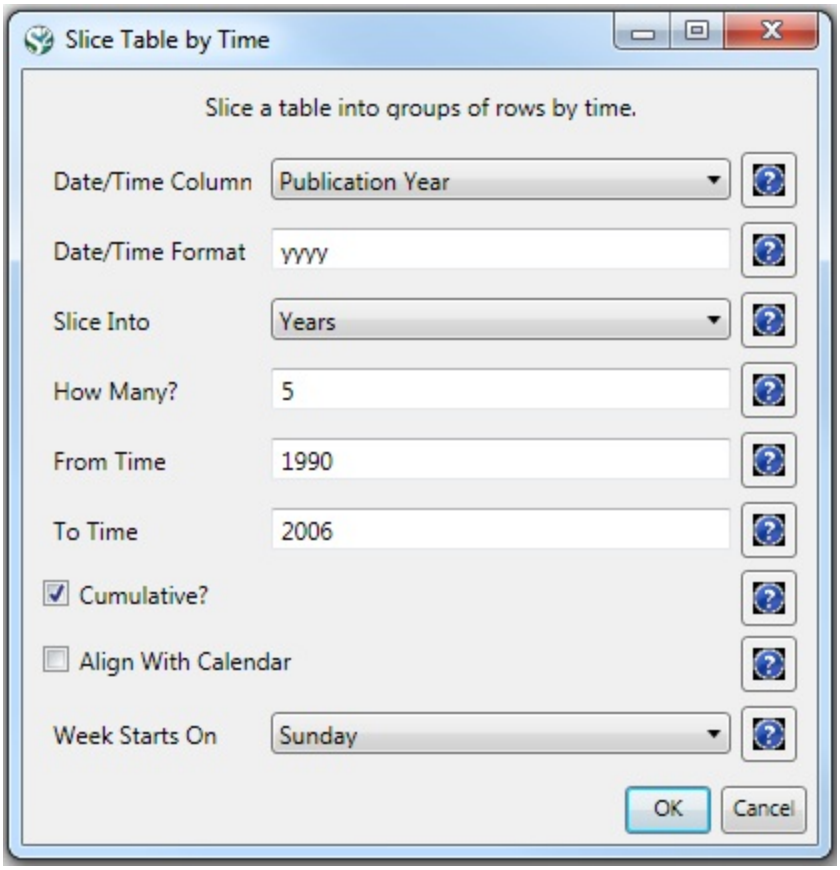
Updated ISI file:



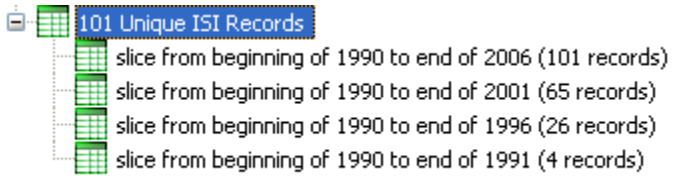
And then Save the file.

The ISI file should now load properly. More information on the ISI file format is available here (http://wiki.cns.iu.edu/display/CISHELL/ISI+%28*.isi%29).

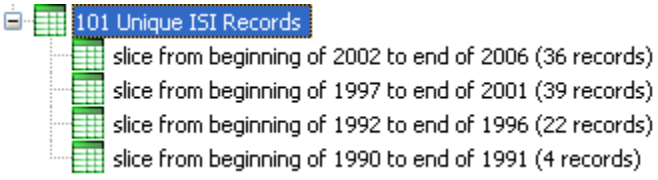
Slice the data into five year intervals from 1990-2006 using *'Preprocessing > Temporal > Slice Table by Time'* and the following parameters:



"Slice Into" allows the user to slice the table by days, weeks, months, quarters, years, decades, and centuries. There are two additional parameters for time slicing: cumulative and align with calendar. The former produces tables containing all data from the beginning to the end of each table's time interval, which can be seen in the Data Manager and below:



The latter option aligns the output tables according to calendar intervals:



Choosing "Years" under "Slice Into" creates multiple tables beginning from January 1st of the first year. If "Months" is chosen, it will start from the first day of the earliest month in the chosen time interval.

To see the evolution of Vespignani's co-authorship network over time, check "cumulative". Then, extract co-authorship networks one at a time for each sliced time table using *'Data Preparation* > *Extract Co-Author Network* ', making sure to select "ISI" from the pop-up window during the extraction. To view each of the Co-Authorship Networks over time using the same graph layout, begin by clicking on longest slice network (the 'Extracted Co-Authorship Network' under 'slice from beginning of 1990 to end of 2006 (101 records)') in the data manager.

Visualize it in GUESS using '*Visualization* > *Networks* > *GUESS*'. From here, run '*Layout* > *GEM*' followed by '*Layou t* > *Bin Pack'*. Run '*Script* > *Run Script ...*' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*'.

The resulting visualization is of the complete co-authorship network over all years. Use *Node Locking* to save the *x, y* coordinates of each node and apply them to the other time slices. In GUESS, select '*File* > *Export Node Positions*' and save the result as '***yoursci2directory****/NodePositions.csv*'. Load the remaining three networks in GUESS using the steps described above and for each network visualization, run '*File* > *Import Node Positions*' and open '***yoursci2 directory****/NodePositions.csv*'. To match the resulting networks stylistically with the original visualization, run '*Script* > *Run Script ...*' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*', followed by '*Layout* > *Bin Pack'*, for each. The evolving network is shown in Figure 5.4.

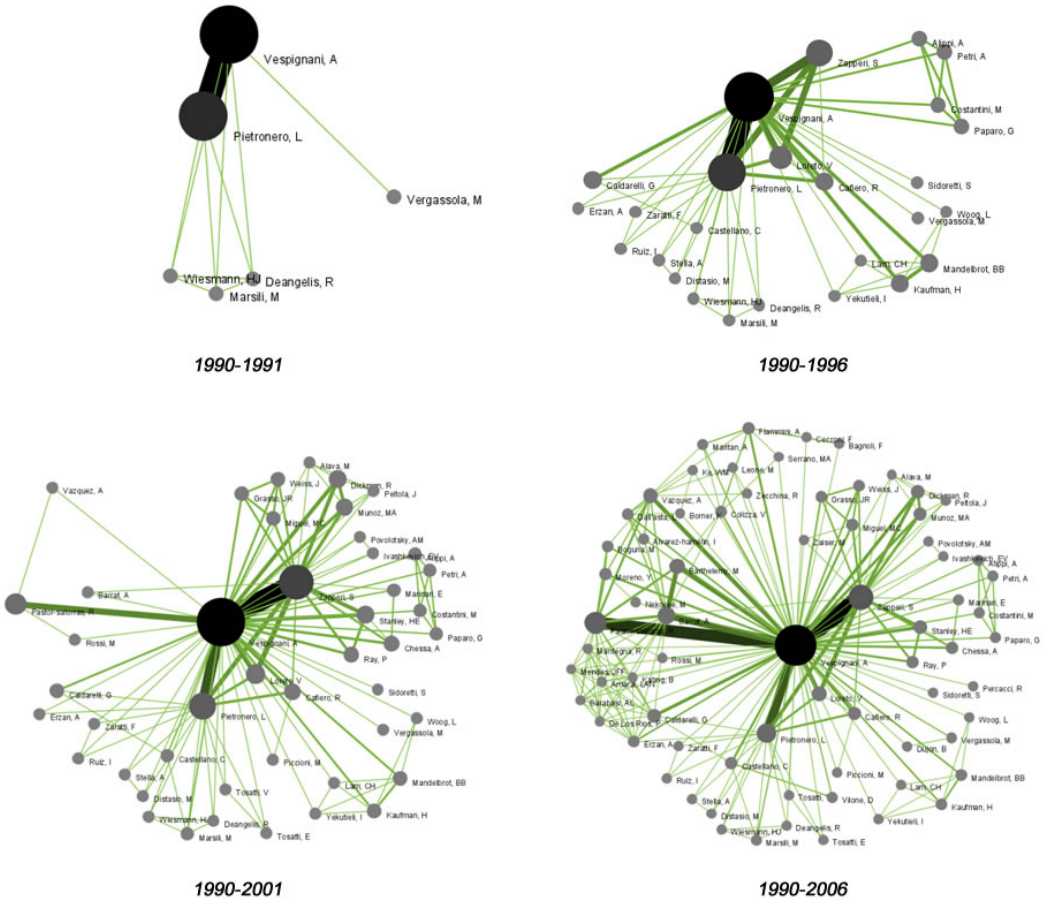


Figure 5.4: Evolving co-authorship network of Vespignani from 1990-2006
The four networks reveal that from 1988-1992, Alessandro Vespignani had one primary co-author and four secondary co-authors. His network expanded considerably to include multiple primary co-authors and more than 200 secondary co-authors in 2006.

To see the log file from this workflow save the 5.1.2 Time Slicing of Co-Authorship Networks (ISI Data) log file.

5.1.3 Funding Profiles of Three Researchers at Indiana University (NSF Data)

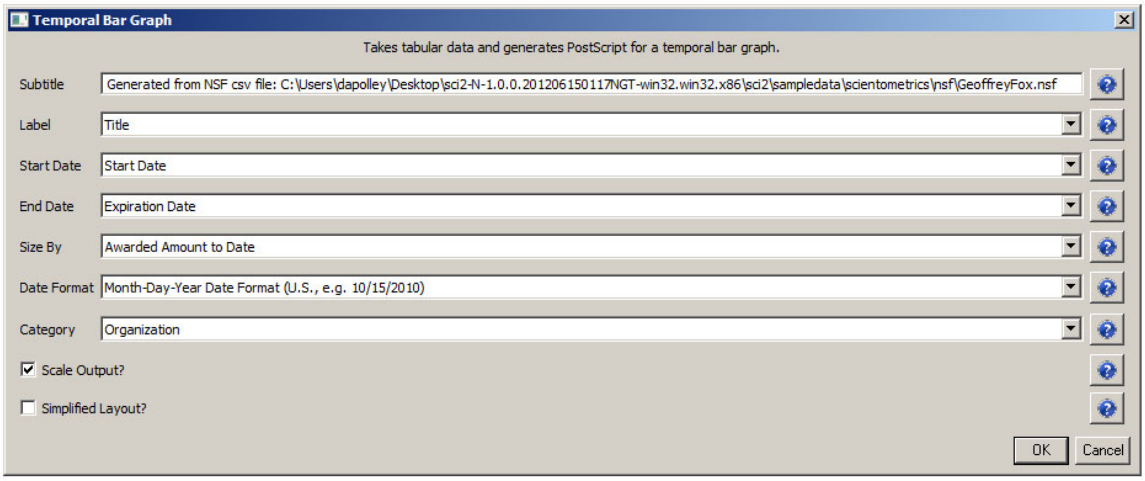
GeoffreyFox.nsf BethPlale.nsf MichaelMcRobbie.nsf	

Time frame:	1978-2010
Region(s):	Indiana University
Topical Area(s):	Informatics, Miscellaneous
Analysis Type(s):	Co-PI Network, Grant Award Summary

It is often useful to compare the profiles of multiple researchers within similar disciplinary or institutional domains. To demonstrate this comparison, load the NSF funding profiles of three Indiana University researchers into the Sci^2 Tool using '*File > Load*' and following this path: '***yoursci2directory****/sampledata/scientometrics/nsf*'. Load *'GeoffreyFox.nsf', 'MichaelMcRobbie.nsf',* and *'BethPlale.nsf'* in NSF csv format (if these files are not in the sample data directory they can be downloaded from 2.5 Sample Datasets). Then run '*Visualization > Temporal > Horizontal Bar Graph*', using the recommended parameters for each.

For instructions on how to save and view the PostScript file generated by the "Horizontal Bar Graph" algorithm, see section 2.4 Saving Visualizations for Publication.

Select *'GeoffreyFox.nsf'* in the Data Manager. Use the following parameters to generate a Horizontal Bar Graph:



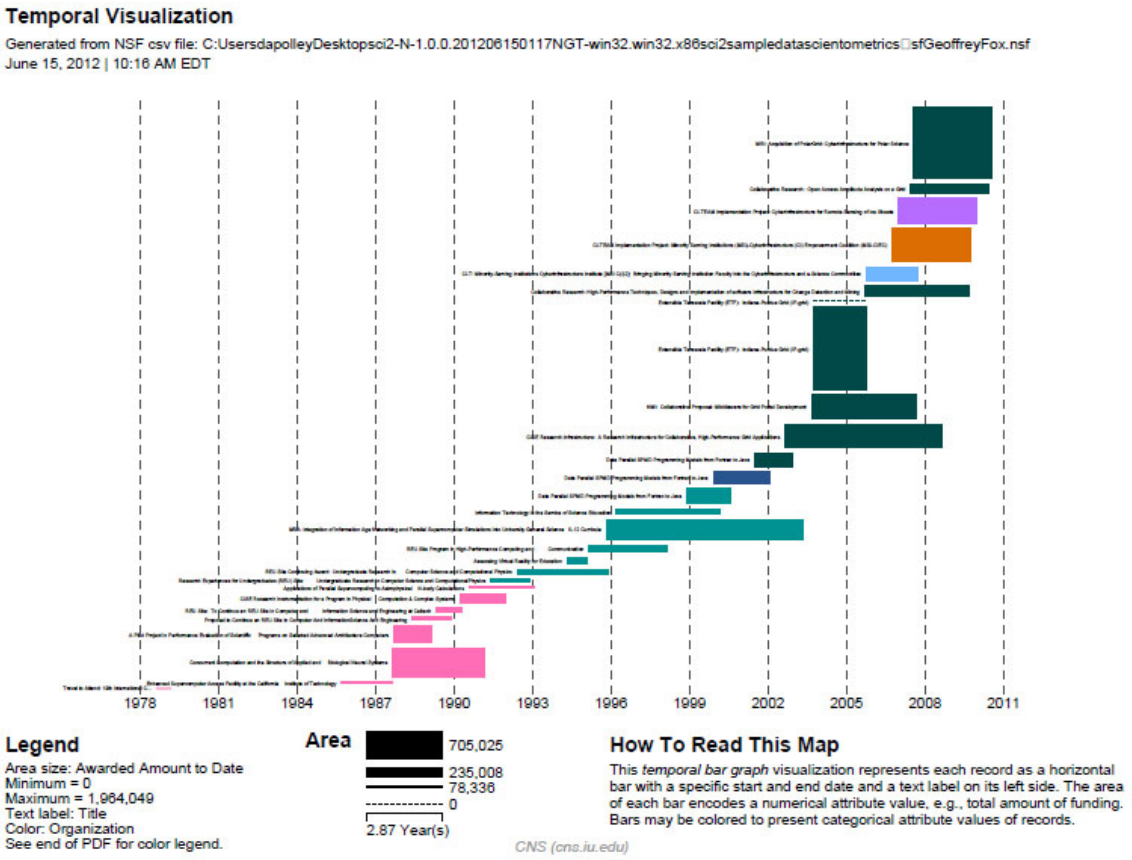


Figure 5.5: Funding profile over time of Geoffrey Fox

Now select *'BethPlale.nsf'* in the Data Manager. Use the following parameters to generate a Horizontal Bar Graph:

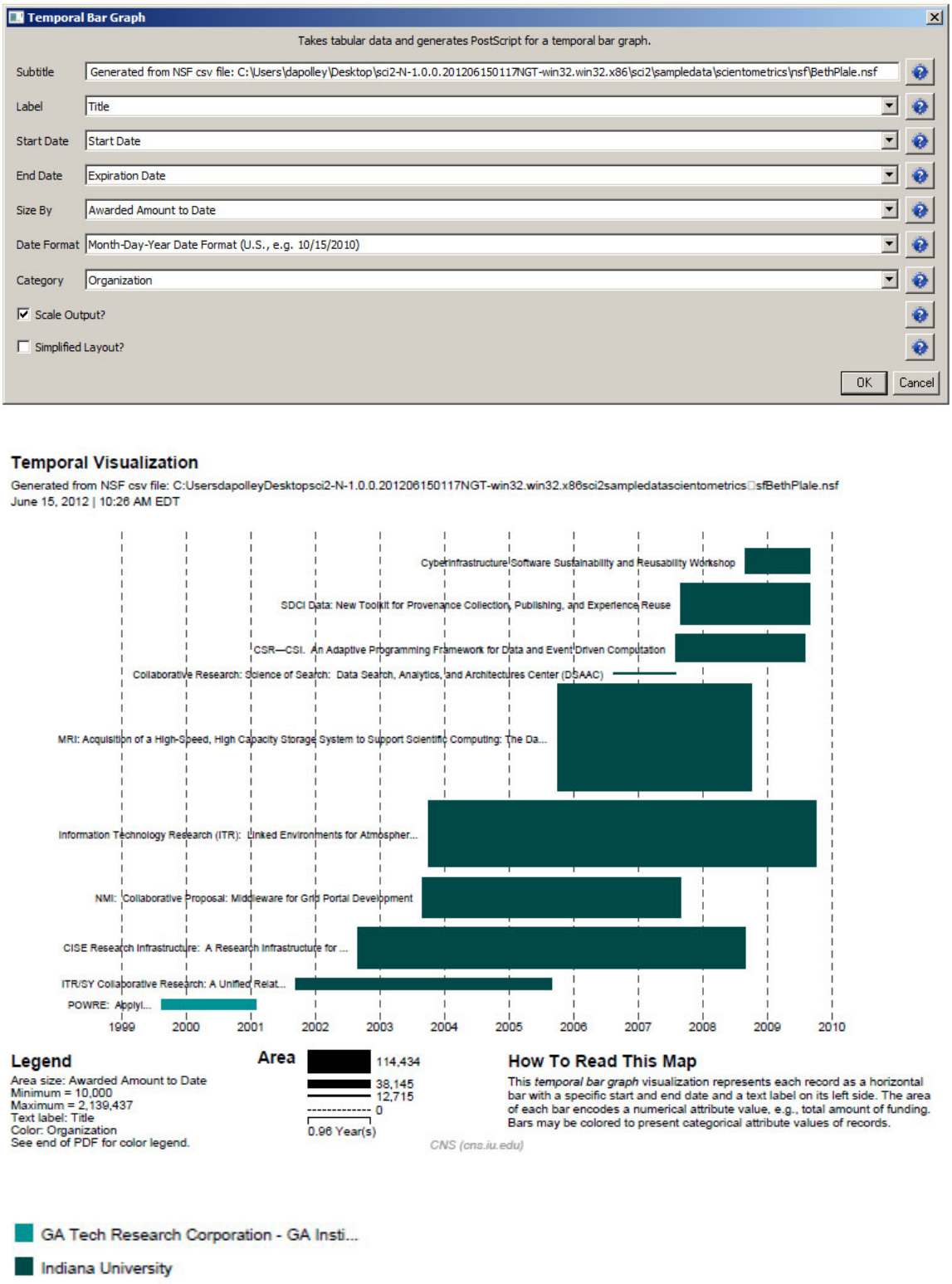
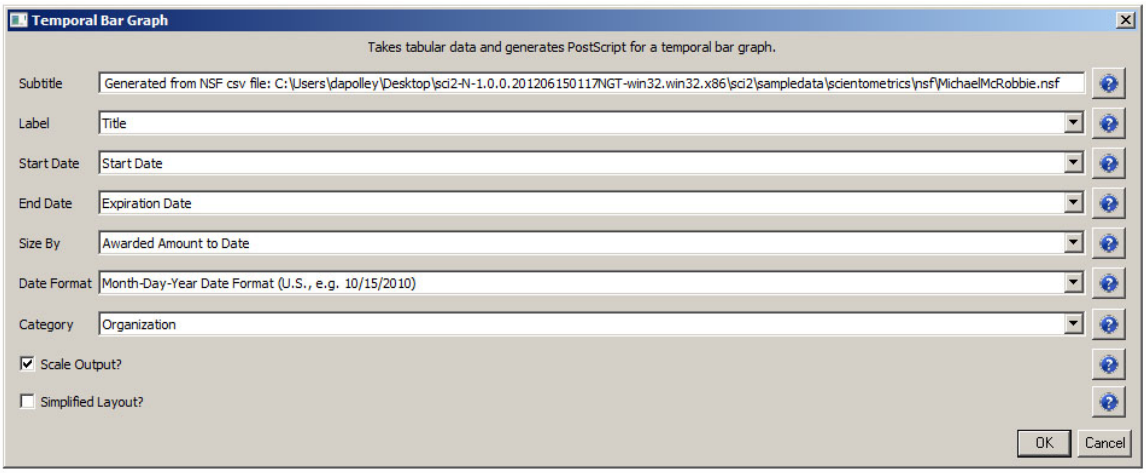


Figure 5.6: Funding profile over time of Beth Plale

Finally, select *'MichaelMcRobbie.nsf'* in the Data Manager. Use the following parameters to generate a Horizontal Bar Graph:



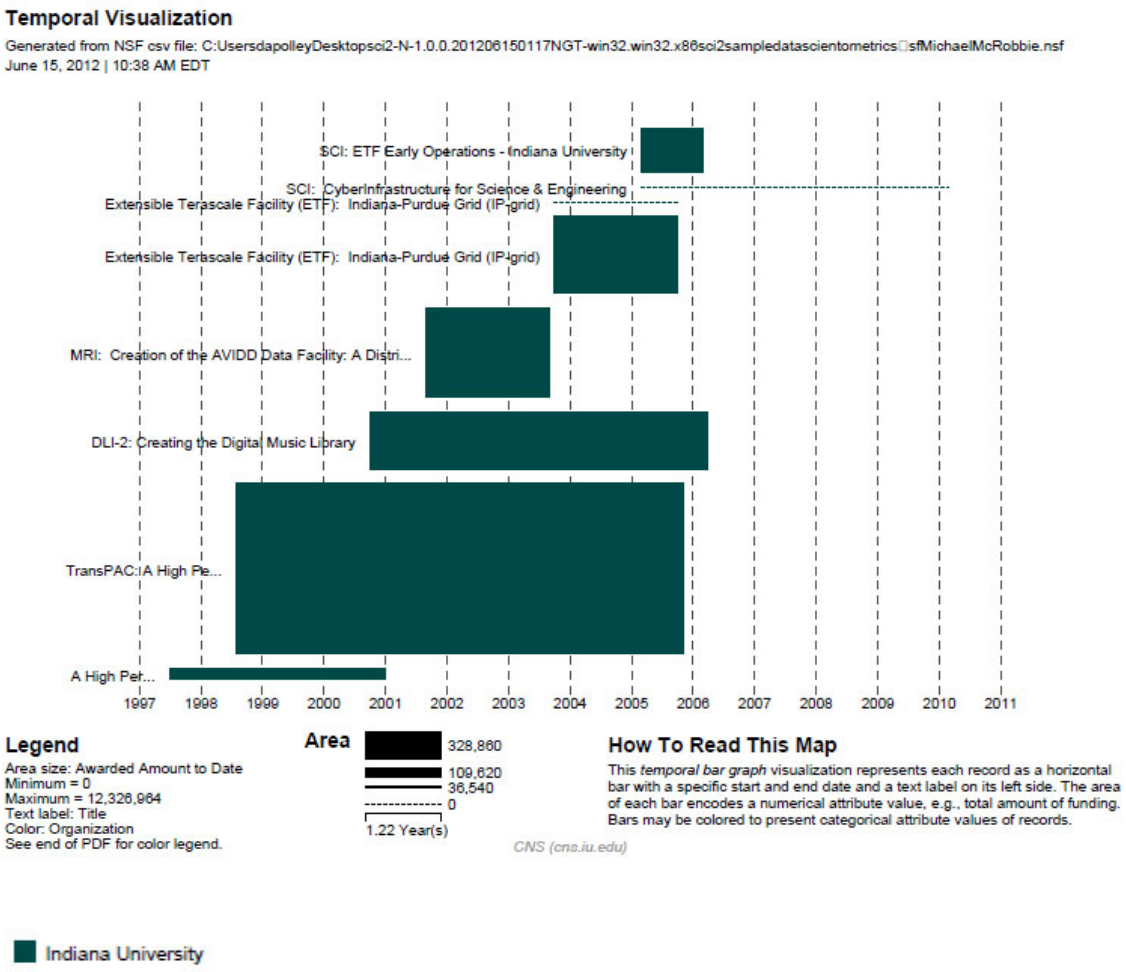
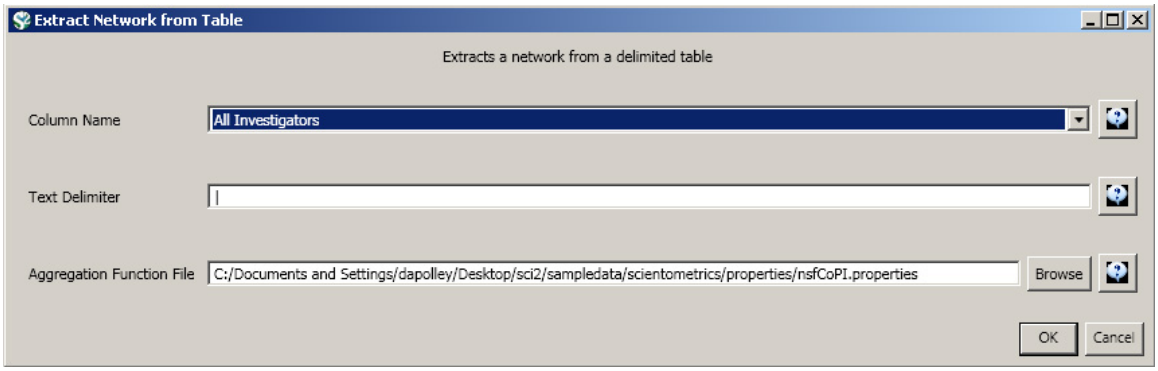


Figure 5.7: Funding profile over time of Michael McRobbie

The horizontal bar graph visualizations in Figures 5.5, 5.6, and 5.7 make it easy to see the timespan of different researchers, as well as the types and volume of grants they generally receive (e.g., many small grants or a handful of large ones). From here, it may be useful to compare their Co-PI networks and look more closely at award totals. Select each dataset in the Data Manager window and run *'Data Preparation > Extract Co-Occurrence Network'* usin g these parameters:

Aggregate Function File

Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



Run *'Visualization > Networks > GUESS'* on each generated network to visualize the resulting Co-PI relationships. Select '*GEM*' from the layout menu to organize the nodes and edges.

To color and size the nodes and edges using the default Co-PI visualization theme, run '***yoursci2directory****/scripts/GUESS/co-PI-nw.py*' from '*Script > Run Script ...*'.

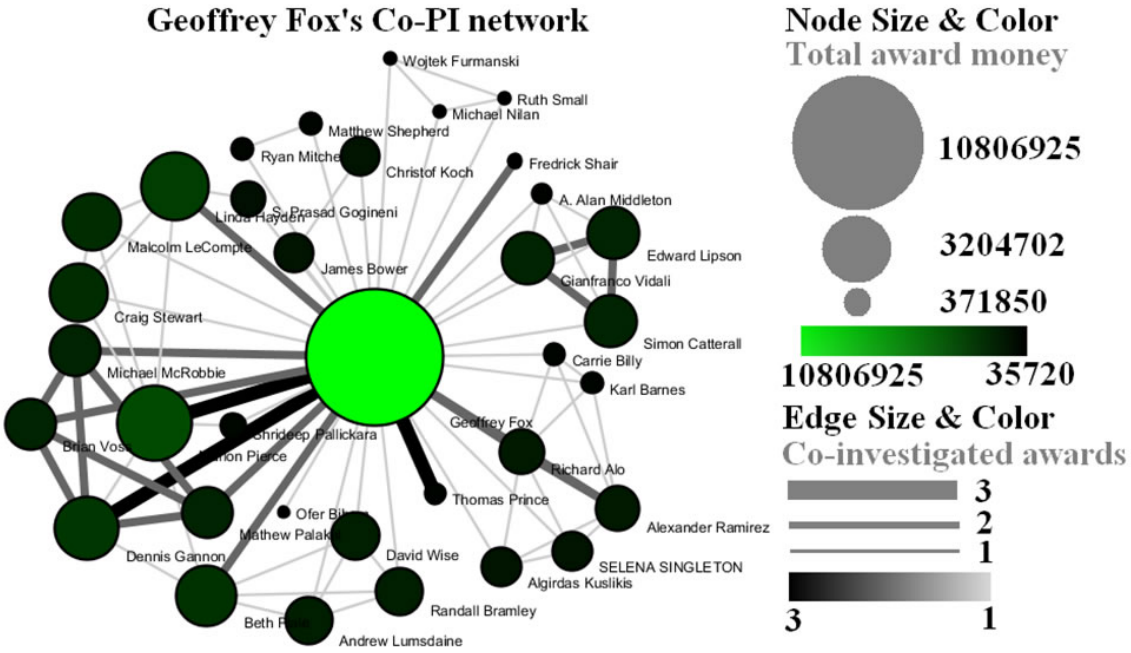


Figure 5.8: Co-PI network of Geoffrey Fox in Indiana University

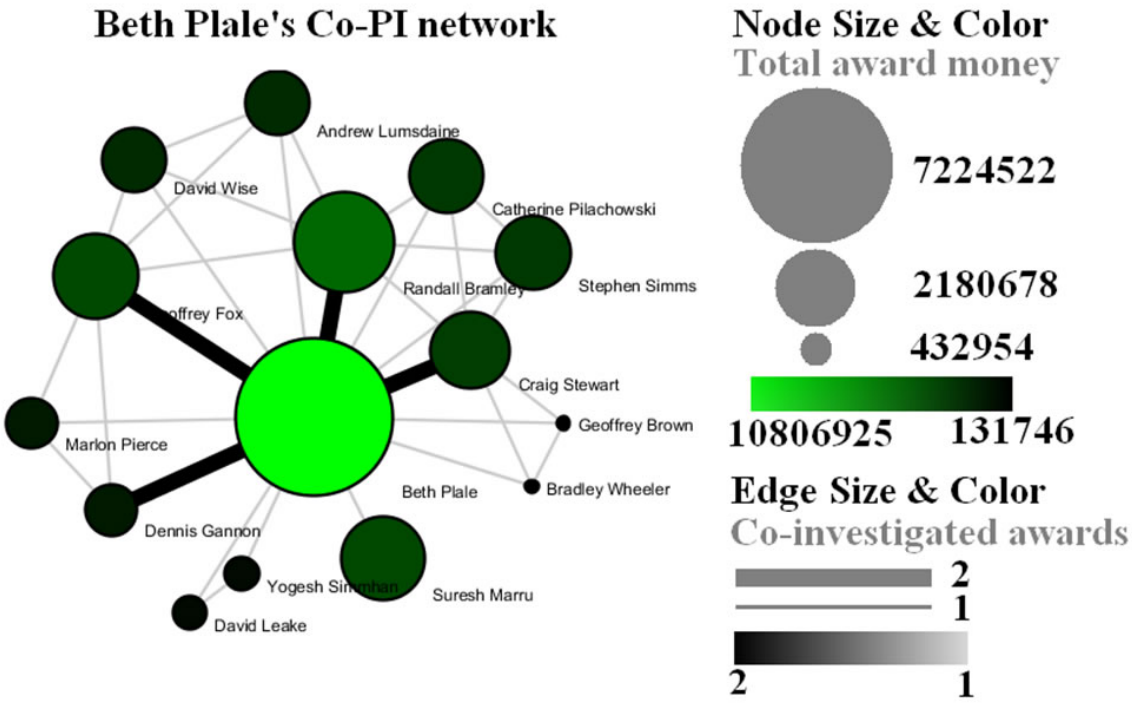


Figure 5.9: Co-PI network of Beth Plale in Indiana University

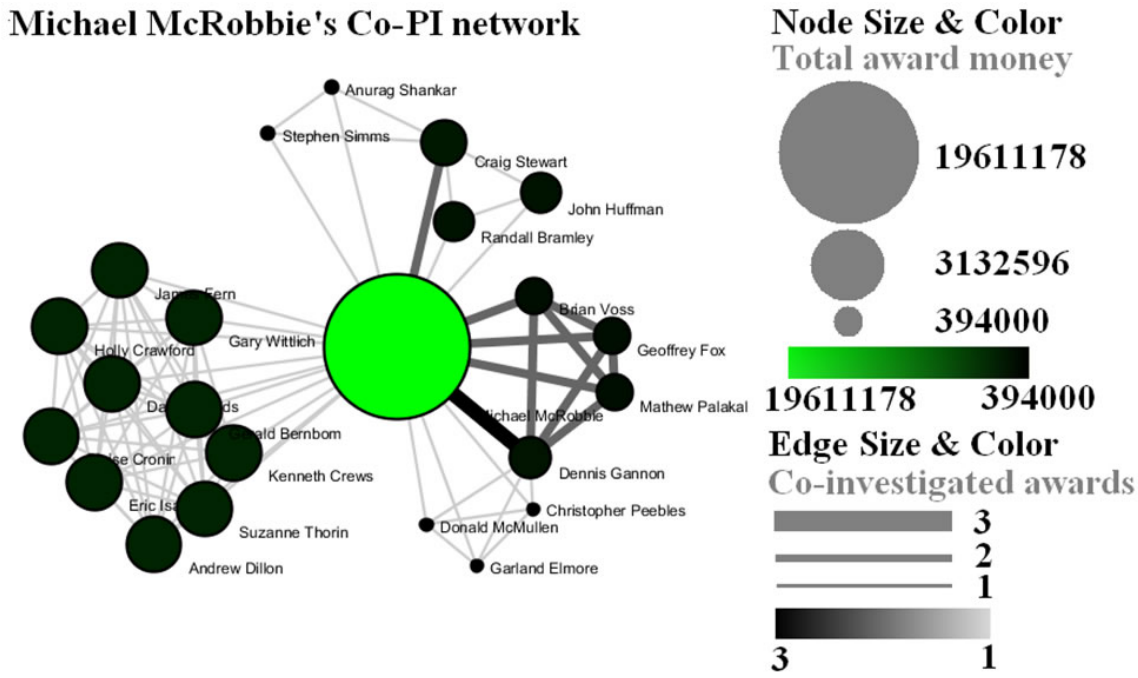


Figure 5.10: Co-PI network of Michael McRobbie in Indiana University

To see the log file from this workflow save the 5.1.3 Funding Profiles of Three Researchers at Indiana University (NSF Data) log file.

5.1.4 Studying Four Major NetSci Researchers (ISI Data)

FourNetSciResearchers.isi	
Time frame:	1955-2007
Region(s):	Miscellaneous
Topical Area(s):	Network Science
Analysis Type(s):	Paper Citation Network, Co-Author Network, Bibliographic Coupling Network, Document Co-Citation Network, Word Co-Occurrence Network

5.1.4.1 Paper-Paper (Citation) Network

Load the file using '*File* > *Load*' and following this path: *'**yoursci2directory**/sampledata/scientometrics/isi/FourNetSciResearchers.isi'* (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). A table of all records and a table of 361 records with unique ISI ids will appear in the Data Manager. In this "clean" file, each original record now has a "Cite Me As" attribute that is constructed from the first author, publication year (PY), journal abbreviation (J9), volume (VL), and beginning page (BP) fields of its ISI record. This "Cite Me As" attribute will be used when matching paper and reference records.

New ISI File Format

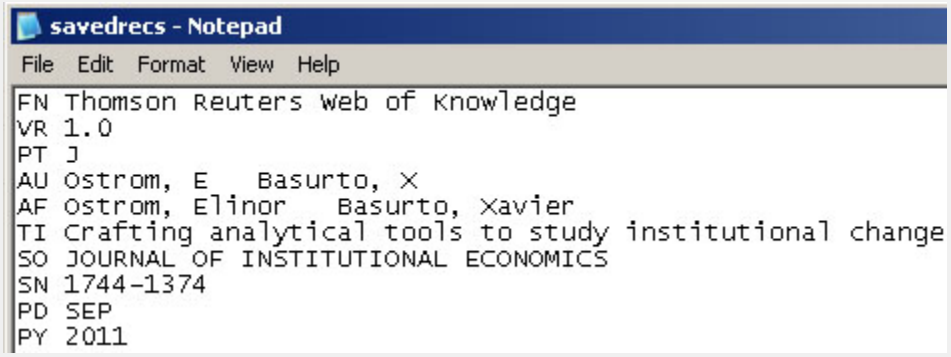
Web of Science made a change to their output format in September, 2011. Older versions of Sci2 tool (Older than v0.5.2 alpha) may refuse to load these new files, with an error like "Invalid ISI format file selected."

Sci2 solution

If you are using an older version of the Sci2 tool, you can download the WOS-plugins.zip file and unzip the JAR files into your sci2/plugins/ directory. Restart Sci2 to activate the fixes. You can now load the downloaded ISI files into the Sci2 without any additional step. If you are using the old Sci2 tool you will need to follow the guidelines below before you can load the new WOS format file into the tool.

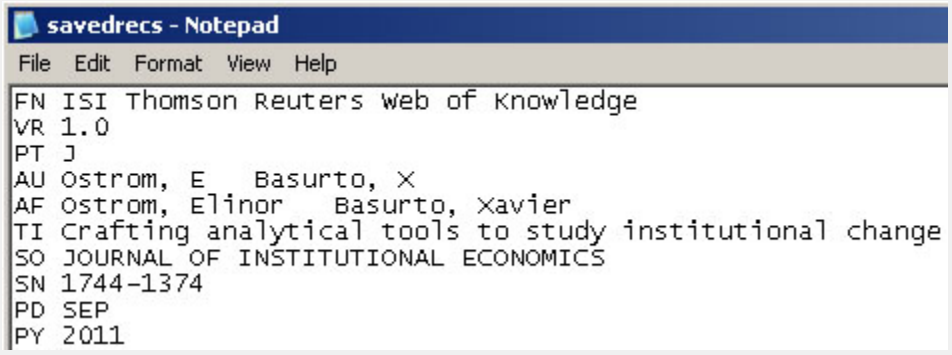
You can fix this problem for individual files by opening them in Notepad (or your favorite text editor). The file will start with the words:

Original ISI file:



Just add the word ISI.

Updated ISI file:



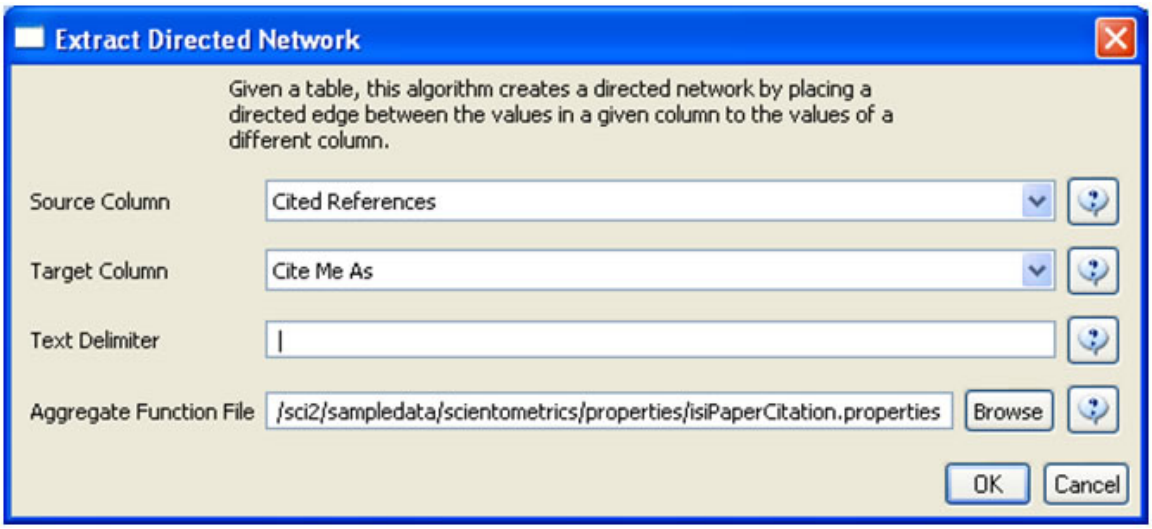
And then Save the file.

The ISI file should now load properly. More information on the ISI file format is available here (http://wiki.cns.iu.edu/display/CISHELL/ISI+%28*.isi%29).

To extract the paper citation network, select the *'361 Unique ISI Records'* table and run *'Data Preparation* > *Extract Directed Network'* using the parameters :

Aggregate Function File

Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



The result is a directed network of paper citations in the Data Manager. Each paper node has two citation counts. The local citation count (LCC) indicates how often a paper was cited by papers in the set. The global citation count (GCC) equals the times cited (TC) value in the original ISI file. Only references from other ISI records count towards an ISI paper's GCC value. Currently, the Sci^2 Tool sets the GCC of references to -1 (except for references that are not also ISI records) to prune the network to contain only the original ISI records.

To view the complete network, select the "Network with directed edges from Cited References to Cite Me As" in the Data Manager and run *'Visualization > Networks > GUESS'* and wait until the network is visible and centered. Because the FourNetSciResearchers dataset is so large, the visualization will take some time to load, even on powerful systems.

Select *' Layout > GEM'* to group related nodes in the network. Again, because of the size of the dataset, this will take some time. Use *'Layout > Bin Pack'* to "pack" the nodes and edges, bringing them closer together for easier analysis and comparison. To size and color-code nodes in the "Graph Modifier," use the following workflow:

1. Resize Linear > Nodes > globalcitationcount> From: 1 To: 50 > When the nodes have no 'globalcitationcount': 0.1 > Do Resize Linear
2. Colorize > Nodes > globalcitationcount > From:  To:  > When the nodes have no 'globalcitationcount': 0.1 >  >Do Colorize
3. Colorize > Edges > weight > From (select the "RGB" tab) 127, 193, 65 To: (select the "RGB" tab) 0, 0, 0
4. Type in Interpreter:

```
>for n in g.nodes:
     n.strokecolor = n.color
```

Or, select the 'Interpreter' tab at the bottom, left-hand corner of the GUESS window, and enter the command lines:

```
> resizeLinear(globalcitationcount,1,50)
> colorize(globalcitationcount,gray,black)
> for e in g.edges:
      e.color="127,193,65,255"
```

Note: The Interpreter tab will have '>>>' as a prompt for these commands. It is not necessary to type '>" at the beginning of the line. You should type each line individually and press "Enter" to submit the commands to the Interpreter.
This will result in nodes which are linearly sized and color coded by their GCC, connected by green directed edges, as shown in Figure 5.11 (left). Any numeric node attribute within the network can be used to code the nodes. To view the available attributes, mouse over a node. The GUESS interface supports pan and zoom, node selection, and details on demand. For more information, refer to the GUESS tutorial at http://nwb.cns.iu.edu/Docs/GettingStartedGUESSNWB.pdf

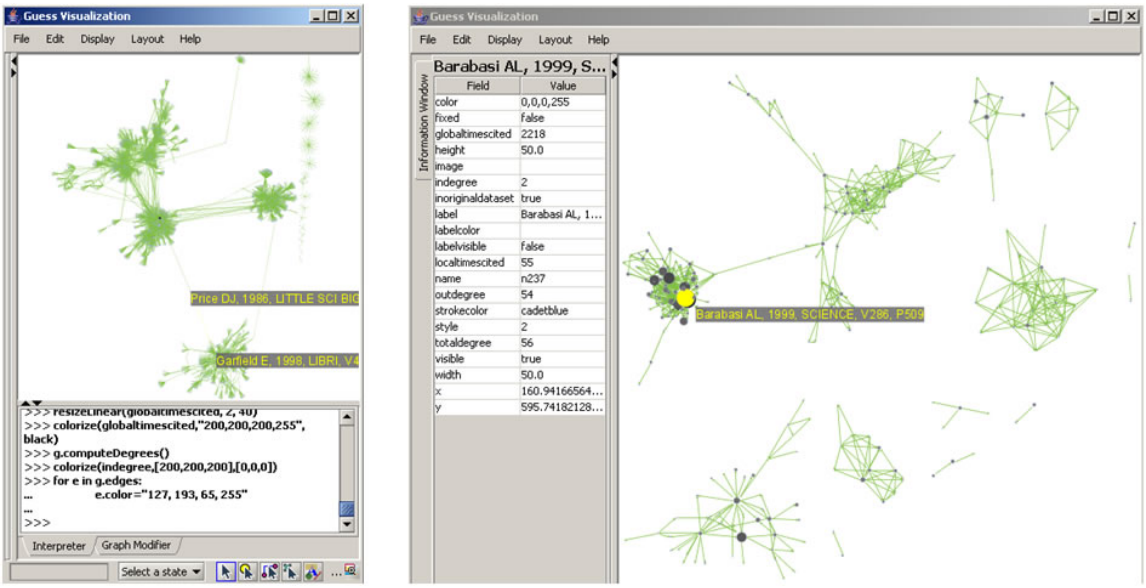
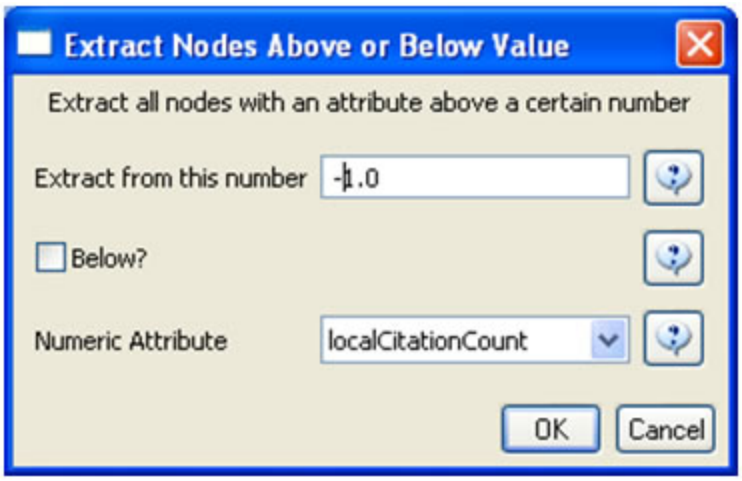


Figure 5.11: Directed, unweighted paper-paper citation network for 'FourNetSciResearchers' dataset with all papers and references in the GUESS user interface (left) and a pruned paper-paper citation network after removing all references and isolates (right)

The complete network can be reduced to papers that appeared in the original ISI file by deleting all nodes that have a GCC of -1. Simply run *'Preprocessing > Networks > Extract Nodes Above or Below Value'* with parameter values:



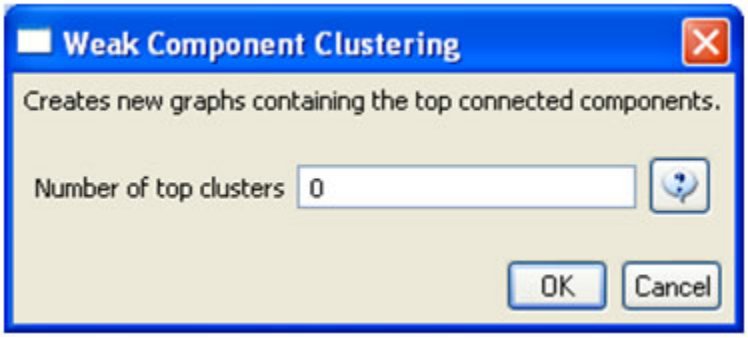
The resulting network is unconnected, i.e., it has many subnetworks many of which have only one node. These single unconnected nodes, also called isolates, can be removed using *'Preprocessing > Networks > Delete Isolates'*. Deleting isolates is a memory intensive procedure. If you experience problems at this step, refer to Section 3.4 Memory Allocation.

The *'FourNetSciResearchers'* dataset has exactly 65 isolates. Removing those leaves 12 networks shown in Figure 5.11 (right) using the same color and size coding as in Figure 5.11 (left). Using *'View > Information Window'* in GUESS reveals detailed information for any node or edge.
Alternatively, nodes could have been color and/or size coded by their degree using, e.g.:

```
> g.computeDegrees()
> colorize(outdegree,gray,black)
```

Note that the outdegree corresponds to the LCC within the given network while the indegree reflects the number of references, helping to visually identify review papers.

The complete paper-paper-citation network can be split into its subnetworks using *'Analysis > Networks > Unweighted & Directed > Weak Component Clustering'* with the default values:



The largest component has 2407 nodes; the second largest, 307; the third, 13; and the fourth has 7 nodes. The largest component is shown in Figure 5.12. The top 20 papers, by times cited in ISI, have been labeled using

```
> toptc = g.nodes[:]
> def bytc(n1, n2):
      return cmp(n1.globalcitationcount, n2.globalcitationcount)
> toptc.sort(bytc)
> toptc.reverse()
> toptc
> for i in range(0, 20):
      toptc[i].labelvisible = true
```

Alternatively, run '*Script > Run Script*' and select '***yoursci2directory****/scripts/GUESS/paper-citation-nw.py*'.

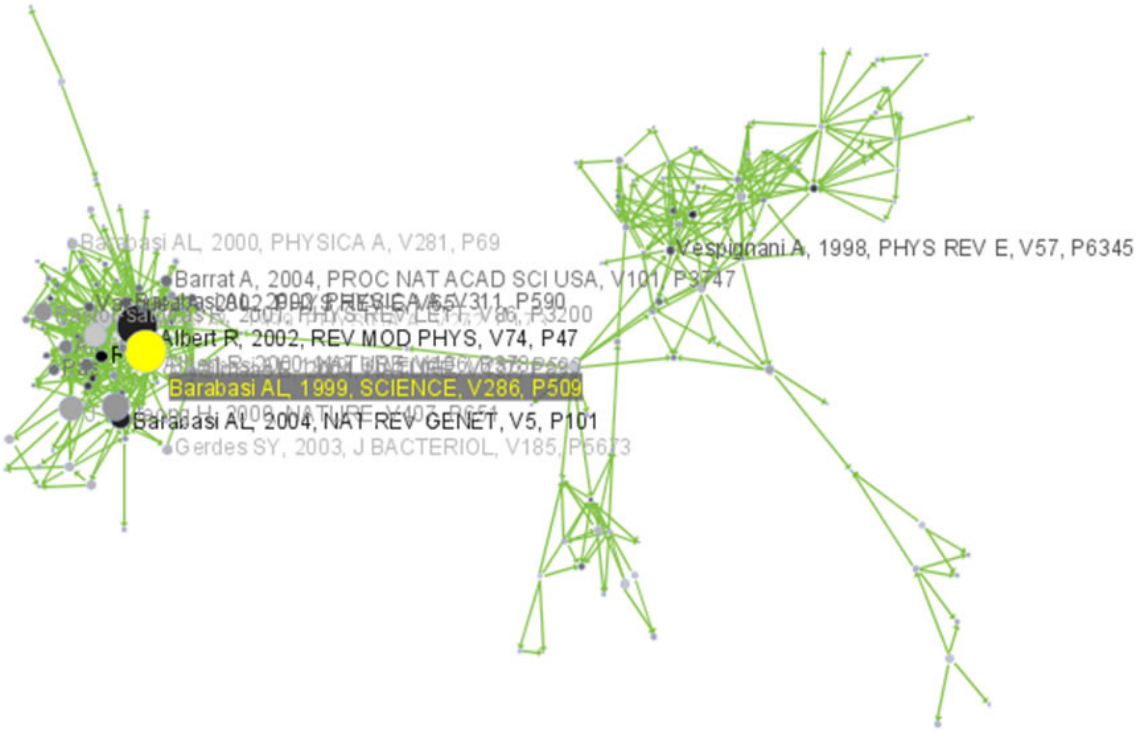


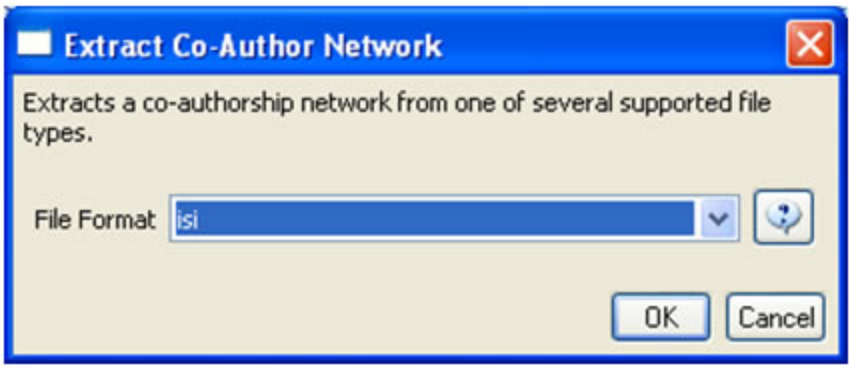
Figure 5.12: Giant components of the paper citation network

Compare the result with Figure 5.11 and note that this network layout algorithm – and most others – are non-deterministic: different runs lead to different layouts. That said, all layouts aim to group connected nodes into spatial proximity while avoiding overlaps of unconnected or sparsely connected sub-networks.

To see the log file from this workflow save the 5.1.4.1 Paper-Paper (Citation) Network log file.

5.1.4.2 Author Co-Occurrence (Co-Author) Network

To produce a co-authorship network in the Sci^2 Tool, select the table of all 361 unique ISI records from the *'FourNet SciResearchers'* dataset in the Data Manager window. Run *'Data Preparation > Extract Co-Author Network'* using the parameter:

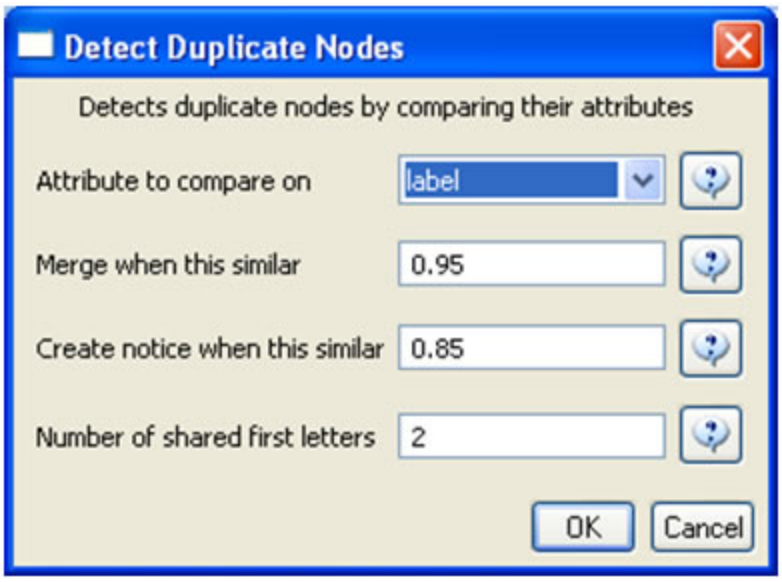


The result is two derived files in the Data Manager window: the "Extracted Co-Authorship Network" and an "Author information" table (also known as a "merge table"), which lists unique authors. In order to manually examine and edit the list of unique authors, open the merge table in your default spreadsheet program. In the spreadsheet, select all records, including "label," "timesCited," "numberOfWorks," "uniqueIndex," and "combineValues," and sort by "label." Identify names that refer to the same person. In order to merge two names, first delete the asterisk ('*') in the "combineValues" column of the duplicate node's row. Then, copy the "uniqueIndex" of the name that should be kept and paste it into the cell of the name that should be deleted. Resave the revised table as a .csv file and reload it. Select both the merge table and the network and run '*Data Preparation* > *Update Network by Merging Nodes*'. Table 5.2 shows the result of merging "Albet, R" and "Albert, R": "Albet, R" will be deleted and all of the node linkages and citation counts will be added to "Albert, R".

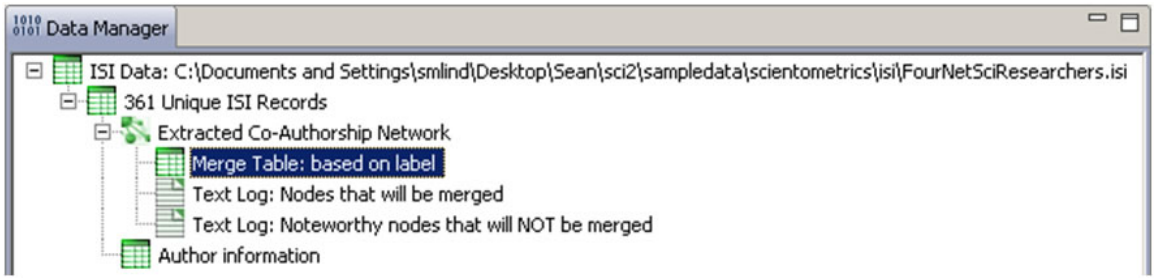
label	timesCited	numberOfWorks	uniqueIndex	combineValues
Abt, HA	3	1	142	*
Alava, M	26	1	196	*
Albert, R	7741	17	60	*
Albet, R	16	1	60	

Table 5.2: Merging of author nodes using the merge table

A merge table can be automatically generated by applying the Jaro distance metric (Jaro, 1989, 1995) available in the open source Similarity Measure Library (http://sourceforge.net/projects/simmetrics/) to identify potential duplicates. In the Sci^2 Tool, simply select the co-author network and run *'Data Preparation* > *Detect Duplicate Nodes '.* using the parameters:



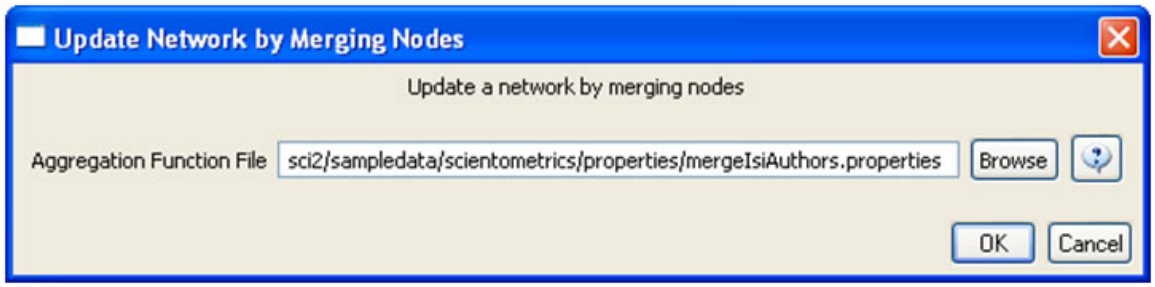
The result is a merge table that has the very same format as Table 5.2, together with two textual log files:



The log files describe, in a more human-readable form, which nodes will be merged or not merged. Specifically, the first log file provides information regarding which nodes will be merged, while the second log file lists nodes which are similar but will not be merged. The automatically generated merge table can be further modified as needed.

In sum, unification of author names can be done manually or automatically, independently or in conjunction with other data manipulation. It is recommended that users create the initial merge table automatically and fine-tune it as needed. Note that the same procedure can be used to identify duplicate references – simply select a paper-citation network and run *'Data Preparation* > *Detect Duplicate Nodes'* using the same parameters as above and a merge table for references will be created.

To merge identified duplicate nodes, select both the "Extracted Co-Authorship Network" and "Merge Table: based on label" by holding down the 'Ctrl' key. Run *'Data Preparation* > *Update Network by Merging Nodes'*. This will produce an updated network as well as a report describing which nodes were merged. To complete this workflow, an aggregation function file must also be selected from the pop-up window:



Aggregation files can be found in the Sci2 sample data. Follow this path *"sci2/sampledata/scientometrics/properties*"

and select the mergeIsiAuthors.properties.

The updated co-authorship network can be visualized using *'Visualization > Networks > GUESS'*, (See section 4.9.4.1 GUESS Visualizations for more information regarding GUESS).
Figure 5.13 shows the layout of the combined *'FourNetSciResearchers'* dataset after it was modified using the following commands in the "Interpreter":

```
> resizeLinear(numberOfWorks,1,50)
> colorize(numberOfWorks,gray,black)
> for n in g.nodes:
      n.strokecolor = n.color
> resizeLinear(numberOfCoAuthored_works, .25, 8)
> colorize(numberOfCoAuthoredworks, "127,193,65,255", black)
> nodesbynumworks = g.nodes[:]
> def bynumworks(n1, n2):
     return cmp(n1.numberofworks, n2.numberofworks)
> nodesbynumworks.sort(bynumworks)
> nodesbynumworks.reverse()
> for i in range(0, 50):
      nodesbynumworks[i].labelvisible = true
```

Alternatively, run '*Script > Run Script ...*' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*'.

For both workflows described above, the final step should be to run *'Layout > GEM'* and then *'Layout > Bin Pack'* to give a better representation of node clustering.

In the resulting visualization, author nodes are color and size coded by the number of papers per author. Edges are color and thickness coded by the number of times two authors wrote a paper together. The remaining commands identify the top 50 authors with the most papers and make their name labels visible.

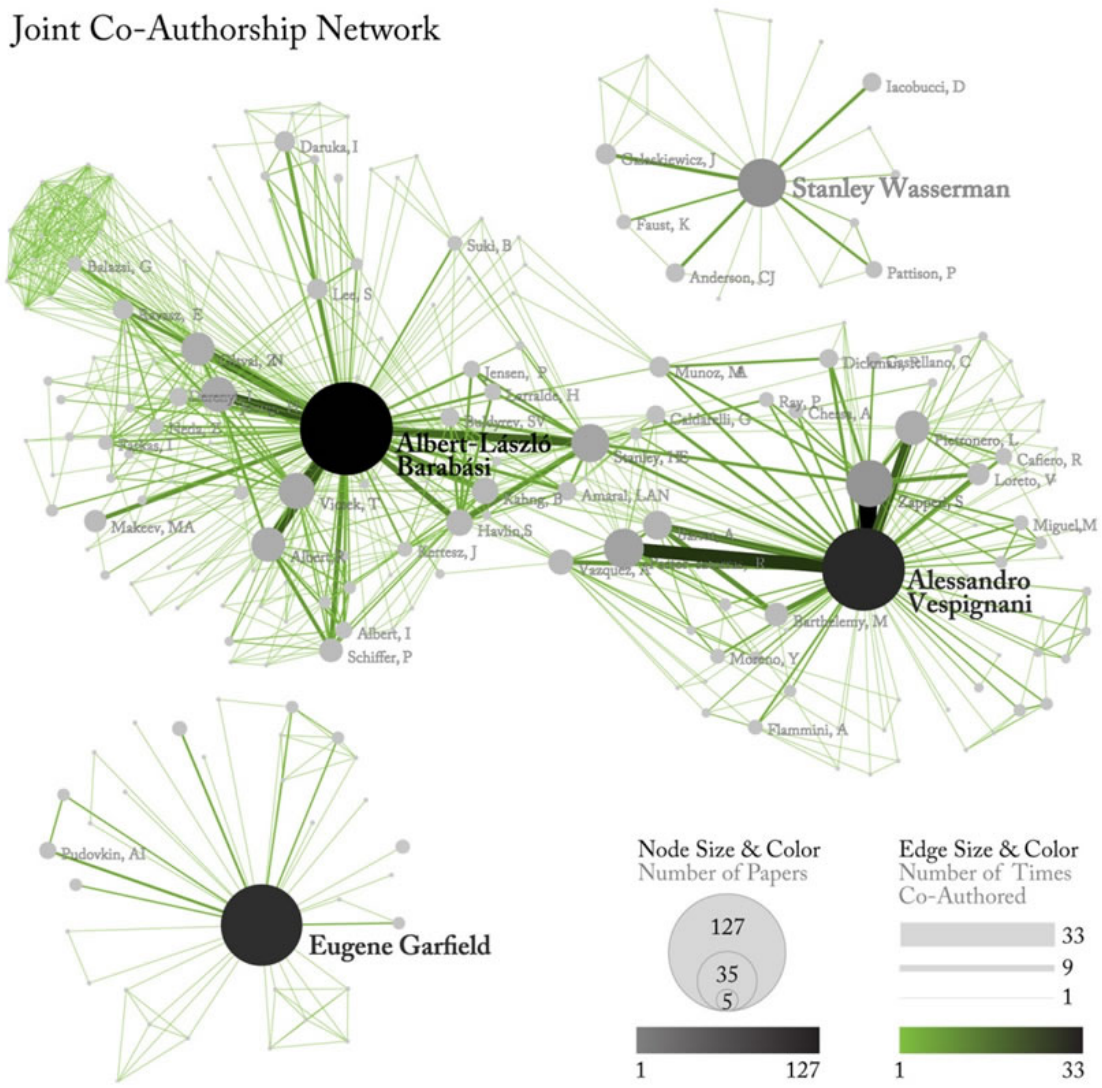


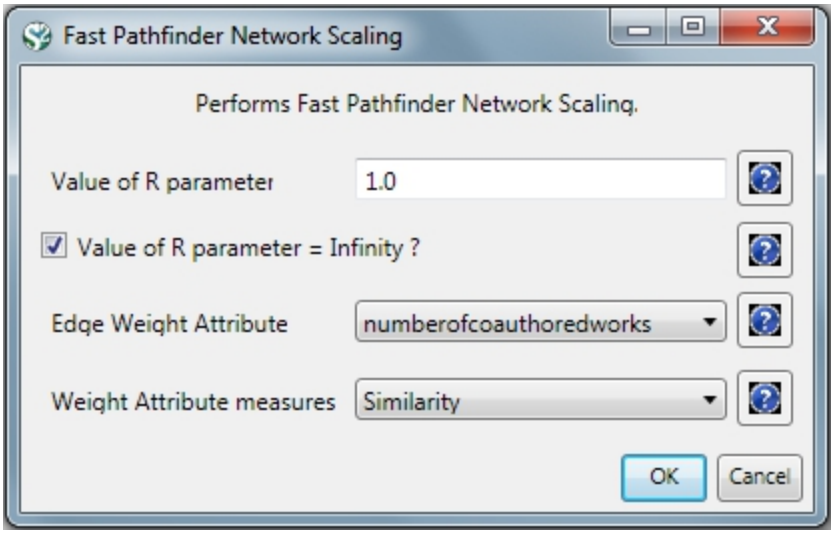
Figure 5.13: Undirected, weighted co-author network for 'FourNetSciResearchers' dataset

GUESS supports the repositioning of selected nodes. Multiple nodes can be selected by holding down the 'Shift' key and dragging a box around specific nodes. The final network can be saved via *'GUESS: File > Export Image'* and opened in a graphic design program to add a title and legend and to modify label sizes. The image above was modified using Photoshop.

Pruning the network

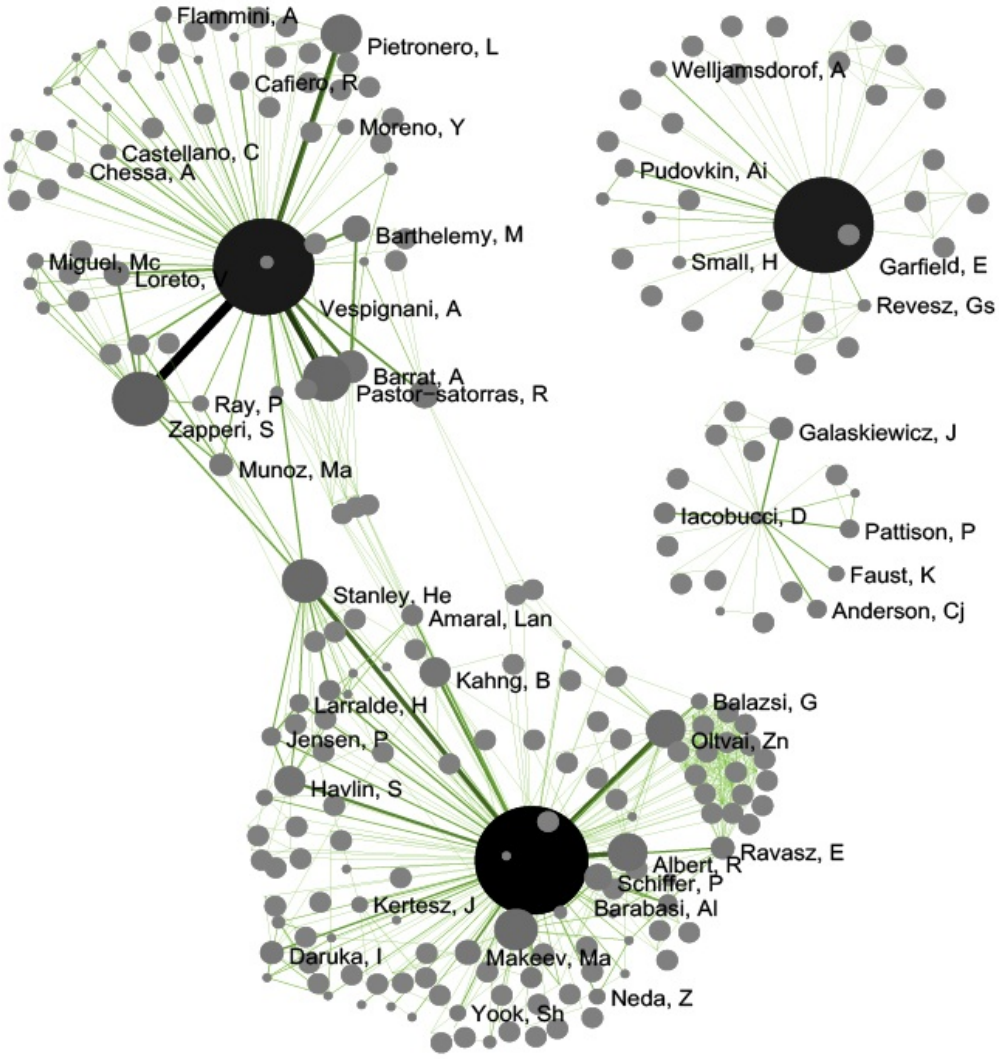
Pathfinder algorithms take estimates of the proximity's between pairs of items as input and define a network representation of the items that preserves only the most important links. The value of Pathfinder Network Scaling for networks lies in its ability to reduce the number of links in meaningful ways, often resulting in a concise representation of clarified proximity patterns.

Given the co-Authorship network extracted at workflow 5.1.4.2, run '*Preprocessing > Networks > Fast Pathfinder Network Scaling' with the following parameters:*

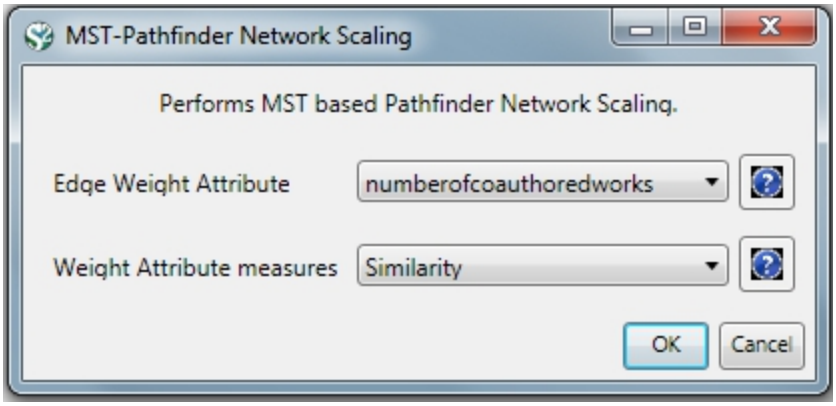


The number of edges reduced from 872 to 731. The updated co-authorship network can be visualized using *'Visualization > Networks >* GUESS*'*.

Inside GUESS, run 'Script > Run Script ...' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*'. Also run *'Layout > GEM'* and then *'Layout > Bin Pack'* to give a better representation of node clustering. The pruned network looks like this:

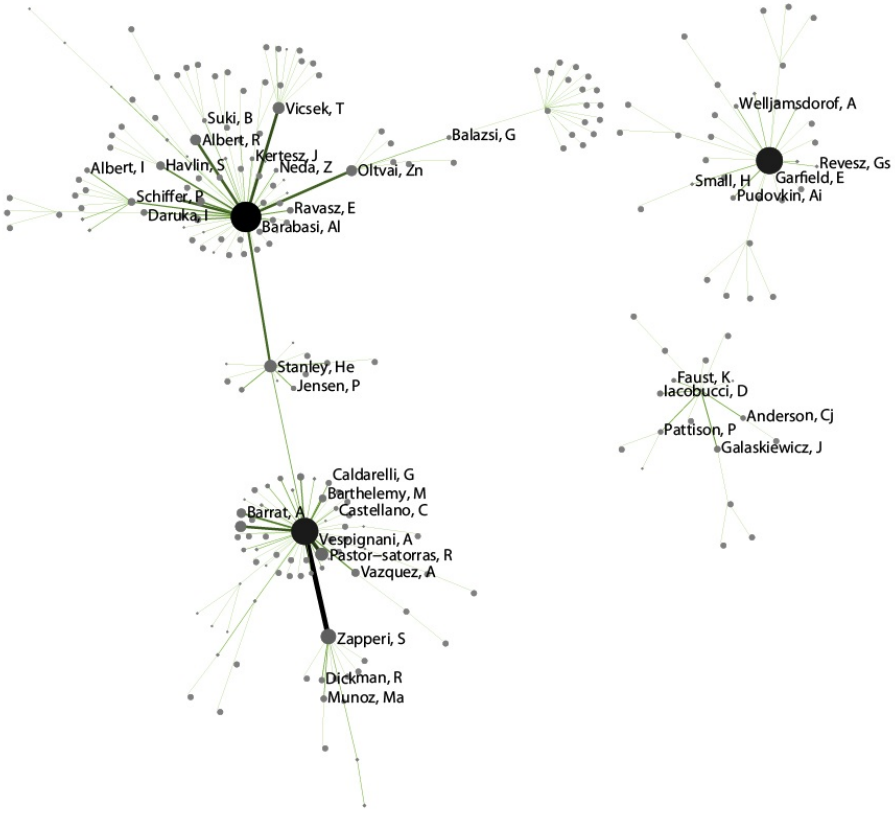


If the network being processed is undirected, which is the case, then MST-Pathfinder Network Scaling can be used to prune the networks. This will give results in 30 times faster than Fast Pathfinder Network Scaling. Also we have found that networks that have a low standard deviation for edge weights or if many of the edge weights are equal to the minimum edge weight, then the network might not be scaled as much as expected when using Fast Pathfinder Network Scaling. To see this behavior, run '*Preprocessing > Networks > MST-Pathfinder Network Scaling*' with the network named 'Updated network' selected with the following parameters:



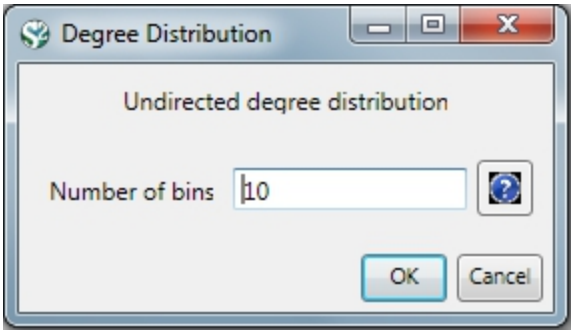
The number of edges reduced from 872 to 235 for the network pruned with the MST-Pathfinder Network Scaling alg orithm, while the reduction for the Fast Pathfinder Network Scaling algorithm was from 872 to 731 only.

Visualize this network using *'Visualization > Networks >* GUESS*'*. Inside GUESS, run 'Script > Run Script ...' and select ' ***yoursci2directory****/scripts/GUESS/co-author-nw.py*' again. Also run *'Layout > GEM'* and then *'Layout > Bin Pack'* to give a better representation of node clustering. The pruned network with the algorithm looks like this:



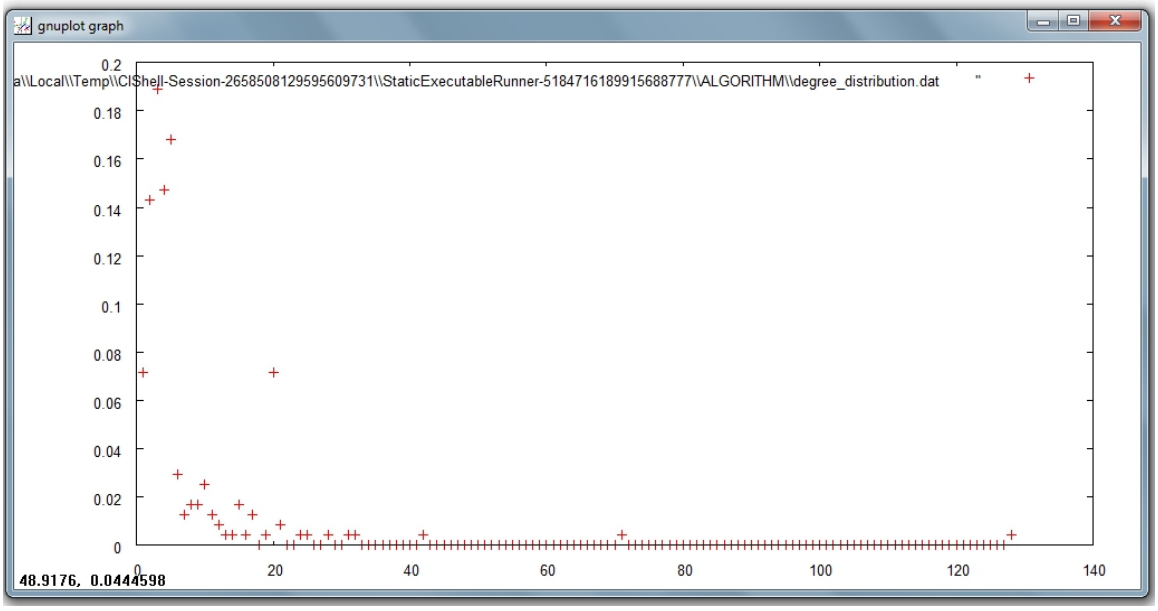
Degree Distribution

Given the co-Authorship network extracted at workflow 5.1.4.2 (name 'Updated Network' at the Data Manager), run ' *Analysis > Networks > Unweighted & Undirected >* Degree Distribution*' with the following parameter:*



This algorithm generates two output files, corresponding to two different ways of partitioning the interval spanned by the values of degree.

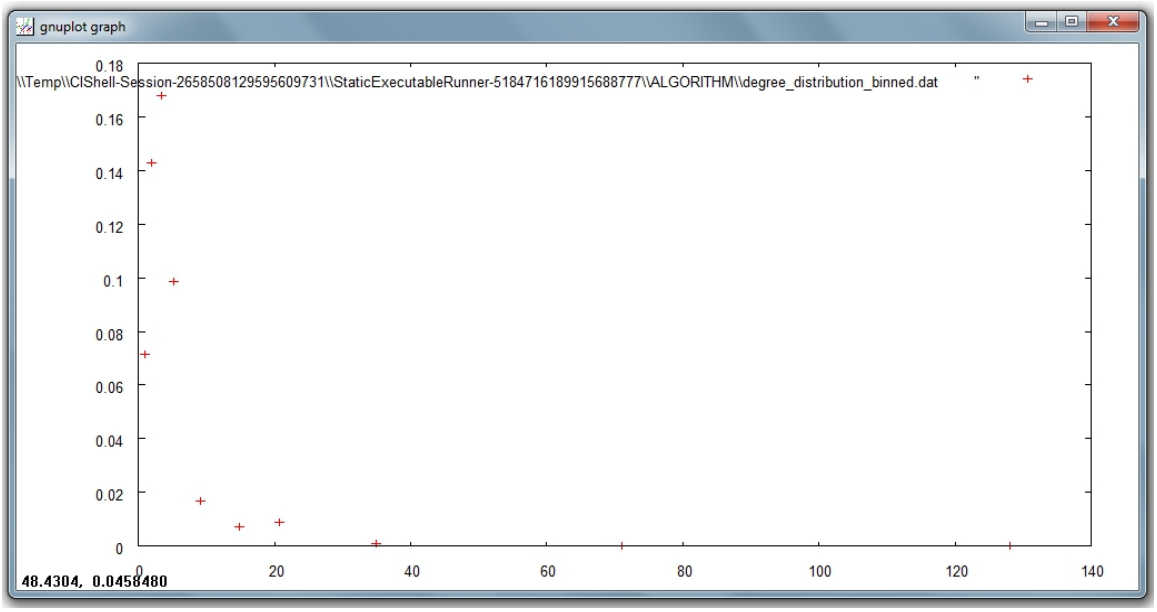
In the first output file named '*Distribution of degree for network at study (equal bins)*', the occurrence of any degree value between the minimum and the maximum is estimated and divided by the number of nodes of the network, so to obtain the probability: the output displays all degree values in the interval with their probabilities. Visualize this first output file with '*Visualization* > *General* > *GnuPlot*':



The second output file named '*Distribution of degree for network at study (logarithmic bins)*' gives the *binned* distribu tion, i.e. the interval spanned by the values of degree is divided into bins whose size grows while going to higher values of the variable. The size of each bin is obtained by multiplying by a fixed number the size of the previous bin. The program calculates the fraction of nodes whose degree falls within each bin. Because of the different sizes of the bins, these fractions must be divided by the respective bin size, to have meaningful averages.

This second type of output file is particularly suitable to study skewed distributions: the fact that the size of the bins grows large for large degree values compensates for the fact that not many nodes have high degree values, so it suppresses the fluctuations that one would observe by using bins of equal size. On a double logarithmic scale, which is very useful to determine the possible power law behavior of the distribution, the points of the latter will appear equally spaced on the x-axis.

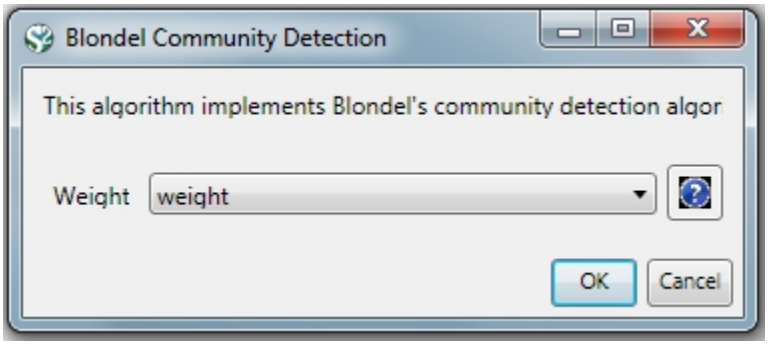
Visualize also this second output file with '*Visualization* > *General* > *GnuPlot*':



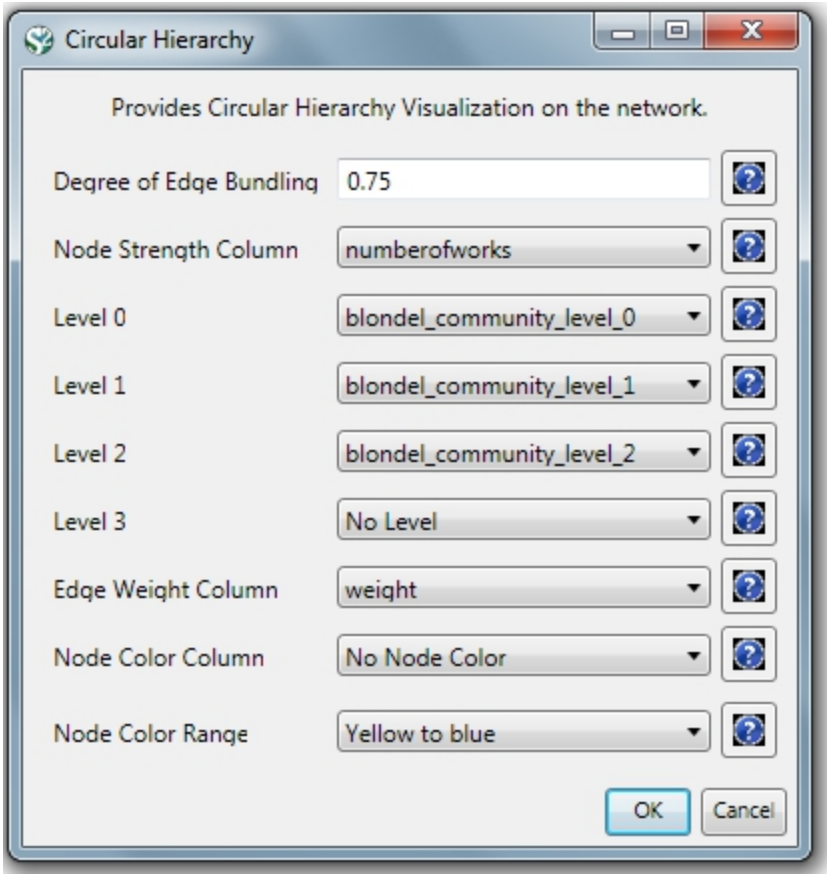
Community Detection

Community Detection algorithms look for subgraphs where nodes are highly interconnected among themselves and poorly connected with nodes outside the subgraph. Many community detection algorithms are based on the optimization of the modularity - a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities. The Blondel Community Detection finds high modularity partitions of large networks in short time and that unfolds a complete hierarchical community structure for the network, thereby giving access to different resolutions of community detection.

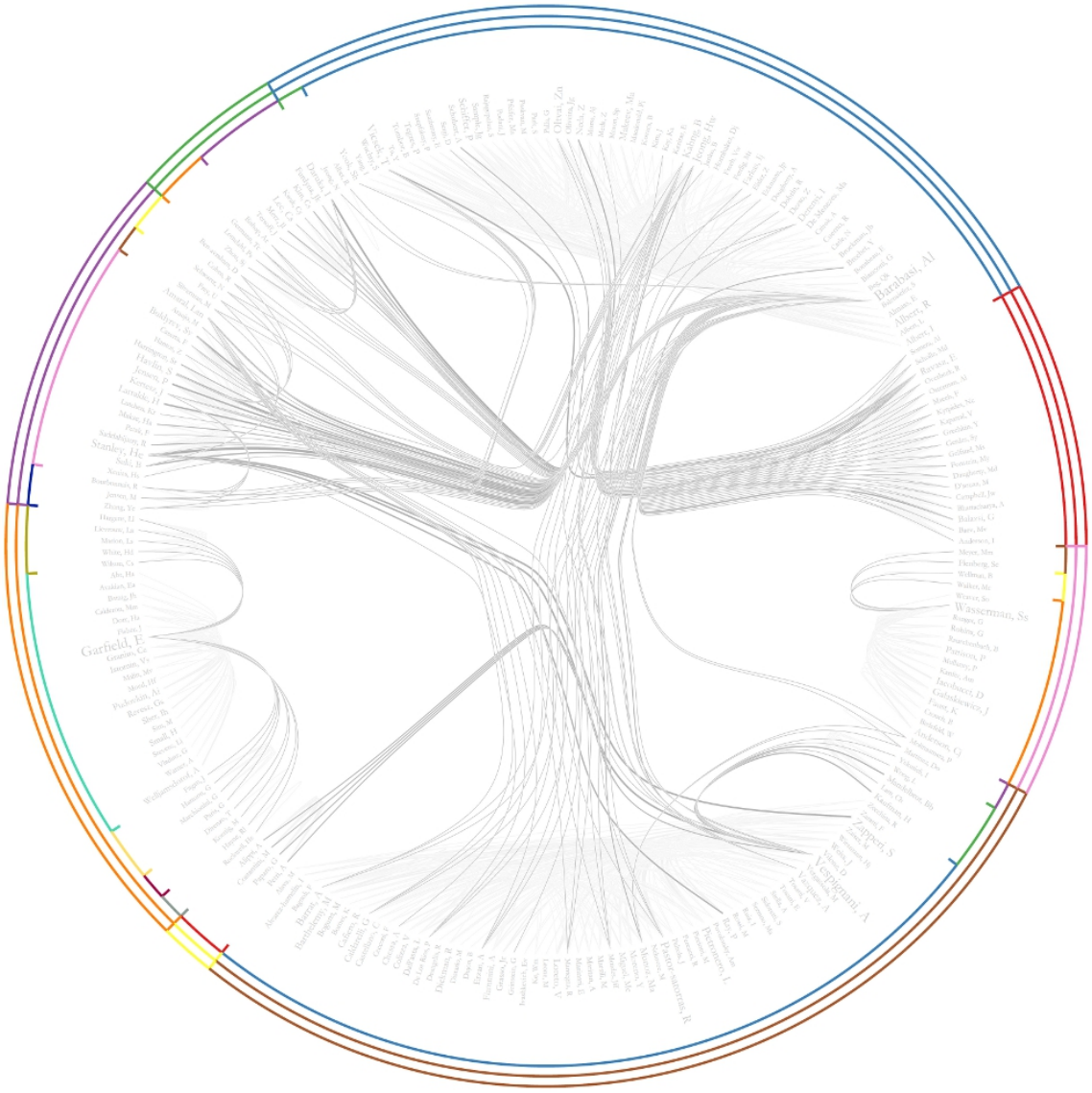
Run 'Analysis > Networks > Weighted & Undirected > Blondel Community Detection' with the following parameter:



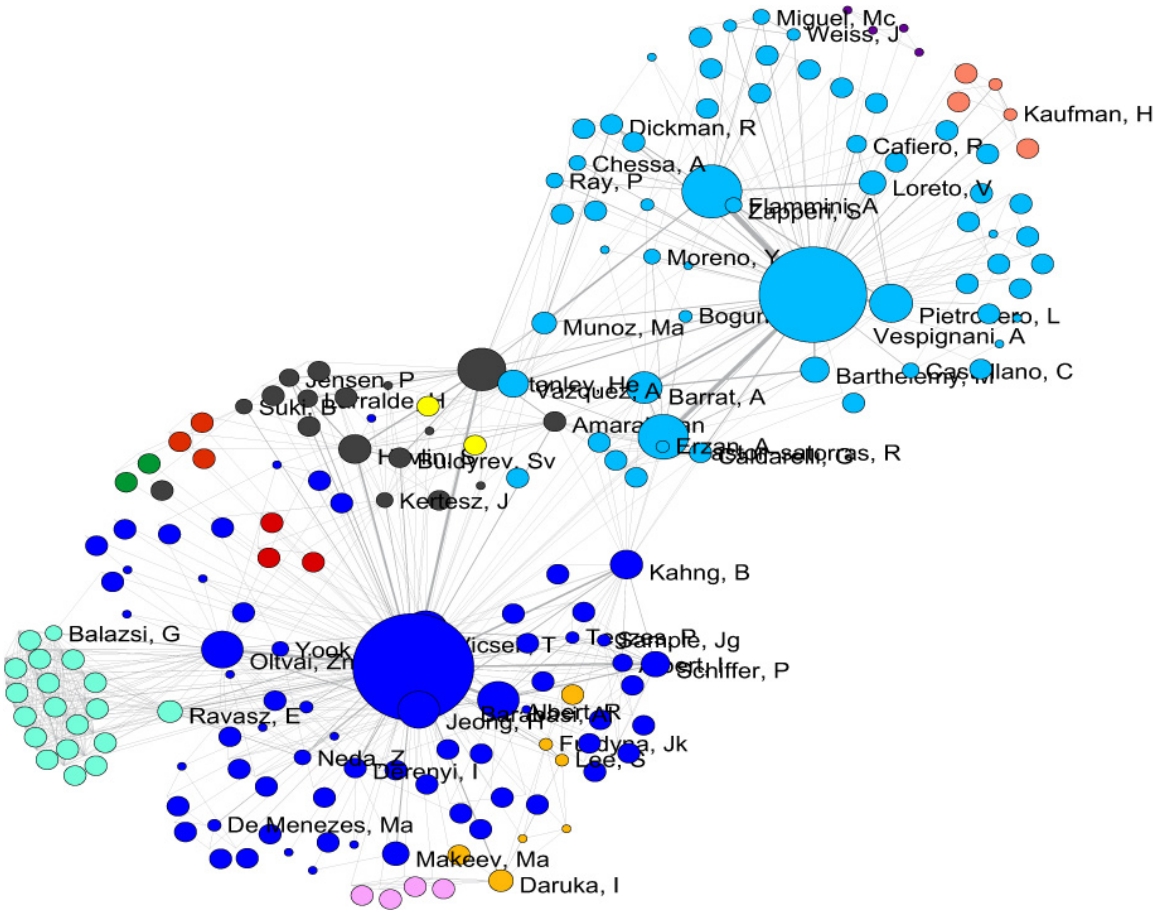
This will generate a network named 'With community attributes' in the Data Manager. To visualize this network run 'Visualization > Networks > Circular Hierarchy' with the following parameters:



The generated postscript file "CircularHierarchy_With community attributes.ps" can be viewed using Adobe Distiller or GhostViewer (see section 2.4 Saving Visualizations for Publication).



To view to the network with community attributes with in GUESS select the network used to generate the image above and run *'Visualization > Networks >* GUESS.'



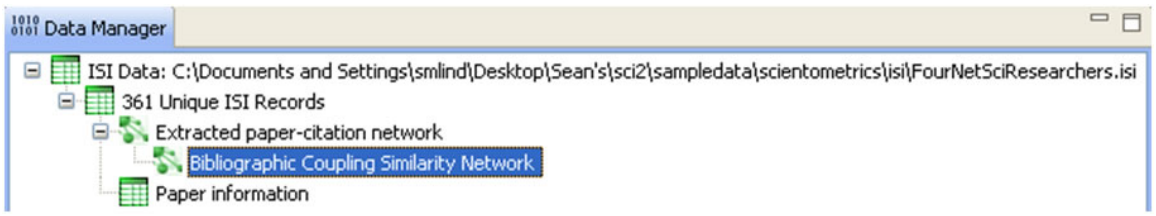
To see the log file from this workflow save the 5.1.4.2 Author Co-Occurrence (Co-Author) Network log file.

5.1.4.3 Cited Reference Co-Occurrence (Bibliographic Coupling) Network

In Sci^2, a bibliographic coupling network is derived from a directed paper citation network (see section 4.9.1.1.1 Document-Document (Citation) Network).

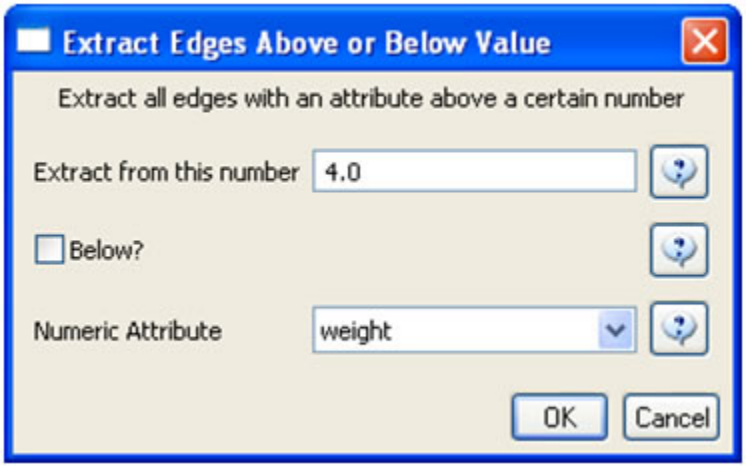
Load the file using '*File* > *Load*' and following this path: *'yoursci2directory/sampledata/scientometrics/isi/FourNetSci Researchers.isi'*. A table of all records and a table of 361 records with unique ISI ids will appear in the Data Manager.

Select the "361 Unique ISI Records" in the Data Manager and run *'Data Preparation* > *Extract Paper Citation Network.'* Select "Extracted Paper Citation Network" and run *'Data Preparation* > *Extract Reference Co-Occurrence (Bibliographic Coupling) Network.'*

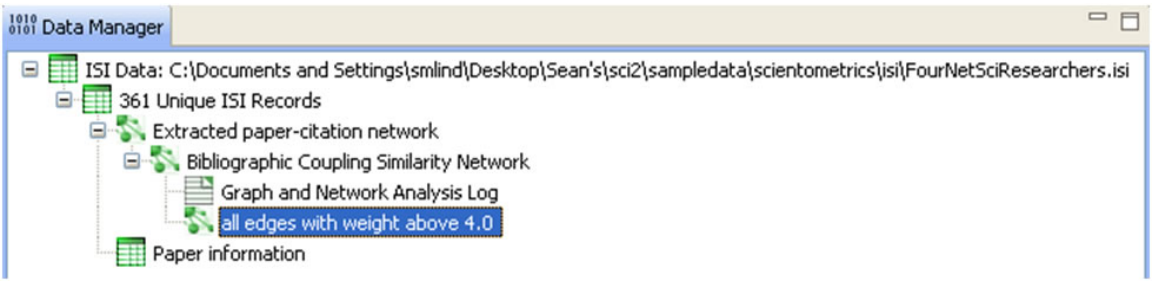


Running *'Analysis* > *Networks* > *Network Analysis Toolkit (NAT)'* reveals that the network has 5,342 nodes (5,013 of which are isolate nodes) and 6,277 edges.

In the "Bibliographic Coupling Similarity Network," edges with low weights can be eliminated by running *'Preprocessing* > *Networks* > *Extract Edges Above or Below Value'* with the following parameter values:



to produce the following network:



Isolate nodes can be removed running '*Preprocessing* > *Networks* > *Delete Isolates*'. The resulting network has 242 nodes and 1,534 edges in 12 weakly connected components.

This network can be visualized in GUESS; see Figure 5.14. Nodes and edges can be color and size coded, and the top 20 most-cited papers can be labeled by entering the following lines in the GUESS "Interpreter":

```
> resizeLinear(globalcitationcount,2,40)
> colorize(globalcitationcount,gray,black)
> resizeLinear(weight,.25,8)
> colorize(weight, "127,193,65,255", black)
> for n in g.nodes:
      n.strokecolor=n.color
> toptc = g.nodes[:]
> def bytc(n1, n2):
      return cmp(n1.globalcitationcount, n2.globalcitationcount)
> toptc.sort(bytc)
> toptc.reverse()
> toptc
> for i in range(0, 20):
      toptc[i].labelvisible = true
```

Alternatively, run 'GUESS: File > Run Script ...' and select '***yoursci2directory****/scripts/GUESS/reference-co-occurence-nw.py*'.

For both workflows described above, the final step should be to run 'Layout > GEM' and then 'Layout > Bin Pack' to give a better representation of node clustering.

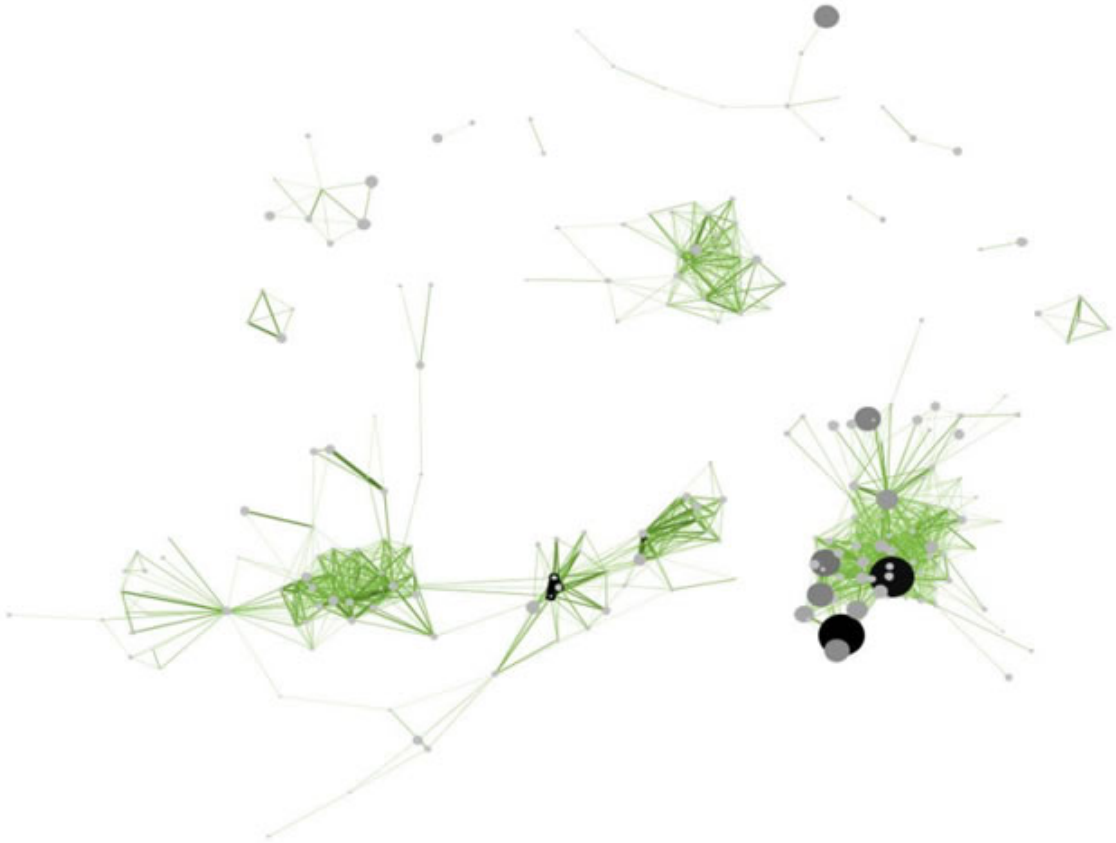


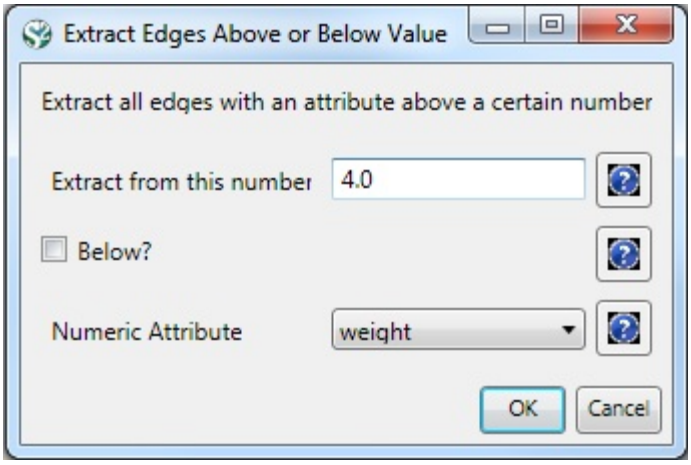
Figure 5.14: Reference co-occurrence network layout for 'FourNetSciResearchers' dataset

To see the log file from this workflow save the 5.1.4.3 Cited Reference Co-Occurrence (Bibliographic Coupling) Network log file.

5.1.4.4 Document Co-Citation Network (DCA)

Load the file using '*Load > File*" and following this path: *'yoursci2directory/sampledata/scientometrics/isi/FourNetSci Researchers.isi'*. A table of all records and a table of 361 records with unique ISI ids will appear in the Data Manager.

Select the "361 Unique ISI Records" and run '*Data Preparation > Extract Paper Citation Network'*. Select the 'Extracted Paper-Citation Network' and run '*Data Preparation > Extract Document Co-Citation Network*.' The co-citation network will have 5,335 nodes (213 of which are isolates) and 193,039 edges. Isolates can be removed by running *'Preprocessing > Networks > Delete Isolates.'* The resulting network has 5122 nodes and 193,039 edges – and is too dense for display in GUESS. Edges with low weights can be eliminated by running '*Preprocessing > Net works > Extract Edges Above or Below Value*' with parameter values:



Here, only edges with a local co-citation count of five or higher are kept. The giant component in the resulting network has 265 nodes and 1,607 edges. All other components have only one or two nodes.

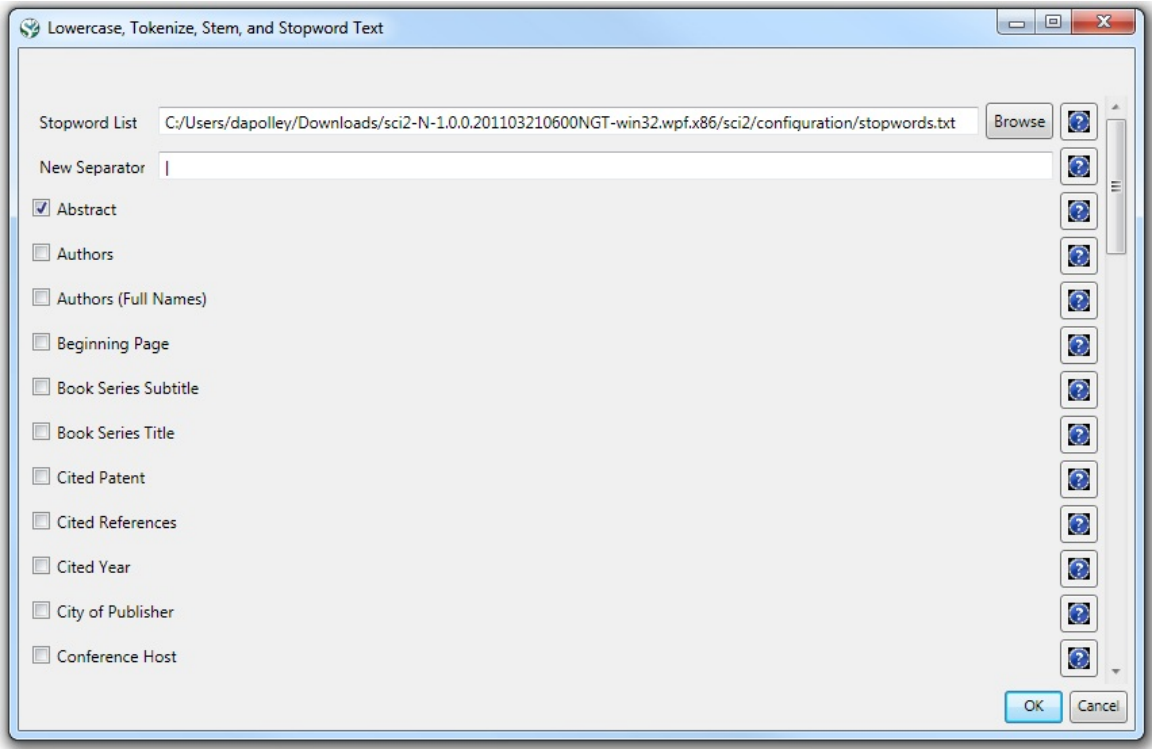
The giant component can be visualized in GUESS, see Figure 5.15 (right); see the above explanation, and use the same size and color coding and labeling as the bibliographic coupling network. Simply run *'GUESS: File > Run Script ...'* and select '***yoursci2directory****/scripts/GUESS/reference-co-occurence-nw.py*'

Figure 5.15: Undirected, weighted bibliographic coupling network (left) and undirected, weighted co-citation network (right) of 'FourNetSciResearchers' dataset, with isolate nodes removed

To see the log file from this workflow save the 5.1.4.4 Document Co-Citation Network (DCA) log file.

5.1.4.5 Word Co-Occurrence Network

In the Sci^2 Tool, select "361 unique ISI Records" from the *'FourNetSciResearchers'* dataset in the Data Manager. Run '*Preprocessing > Topical >* *Lowercase, Tokenize, Stem, and Stopword Text*' using the following parameters:

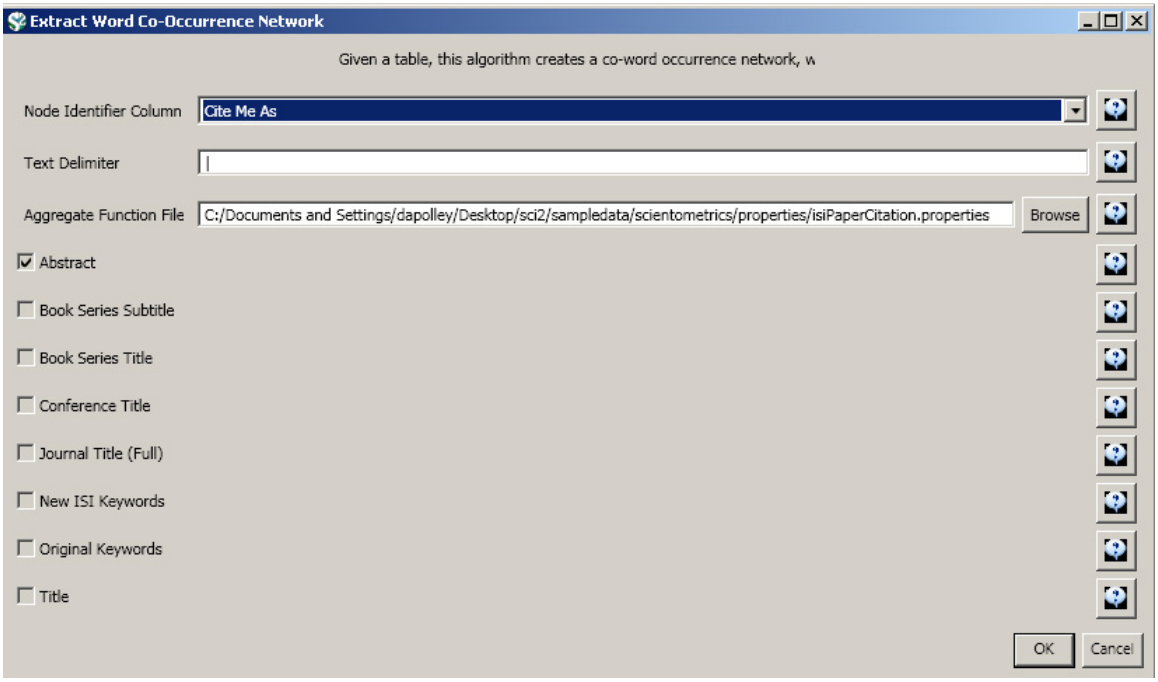


Text normalization utilizes the Standard Analyzer provided by Lucene (http://lucene.apache.org|). It separates text into word tokens, normalizes word tokens to lower case, removes "s" from the end of words, removes dots from acronyms, deletes stop words, and applies the English Snowball stemmer (http://snowball.tartarus.org/algorithms/english/stemmer.html), which is a version of the Porter2 stemmer designed for the English language..

The result is a derived table – "with normalized Abstract" – in which the text in the abstract column is normalized. Select this table and run '*Data Preparation* > *Extract Word Co-Occurrence Network*' using parameters:

Aggregate Function File

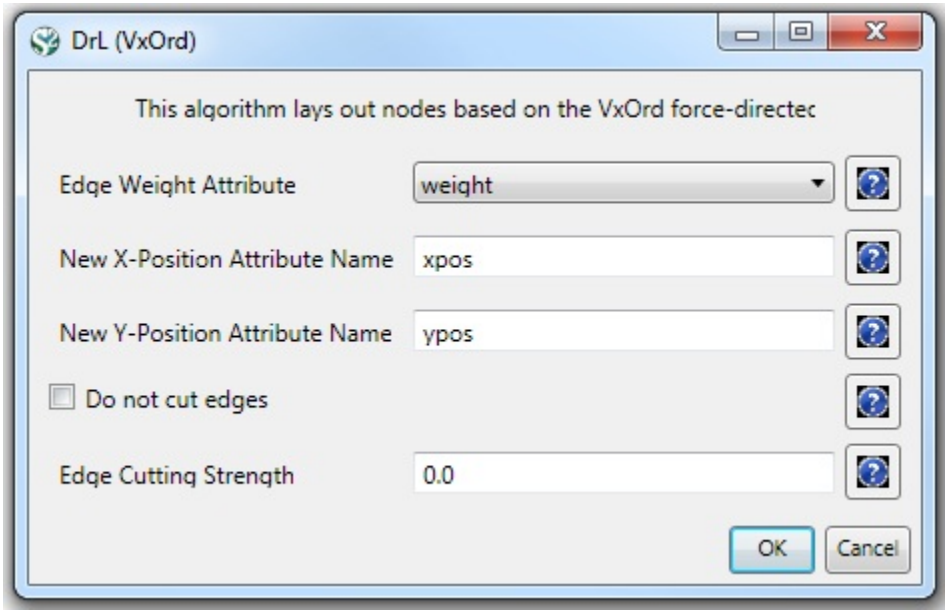
Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



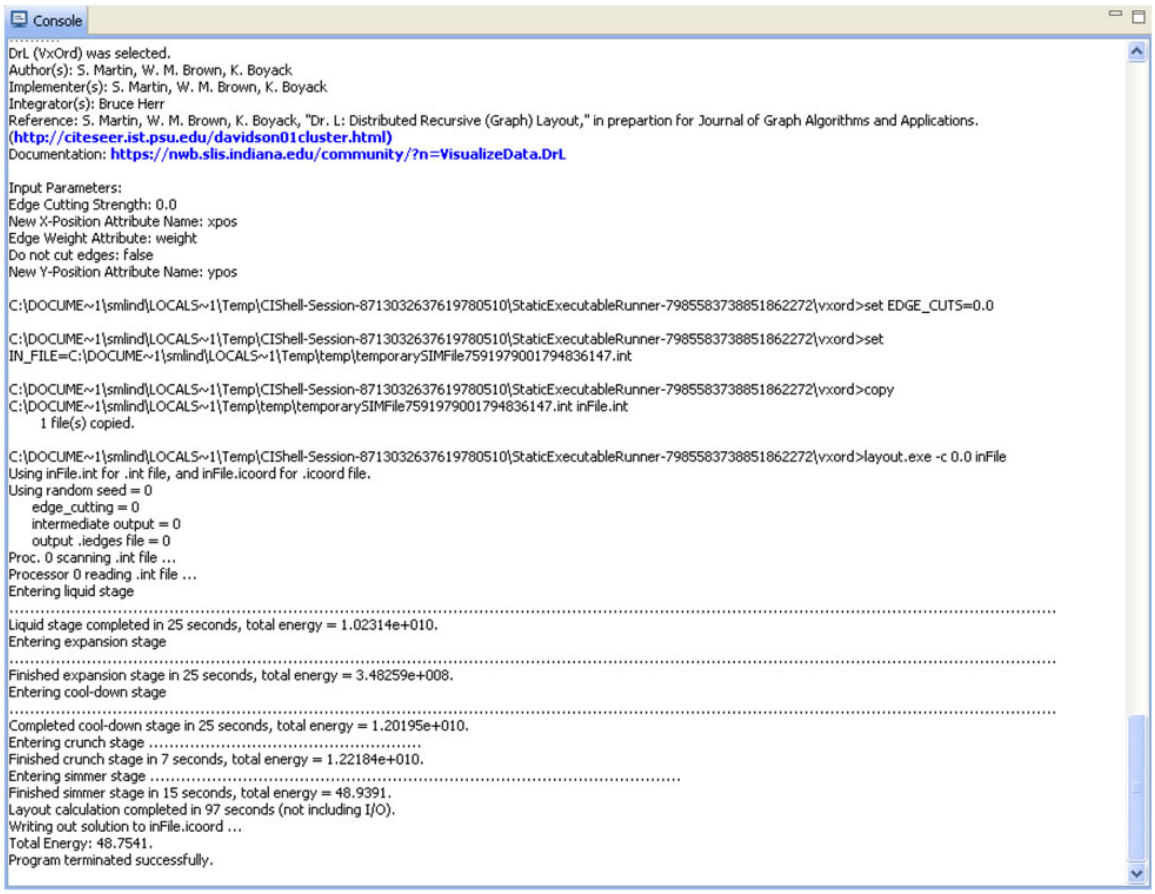
The outcome is a network in which nodes represent words and edges and denote their joint appearance in a paper. Word co-occurrence networks are rather large and dense. Running the *'Analysis > Networks > Network Analysis Toolkit (NAT)'* reveals that the network has 2,821 word nodes and 242,385 co-occurrence edges.

There are 354 isolated nodes that can be removed by running '*Preprocessing > Networks > Delete Isolates'* on the Co-Word Occurrence network. Note that when isolates are removed, papers without abstracts are removed along with the keywords.

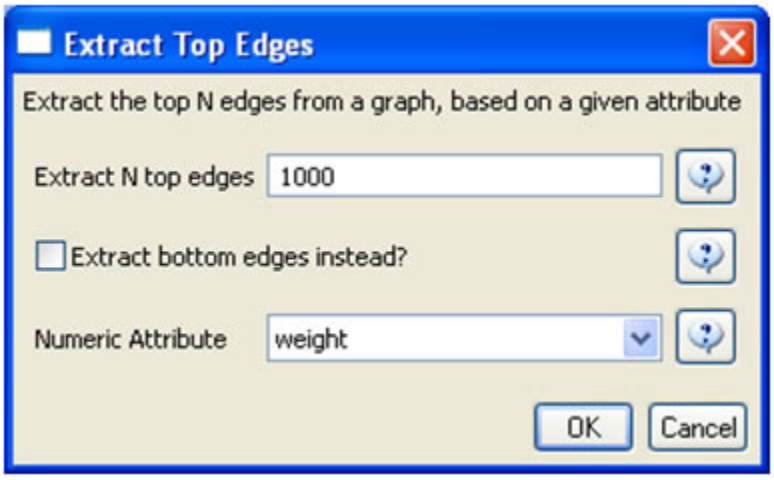
The result is one giant component with 2,467 nodes and 242,385 edges. To visualize this rather large network, begin by running *'Visualization > Networks > DrL (VxOrd)'* with default values:



Note that the DrL algorithm requires extensive data processing and requires a bit of time to run, even on powerful systems. See the console window for details:



To keep only the strongest edges in the "Laid out with DrL" network, run '*Preprocessing > Networks >* [*Extract Top Edges*]*'* on the new network using the following parameters:



Once edges have been removed, the network "top 1000 edges by weight" can be visualized by running *'Visualizatio n > Networks >* [*GUESS*]*'*. In GUESS, run the following commands in the Interpreter:

```
> for node in g.nodes:
      node.x = node.xpos * 40
      node.y = node.ypos * 40
> resizeLinear(references, 2, 40)
> colorize(references,[200,200,200],[0,0,0])
> resizeLinear(weight, .1, 2)
> g.edges.color = "127,193,65,255"
```

The result should look something like Figure 5.16.

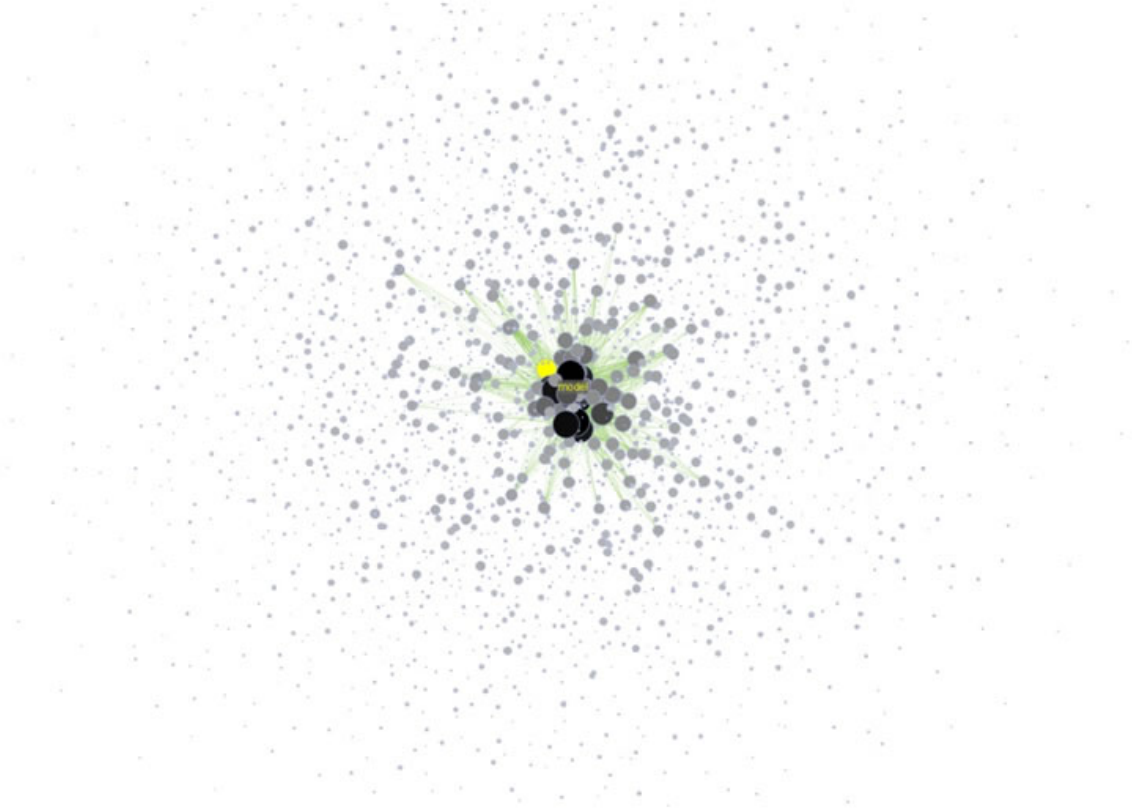


Figure 5.16: Undirected, weighted word co-occurrence network visualization for the DrL-processed 'FourNetSciResearchers' dataset

Currently, when you resize large networks in the GUESS visualization tool, the network visualizations can become "uncentered" in the display window. Running '*View* > *Center*' does not solve this problem. Users should zoom out to find the visualization, center it, and then zoom back in.

Note that only the top 1000 edges (by weight) in this large network appear in the above visualization, creating the impression of isolate nodes. To remove nodes that are not connected by the top 1000 edges (by weight), run '*Prepr ocessing* > *Networks* > *Delete Isolates*' on the "top 1000 edges by weight" network and visualize the result using the workflow described above.

To see the log file from this workflow save the 5.1.4.5 Word Co-Occurrence Network log file.

Database Extractions

Extended Version

This workflow uses the extended version of the Sci^2 Tool. To know how to extend Sci^2 view Section 3.2 Additional Plugins.

The Sci^2 Tool supports the creation of databases from ISI files. Database loading improves the speed and functionality of data preparation and preprocessing. While the initial loading can take quite some time for larger datasets (see sections 3.4 Memory Allocation and 3.5 Memory Limits) it results in vastly faster and more powerful data processing and extraction.

Once again load *'**yoursci2directory**/sampledata/scientometrics/isi/FourNetSciResearchers.isi'*, this time using '*File > Load*' instead of '*File > Load.*' Choose 'ISI database' from the load window. Right-click to view the database schema:

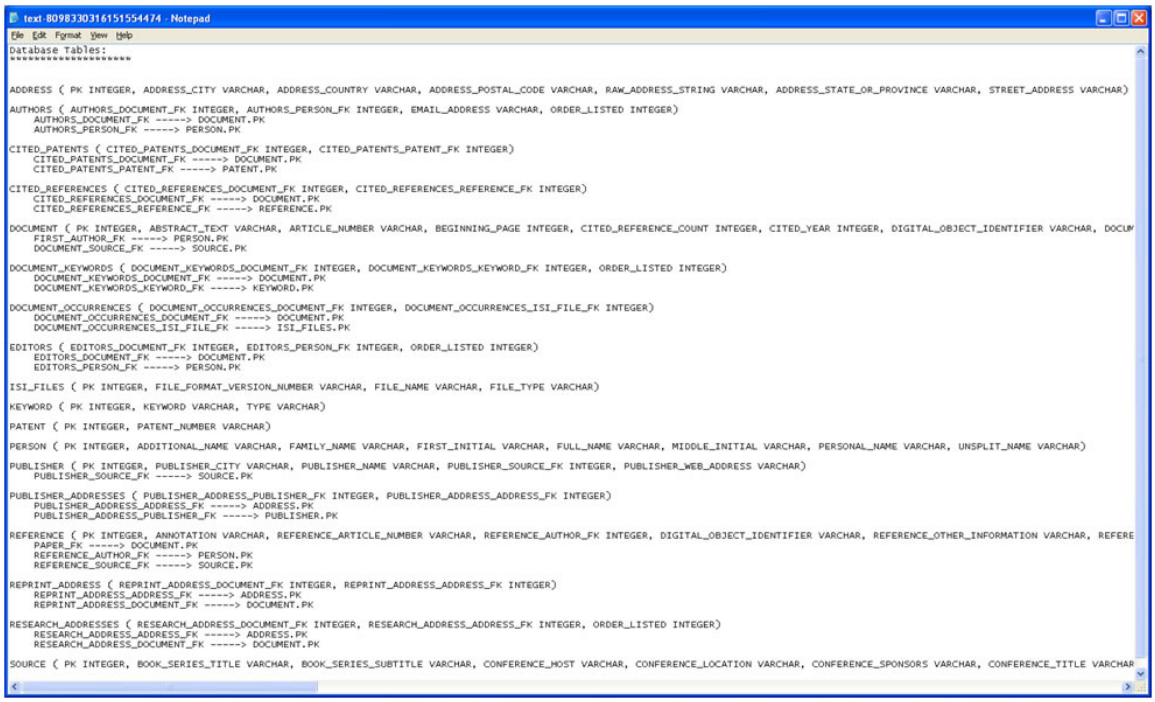


Figure 5.17: The database schema as viewed in Notepad

As before, it is important to clean the database before running any extractions by merging and matching authors, journals, and references. Run '*Data Preparation > Database > ISI > Merge Identical ISI People*', followed by '*Data Preparation > Database > ISI > Merge Document Sources''* and '*Data Preparation > Database > ISI > Match References to Papers*'. Make sure to wait until each cleaning step is complete before beginning the next one.

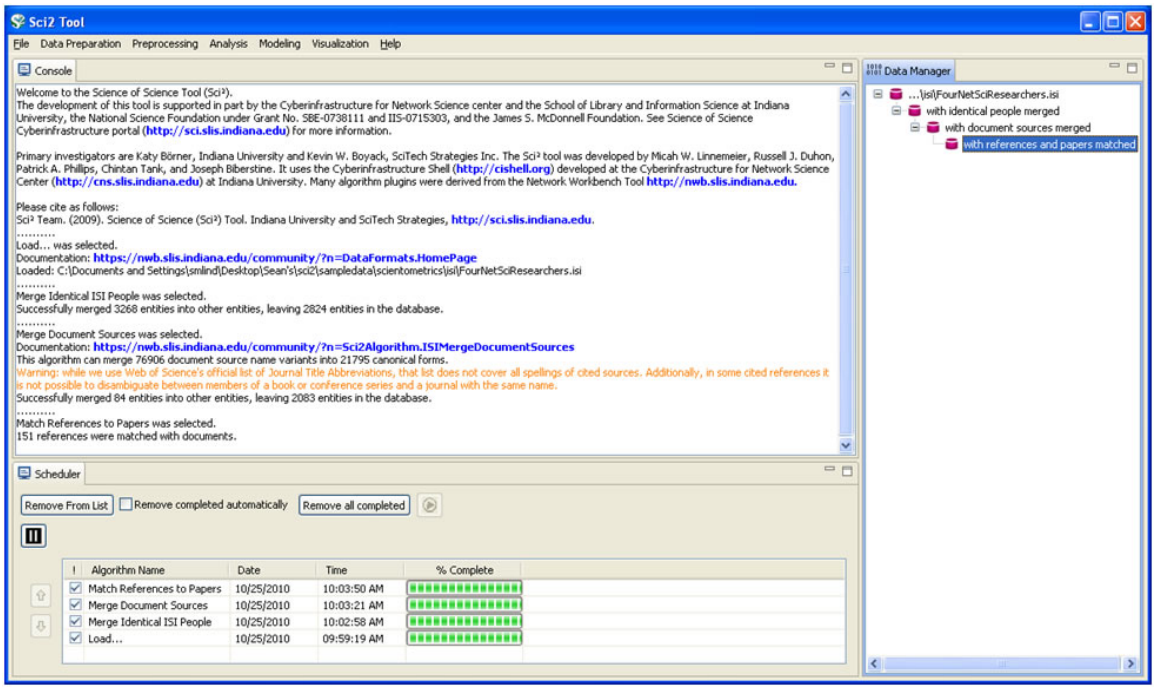
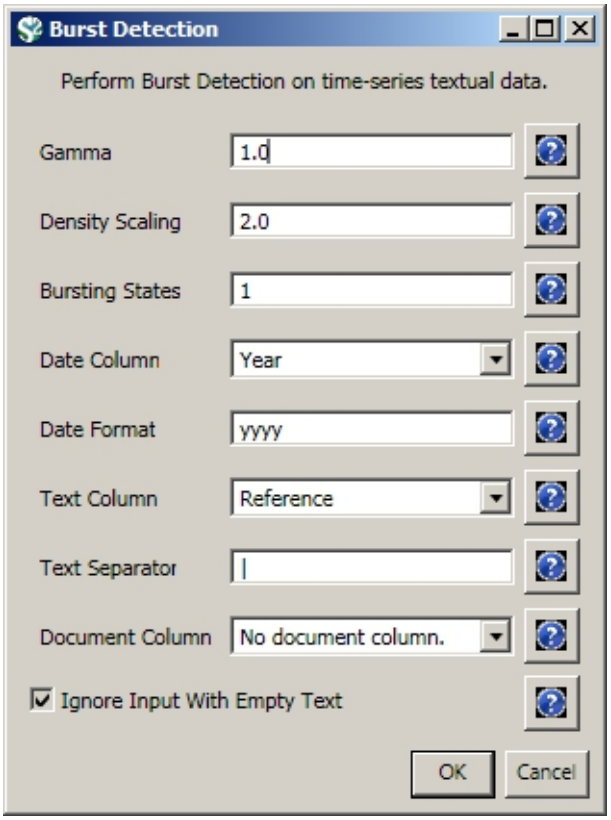


Figure 5.18: Cleaned database of 'FourNetSciResearchers'

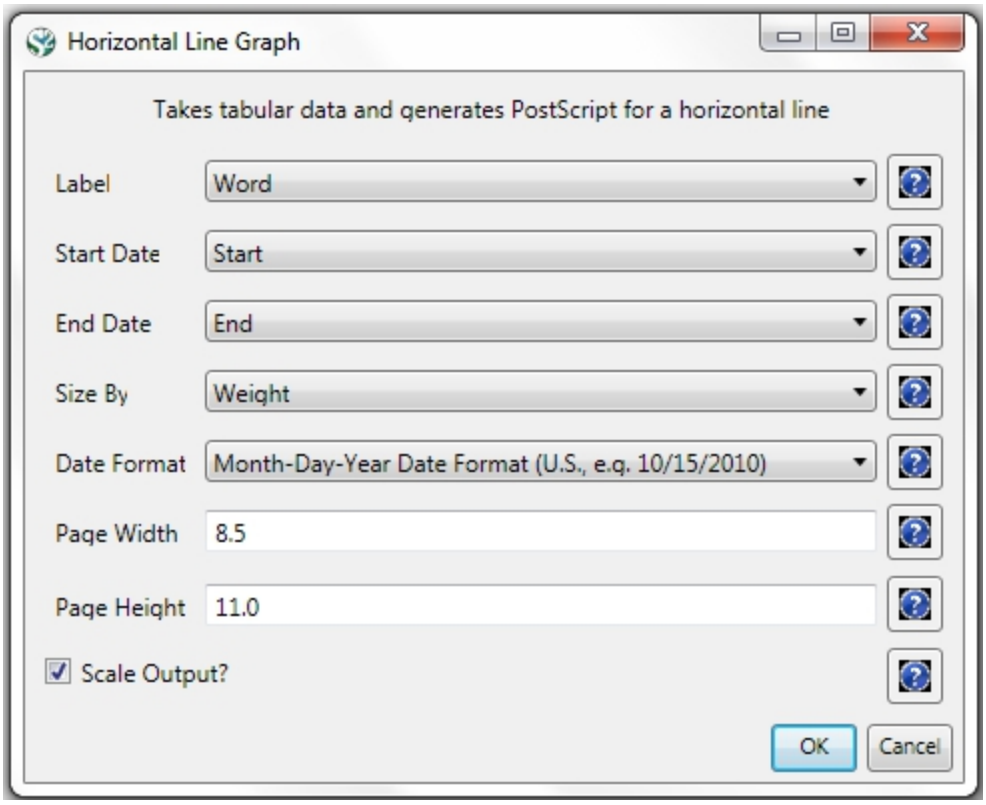
Extracting different tables will provide different views of the data. Run '*Data Preparation* > *Database* > *ISI* > *Extract Authors*' to view all the authors from FourNetSciResearchers.isi. The table includes the number of papers each person in the dataset authored, their Global Citation Count (how many times they have been cited according to ISI), and their Local Citation Count (how many times they were cited in the current dataset.)
The queries can also output data specifically tailored for the burst detection algorithm (see section 4.6.1 Burst Detection).
Run '*Data Preparation* > *Database* > *ISI* > *Extract References by Year for Burst Detection*' on the cleaned "with references and papers matched" database, followed by '*Analysis* > *Topical* > Burst Detection' with the following parameters:



View the file "Burst detection analysis (Publication Year, Reference): maximum burst level 1". On a PC running Windows, right click on this table and select view to see the data in Excel. On a Mac or a Linux system, right click and save the file, then open using the spreadsheet program of your choice. See Burst Detection for the meaning of each field in the output.

A empty value in the "End" field indicates that the burst lasted until the last date present in the dataset. Where the "End" field is empty, put manually the last year present in the dataset. In this case, 2007.

After you add manually this information, save this .csv file somewhere in your computer. Load back this .csv file into Sci2 using '*File* > *Load*'. Select 'Standart csv format' int the pop-up window. A new table will appear in the Data Manager. To visualize these table that contains the results of the Burst Detection algorithm, select the table you just loaded in the Data Manager and run '*Visualization* > *Temporal* > Horizontal Bar Graph' with the following parameters:



See section [2.4 Saving Visualizations for Publication] to save and view the graph.

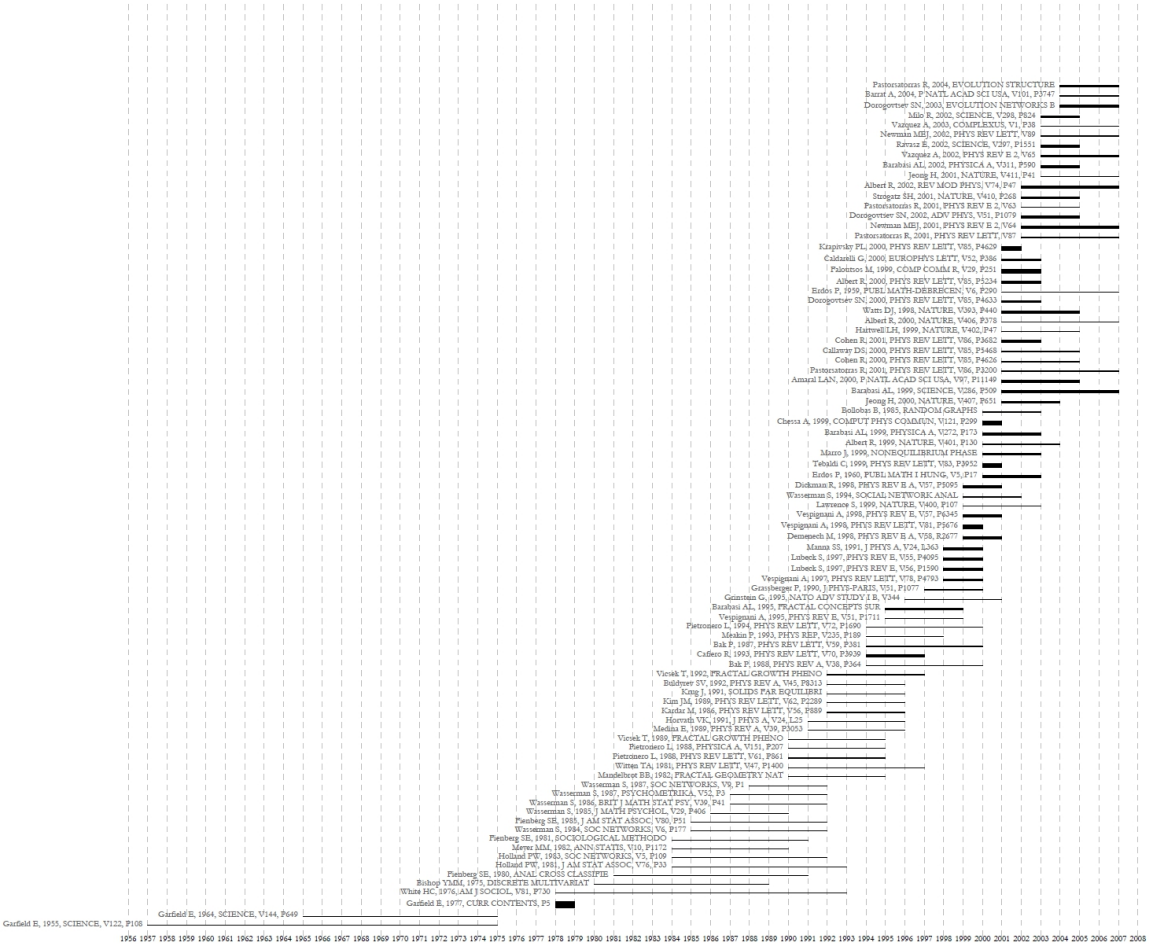


Figure 5.19: Top reference bursts in the 'FourNetSciResearchers' dataset

For temporal studies, it can be useful to aggregate data by year rather than by author, reference, etc. Running '*Data Preparation > Database > ISI > Extract Longitudinal Summary*' will output a table which lists metrics for every year mentioned in the dataset. The longitudinal study table contains the volume of documents and references published per year, as well as the total number of references, the number of distinct references, distinct authors, distinct sources, and distinct keywords per year. The results are graphed in Figure 5.20 (the graph was created using Excel).

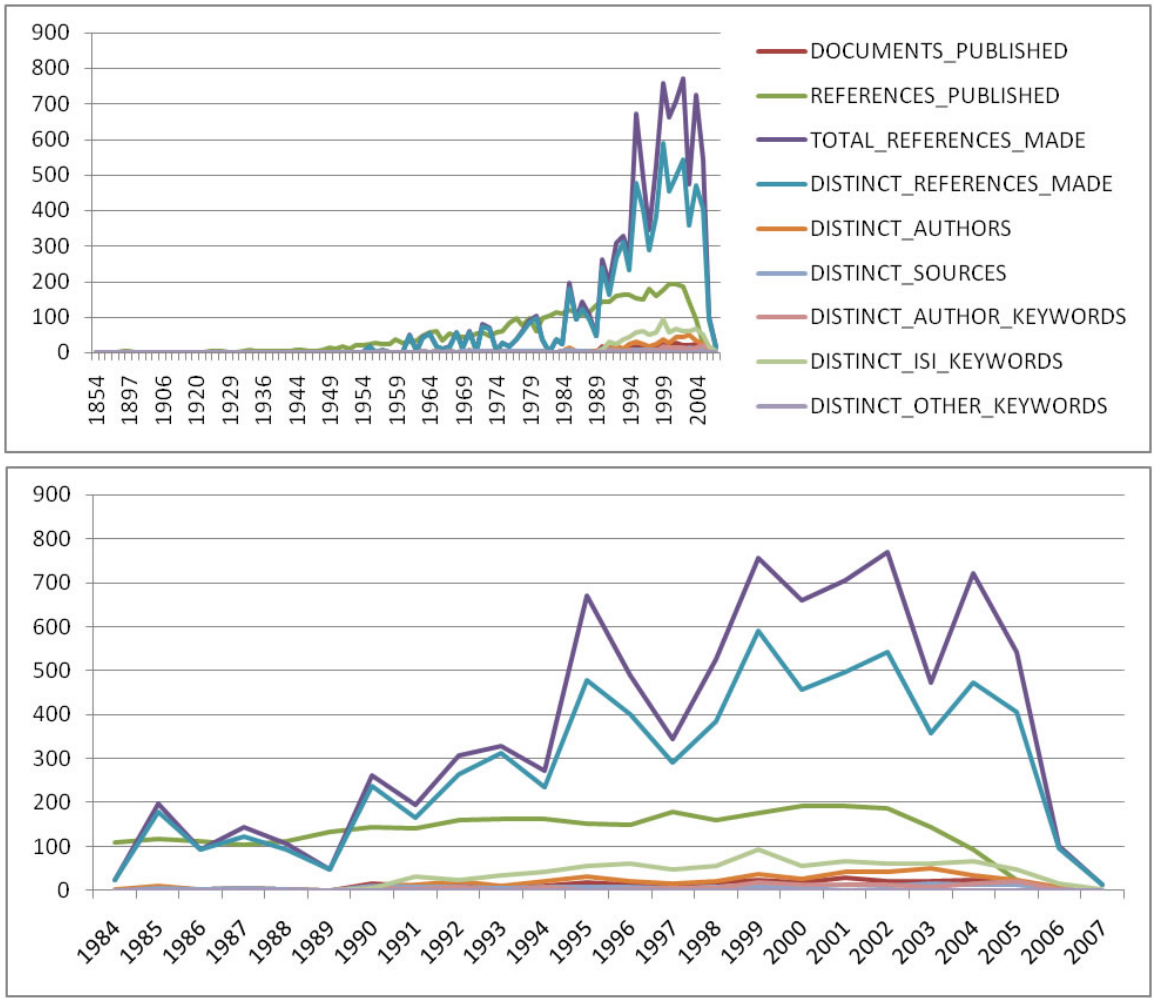


Figure 5.20: Longitudinal study of 'FourNetSciResearchers'

The largest speed increases from the database functionality can be found in the extraction of networks. First, compare the results of a co-authorship extraction with those from section 5.1.4.2 Author Co-Occurrence (Co-Author) Network. Run '*Data Preparation > Database > ISI > Extract Co-Author Network*' followed by '*Analysis > Networks > Network Analysis Toolkit (NAT)*'. Notice that both networks have 247 nodes and 891 edges. Visualize the extracted co-author network in GUESS using '*Visualization > Networks > GUESS*' and reformat the visualization using '*Layout*

> *GEM'* and *'Layout > Bin Pack*.' To apply the default co-authorship theme, go to 'Script > Run Script' and find *'**your sci2directory**/scripts/GUESS/co-author-nw_database.py'*. The resulting network will look like Figure 5.21.

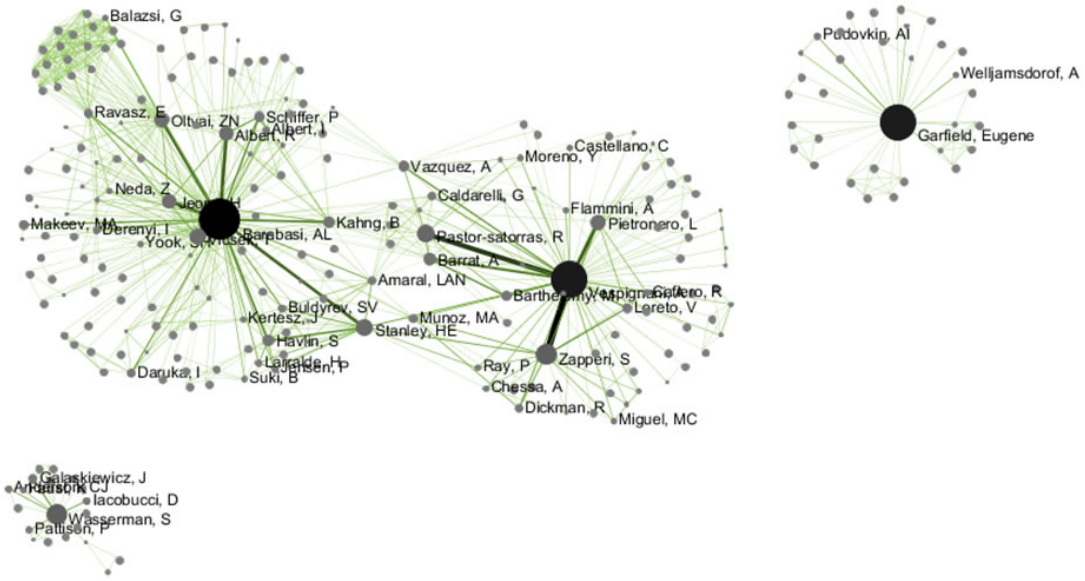
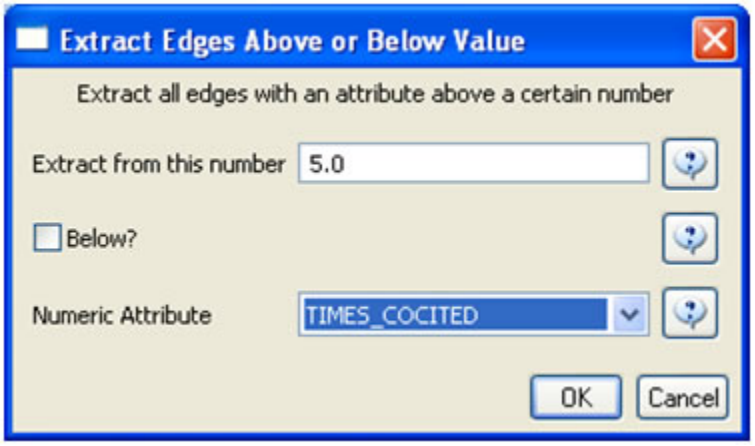


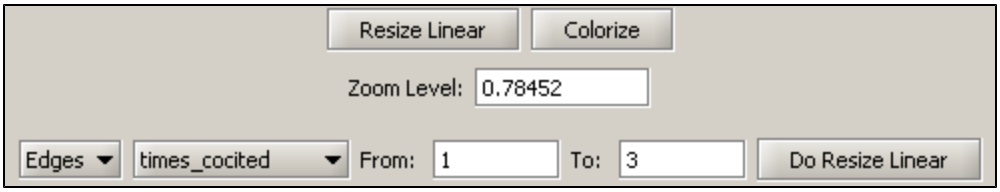
Figure 5.21: Longitudinal study of 'FourNetSciResearchers,' visualized in GUESS

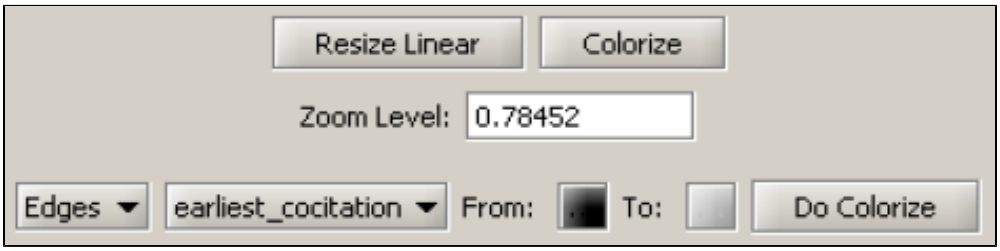
Using Sci^2's database functionality allows for several network extractions that cannot be achieved with the text-based algorithms. For example, extracting journal co-citation networks reveals which journals are cited together most frequently. Run '*Data Preparation > Database > ISI >* *Extract Document Co-Citation Network (Core and References)*' on the database to create a network of co-cited journals, and then prune it using '*Preprocessing > Networks >* *Extract Edges Above or Below Value*' with the parameters:



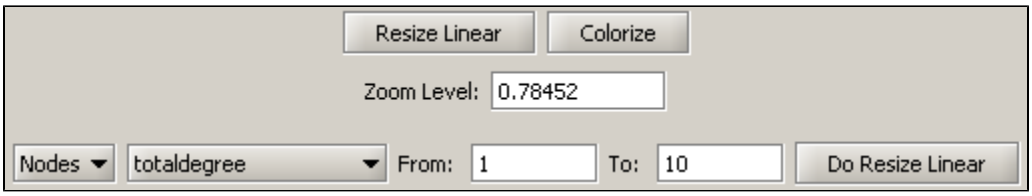
Now remove isolates ('*Preprocessing > Networks >* *Delete Isolates*') and append node degree attributes to the network ('*Analysis > Networks > Unweighted & Undirected >* *Node Degree*'). The workflow in your Data Manager should look like this:

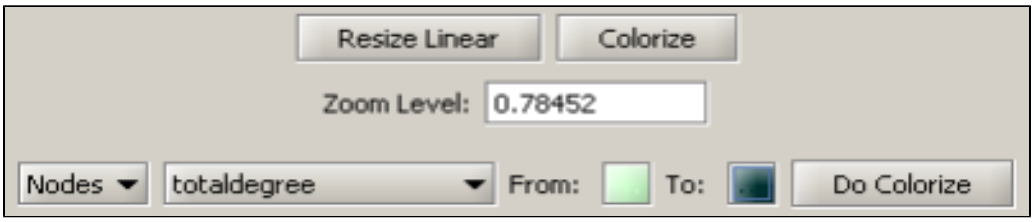
View the network in GUESS using '*Visualization > Networks >* *GUESS*.' Use '*Layout > GEM'* and *'Layout > Bin Pack'* to reformat the visualization. Resize and color the edges to display the strongest and earliest co-citation links using the following parameters:





Resize, color, and label the nodes to display their degree using the following parameters:





The resulting Journal Co-Citation Analysis (JCA) appears in Figure 5.22.

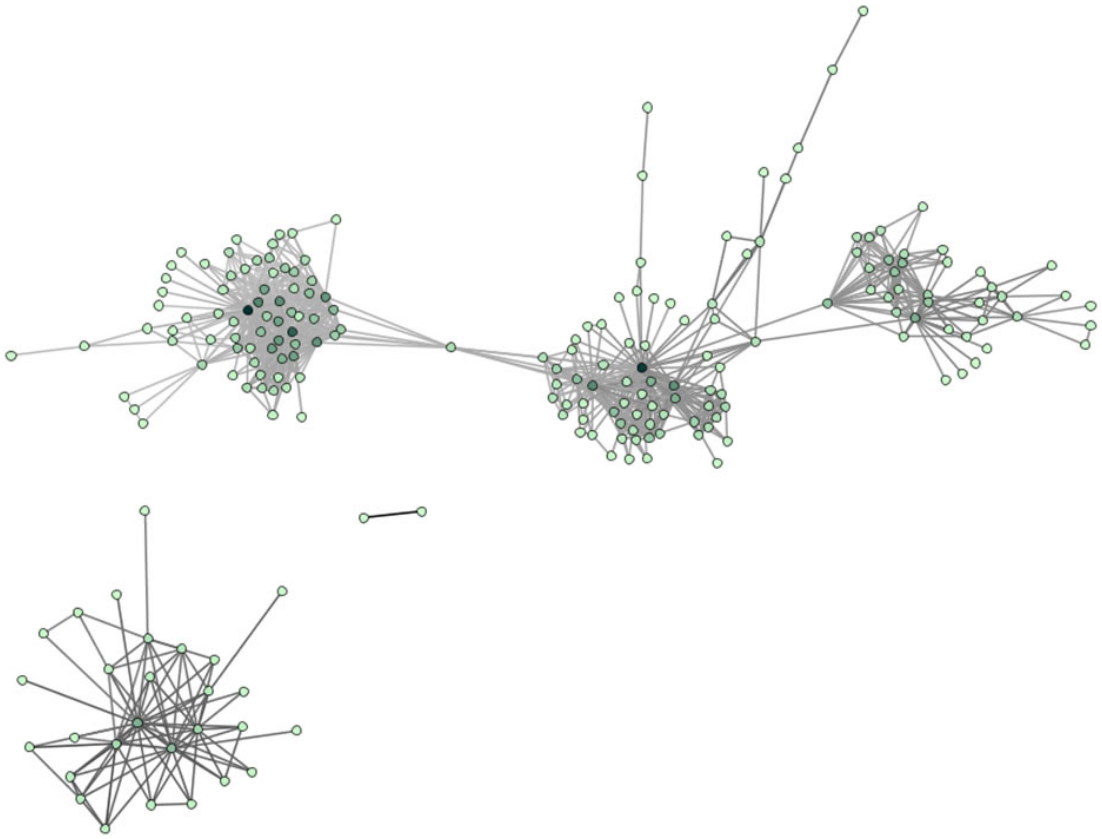


Figure 5.22: Journal co-citation analysis of 'FourNetSciResearchers,' edges with 5 or more cocitations and nodes with node degree added, visualized in GUESS

5.2 Institution Level Studies - Meso

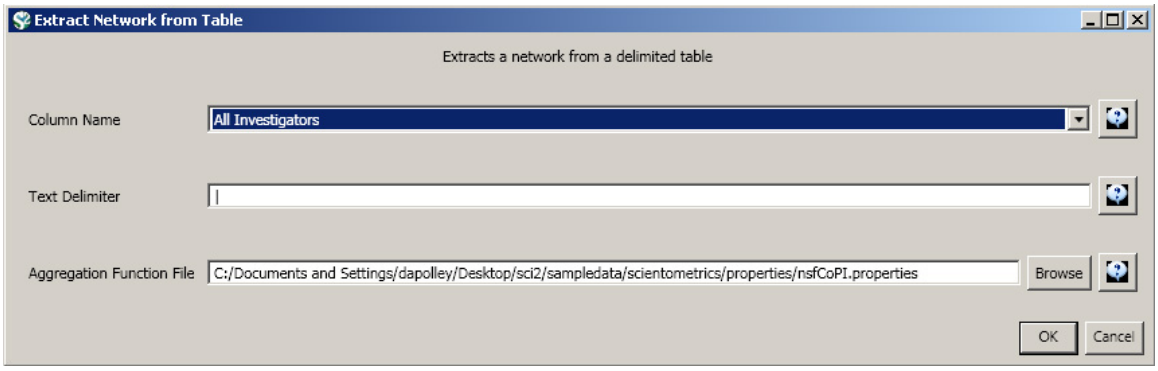
5.2.1 Funding Profiles of Three Universities (NSF Data)

Cornell.nsfIndiana.nsfMichigan.nsf	
Time frame:	2000-2009

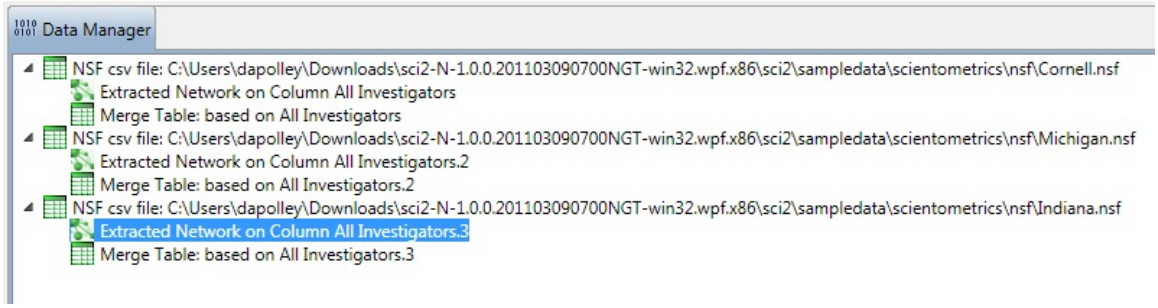
Region(s):	Cornell University, Indiana University, Michigan University
Topical Area(s):	Miscellaneous
Analysis Type(s):	Co-PI Network

Load '*Cornell.nsf*', '*Michigan.nsf*', and '*Indiana.nsf*' using '*File > Load*'' and following this path: '***yoursci2directory****/sampledata/scientometrics/nsf*' (if these files are not in the sample data directory they can be downloaded from 2.5 Sample Datasets). Use the following workflow for each of the three nsf files loaded. Select each of the datasets in the Data Manager window and run *'Data Preparation > Extract Co-Occurrence Network'* using the following parameters (Note that the Aggregation Function File is '***yoursci2directory****/sampledata/scientometrics/properties/nsfCoPI.properties'*):

> **Aggregate Function File**
>
> Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



Two derived files per dataset will appear in the Data Manager window: the co-PI network and a merge table.



In the network, nodes represent investigators and edges denote their co-PI relationships. (To learn how the merge table can be used to further clean PI names, see section 5.1.4.2 Author Co-Occurrence (Bibliographic Coupling) Network.)

Choose the "Extracted Network on Column All Investigators" and run *'Analysis > Networks > Network Analysis Toolkit (NAT)'* for each dataset. This will display the amount of nodes and edges, as well as the amount of isolate nodes that can be removed by running *'Preprocessing > Networks > Delete Isolates'.*

Select *'Visualization > Networks > GUESS'* and run '*Layout > GEM'* followed by '*Layout > Bin Pack'* to visualize the network. Run the '***yoursci2directory****/scripts/GUESS/co-PI-nw.py'* script. Visualizations of the three university's co-PI networks are shown in Figure 5.23.

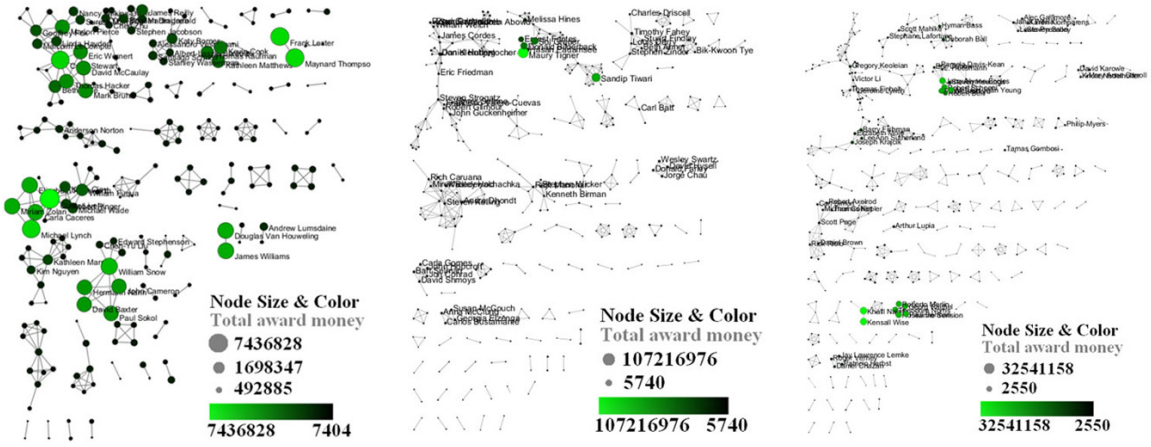
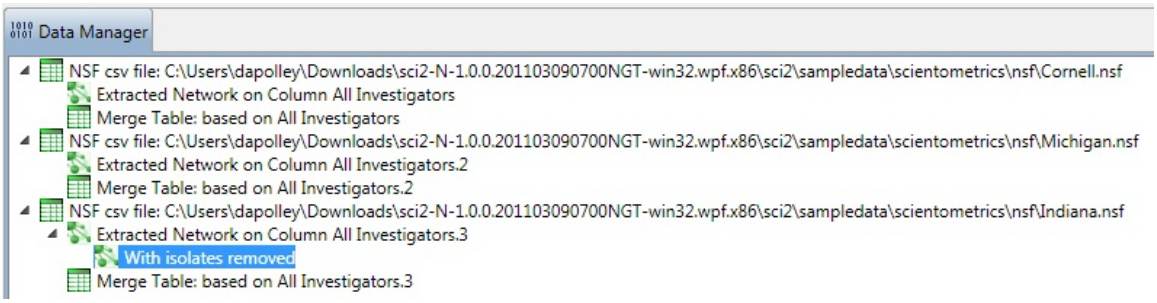
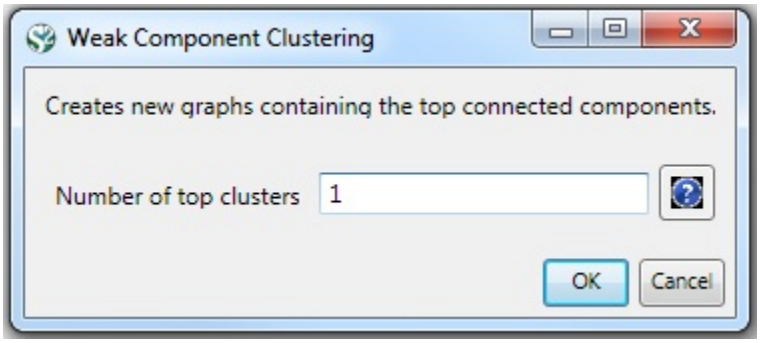


Figure 5.23: Co-PI network of Indiana University (top, left), Cornell University (top, right), University of Michigan (middle).

To see a more detailed view of any of the components in the network (e.g. the largest Indiana component) select the network with deleted isolates in the Data Manager:



Then, run *'Analysis > Networks > Unweighted & Undirected >* Weak Component Clustering' with the parameter:



Indiana's largest component has 19 nodes, Cornell's has 67 nodes, and Michigan's has 55 nodes. Visualize Indiana's network in GUESS using the '***yoursci2directory****/scripts/GUESS/co-PI-nw.py'* script. Save the file as a jpg by selecting *'File > Export Image'*. Use the "Browse..." option in the "Export Image – GUESS" popup window to select the folder in which you would like to save the image.

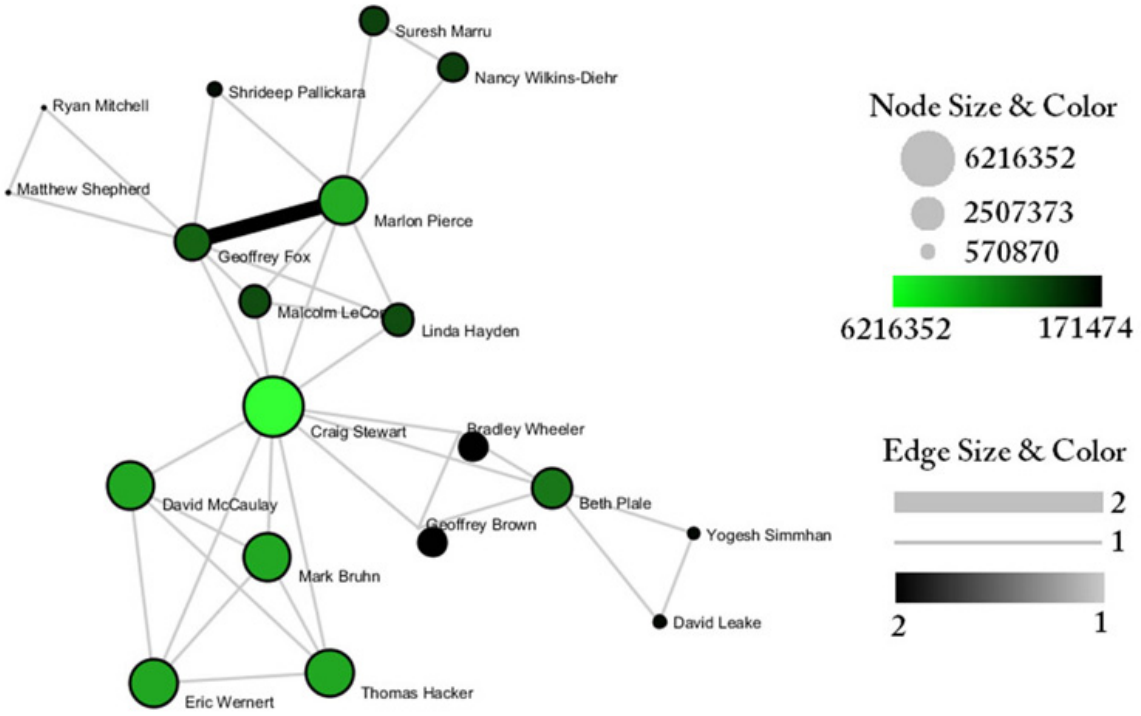


Figure 5.24: Largest component of Indiana University co-PI network. Node size and color display the total a ward amount.

To see the log file from this workflow save the 5.2.1 Funding Profiles of Three Universities (NSF Data) file.

5.2.1.1 Database Extractions

> **Extended Version**
>
> This workflow uses the extended version of the Sci^2 Tool. To know how to extend Sci^2 view Section 3.2 Additional Plugins.

The Sci^2 Tool supports the creation of databases for NSF files. Database loading improves the speed and functionality of data preparation and preprocessing. To load the Indiana NSF file as a database, go to *'File > Load >* and select ***yoursci2directory****/sampledata/scientometrics/nsf/Indiana.nsf.'* In the "Load" pop-up window, choose "NSF database." Cleaning should be performed before any other task using '*Data Preparation > Database > NSF >* Merge Identical NSF People'.
To view a breakdown of each investigator from Indiana, run '*Data Preparation > Database > NSF >* Extract Investigators'. View the table – notice that next to each investigator will be listed their total number of awards, total as the PI and as a Co-PI, the total amount awarded to date, as well as their earliest award start date and latest award expiration date.

To create Co-PI networks like those from the previous workflows, simply run '*Data Preparation > Databases > NSF >* Extract Co-PI Network' on the cleaned database. Delete the isolates by running *'Preprocessing > Networks >* *Delete Isolates'*.
As before, to visualize the network, select *'Visualization > Networks > GUESS'* and run '*Layout > GEM'* followed by '*Layout > Bin Pack'*. Run the '***yoursci2directory****/scripts/GUESS/co-PI-nw_database.py'* script to apply the standard Co-PI network theme.

Figure 5.25: Indiana University Co-PI network using databases.

5.2.2 Mapping CTSA Centers (NIH RePORTER Data)

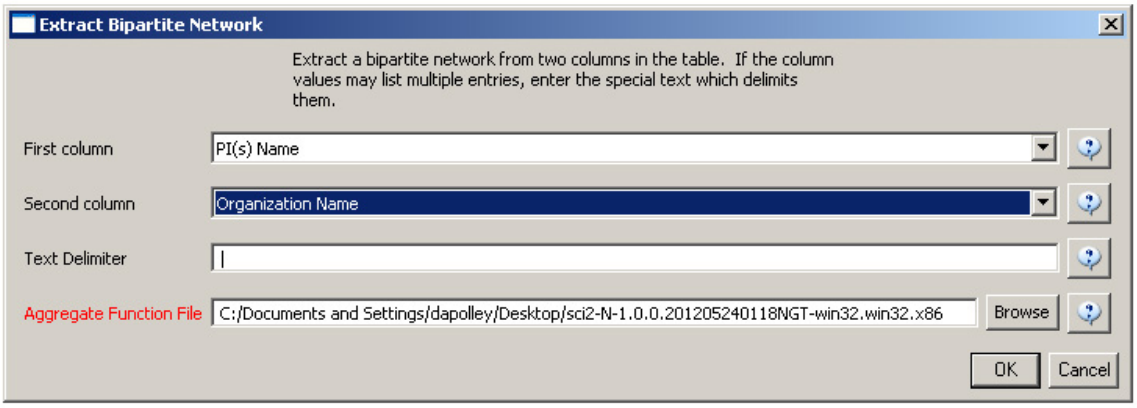
CTSA2005-2009.xls	
Time frame:	2005-2009
Region(s):	Miscellaneous
Topical Area(s):	Clinical and Translational Science
Analysis Type(s):	PI-Institution Network, Co-Authorship Network

Data doesn't always come in easily parsible formats. This study, a search for all recent grants to CTSA centers, requires some advanced searching and data manipulation to prepare the required data. The data comes from the union of NIH RePORTER downloads (see section 4.2.2.2 NIH RePORTER) and NIH ExPORTER data dumps (http://projectreporter.nih.gov/exporter/). Each CTSA Center grant was found and matched with its associated publications using a project-specific ID.

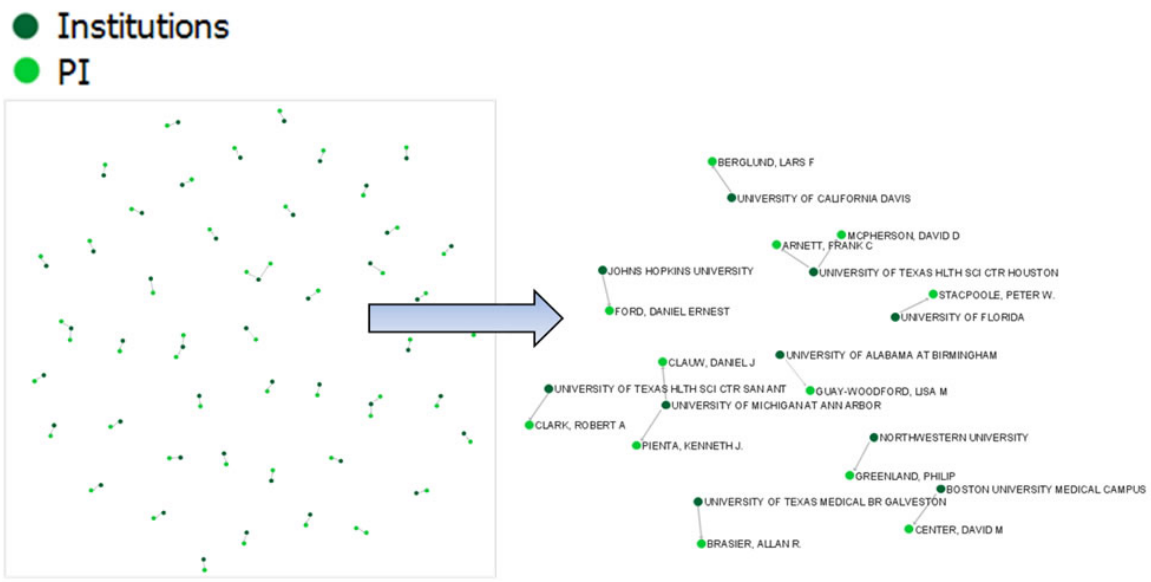
The resulting file, which contains all NIH Clinical and Translational Science Awards and their corresponding details from 2005-2009, is saved in an Excel file in '***yoursci2directory****/sampledata/scientometrics/nih/CTSA2005-2009*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). The file contains two spreadsheets, one with publications and one with grants. Save each spreadsheet out as *grants.csv* and *publications*

.csv.

First load *grants.csv* in the Sci^2 Tool using '*File* > *Load*' and *'Standard csv format'* in the "Load" pop-up. To view a bimodal network visualizing which main PIs associate with which institution, run '*Data Preparation* > *Extract Bipartite Network* *'* with the following parameters:

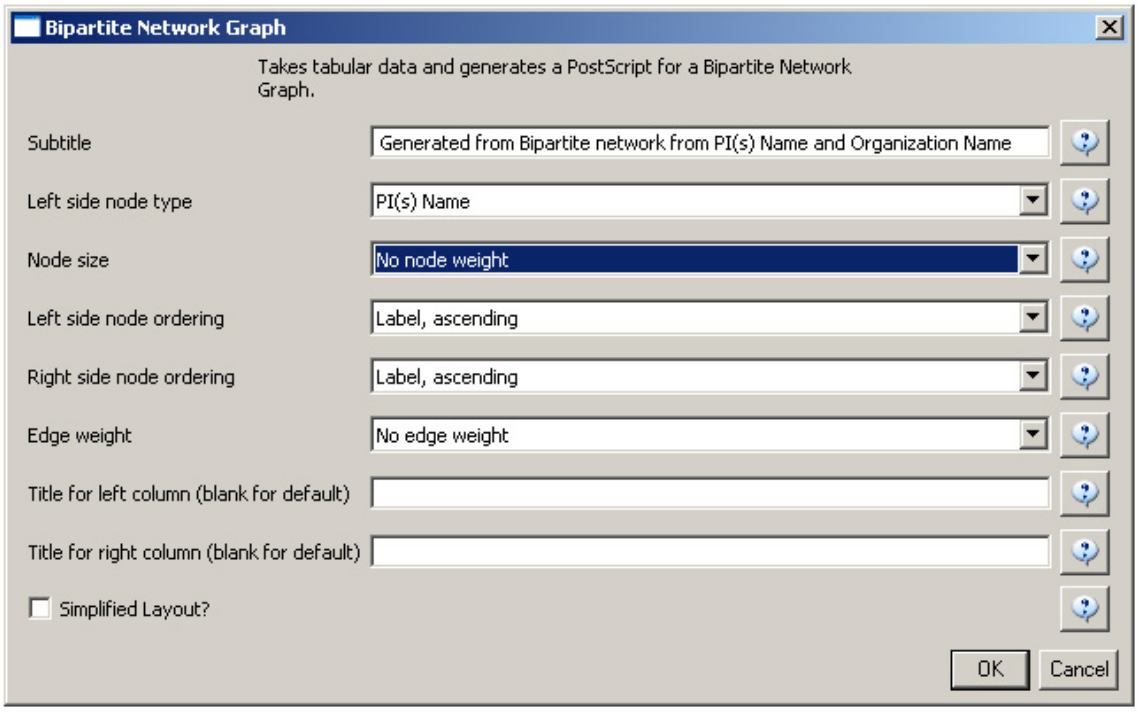


The resulting network can be visualized in GUESS and laid out using GEM, see Figure 5.26



.

The network can also be visualized with the Bipartite-specific visualization. Run '*Visualization* > *Networks* > *Bipartite Network Graph'* with the following parameters:



Sci2 with generate a visualization titled "Bipartite Network Graph PS" in the Data Manager. This network visualization can be saved as a PostScript file and the resulting visualization will look like this:

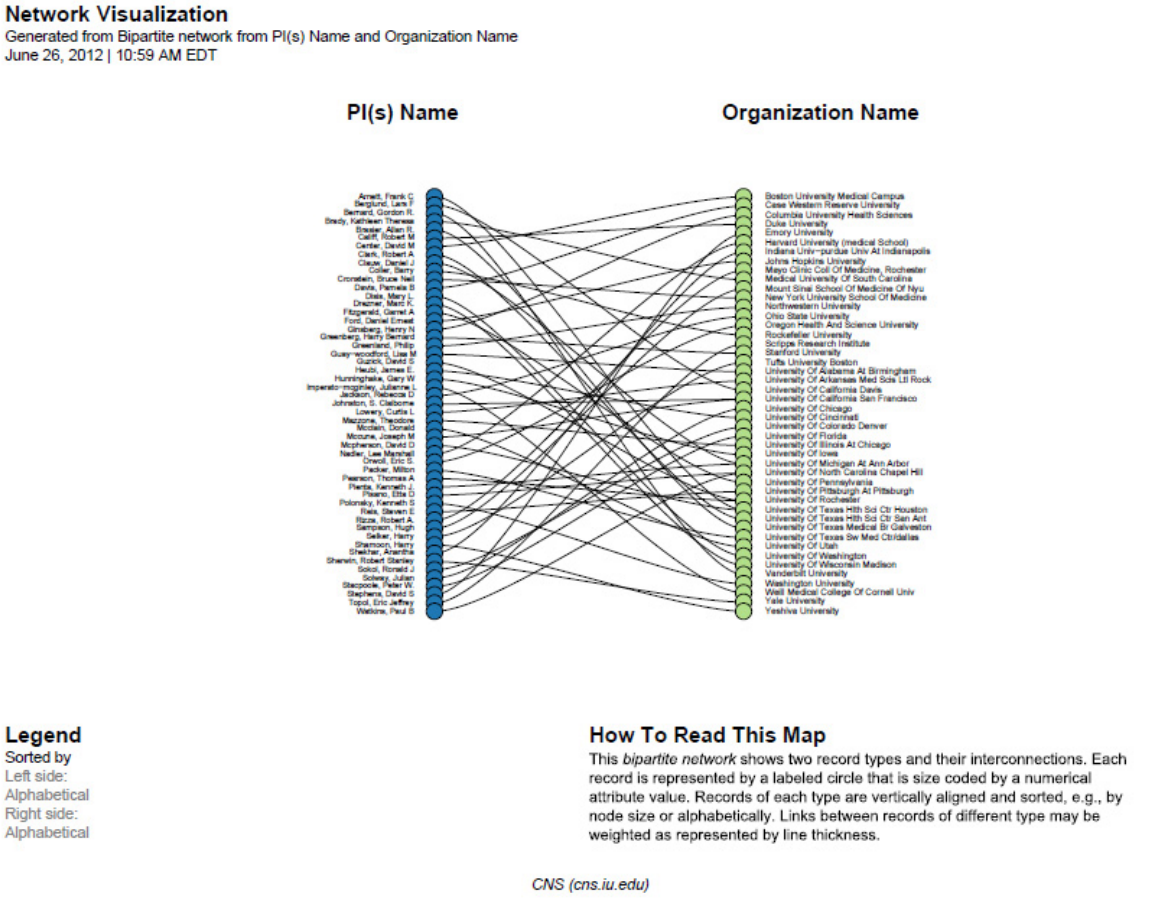


Figure 5.26: Bimodal institution-PI network for CTSA Centers.

Now load *publications.csv* as a standard csv and create a co-authorship network by running '*Data Preparation* > *Extract Co-Occurrence Network*' with text delimiter set to " **;** ". The parameter for Column Name should be set to "author." The resulting co-authorship network has 8,668 nodes, 26 isolates, and 50,129 edges (see Figures 5.27 and 5.28).

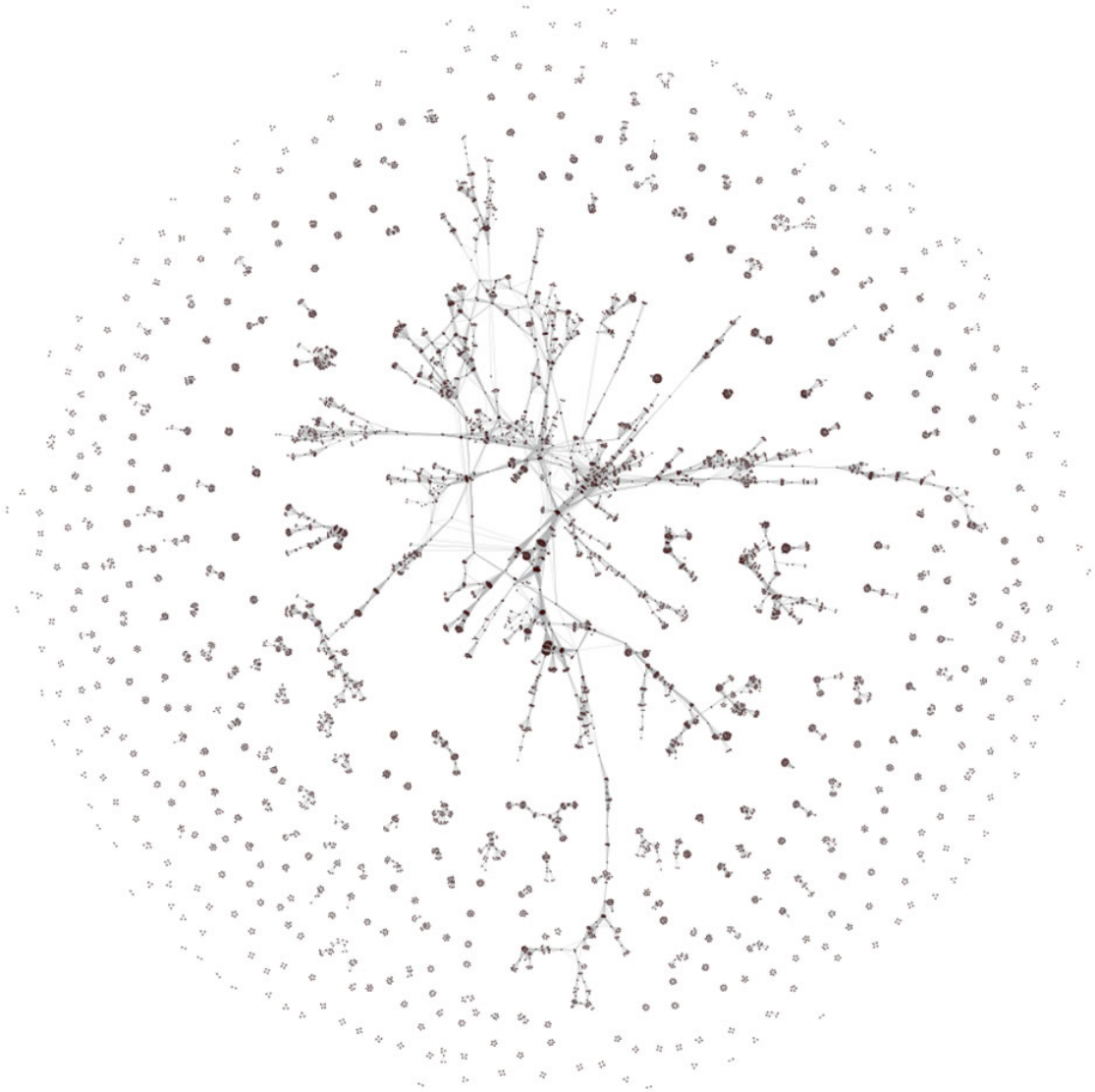


Figure 5.27: Co-authorship network of CTSA Center publications

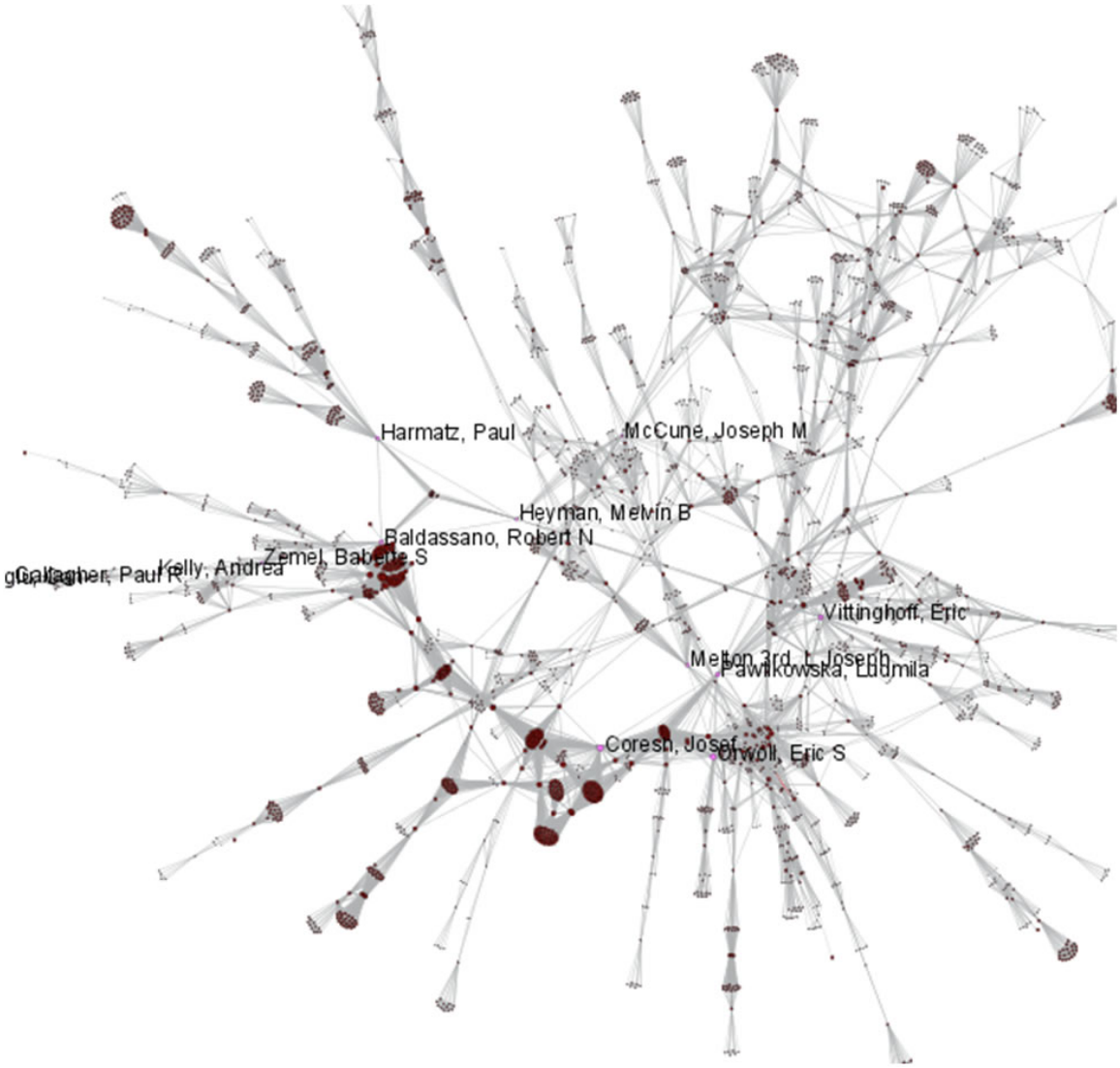


Figure 5.28: Largest connected component of CTSA Center publication co-authorship network

To see the log file from this workflow save the 5.2.2 Mapping CTSA Centers (NIH RePORTER Data) log file.

5.2.3 Biomedical Funding Profile of NSF (NSF Data)

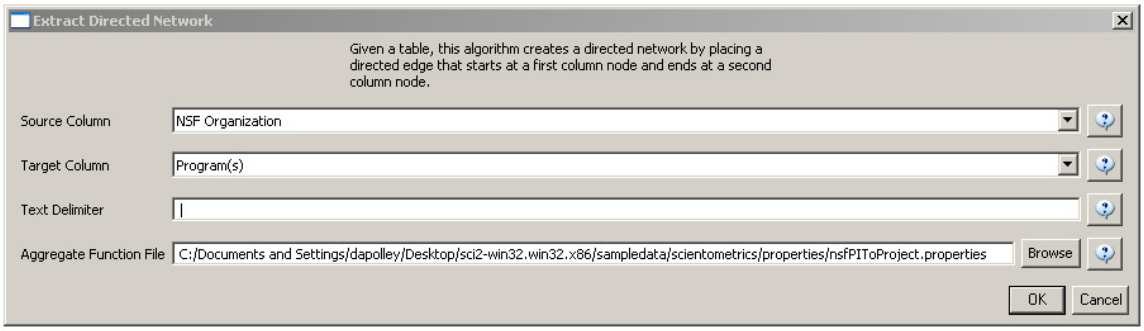
MedicalAndHealth.nsf	
Time frame:	2003-2010
Region(s):	Miscellaneous
Topical Area(s):	Biomedical
Analysis Type(s):	NSF Organization-Program Network

A NSF Organization-to-Program(s) bimodal network is exhibited here using data downloaded from the NSF Awards Search at (http://www.nsf.gov/awardsearch) on Nov 23th, 2009, using the query "medical AND health" in the title,

abstract, and awards field, with "Active awards only" checked. This query yielded 286 awards (see section 4.2.2.1 NSF Award Search for data retrieval details).

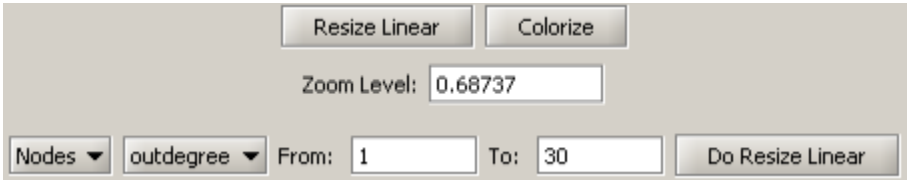
Load '***yoursci2directory***/*sampledata/scientometrics/nsf/MedicalAndHealth.nsf*' in NSF csv format (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Then run *'Data Preparation > Extract Directed Network'* with parameters:

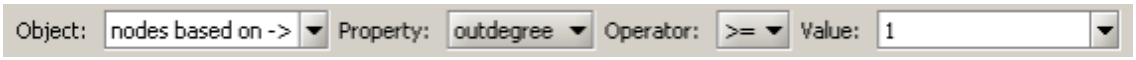
Make sure to select the nsfPIToProject.properties for the aggregate function file.



Select "Network with directed edges from NSF Organization to Program(s)" in the Data Manager and run *'Analysis > Networks > Unweighted and Directed > Node Indgree*. Then select "Network with indegree attribute added to node list" in the Data Manager and run '*Analysis > Networks > Unweighted and Directed > Node Outdegree'*. Select the resulting "Network with outdegree attribute added to node list" and run *'Visualization > Networks > GUESS'* followed by *' Layout > GEM'* and *'Layout > Bin Pack'*.

In graph modifier interface:





Select "Colour" from the options directly below and choose desired color.

Then select "Show Label".

Select "Colour" and choose desired color.

The resulting network is visualized in Figure 5.29.

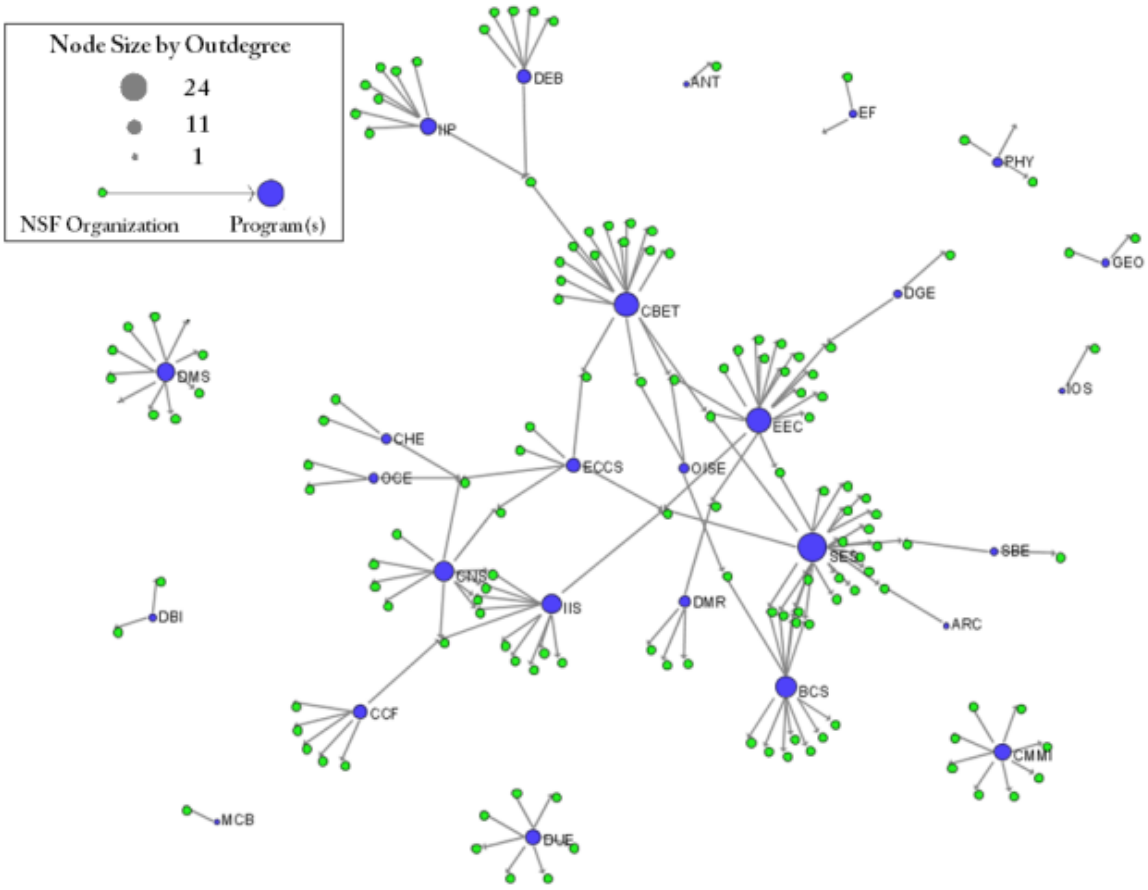


Figure 5.29: Bimodal Network of NSF Organization to Program(s) of NSF Medical + Health Funding

To see the log file from this workflow save the 5.2.3 Biomedical Funding Profile of NSF (NSF Data) log file.

5.2.4 Mapping Scientometrics (ISI Data)

5.2.4.1 Document Co-Citation

Scientometrics.isi	
Time frame:	1978-2008
Region(s):	Miscellaneous
Topical Area(s):	Scientometrics
Analysis Type(s):	Document Co-Citation Network

Scientometrics is a discipline which uses statistical and computational techniques in order to understand the structure and dynamics of science. Here we use ISI data from the journal "Scientometrics" and Science of Science and Innovation Policy (SciSIP) data from NSF Awards Search.

Download Scientometrics.isi. Load the file using '*File* > *Load*' and locating the downloaded file. This domain level dataset is ideal for document co-citation analysis, as the scale is large enough that the resulting network will paint a

fairly accurate picture of document similarity within the domain of scientometrics.

New ISI File Format

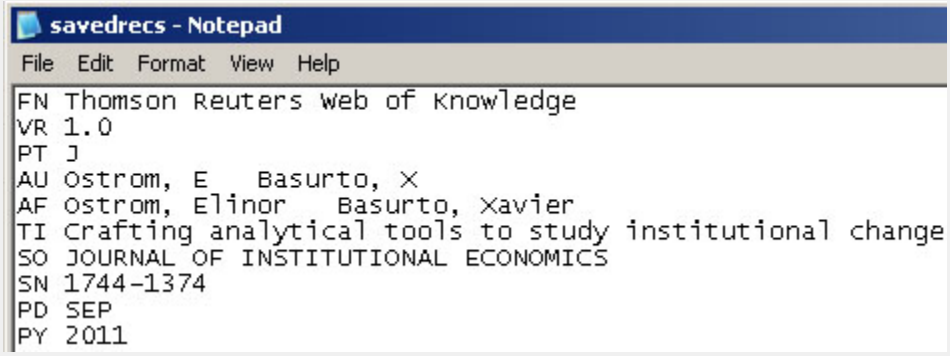
Web of Science made a change to their output format in September, 2011. Older versions of Sci2 tool (older than v0.5.2 alpha) may refuse to load these new files, with an error like "Invalid ISI format file selected."

Sci2 solution

If you are using an older version of the Sci2 tool, you can download the WOS-plugins.zip file and unzip the JAR files into your sci2/plugins/ directory. Restart Sci2 to activate the fixes. You can now load the downloaded ISI files into the Sci2 without any additional step. If you are using the old Sci2 tool you will need to follow the guidelines below before you can load the new WOS format file into the tool.

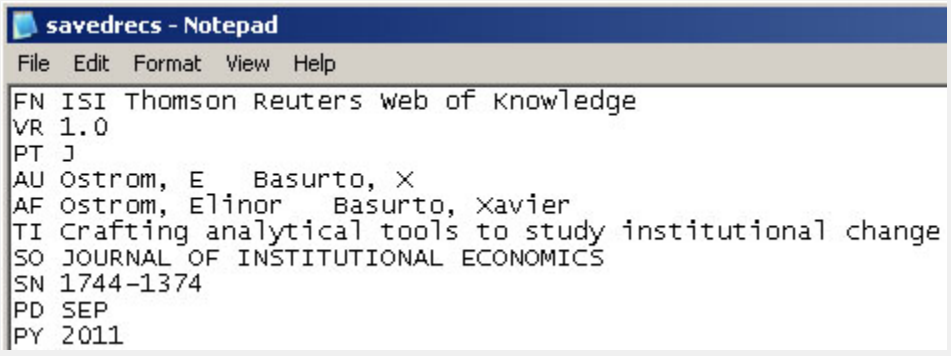
You can fix this problem for individual files by opening them in Notepad (or your favorite text editor). The file will start with the words:

Original file:



Just add the word ISI.

Updated file:



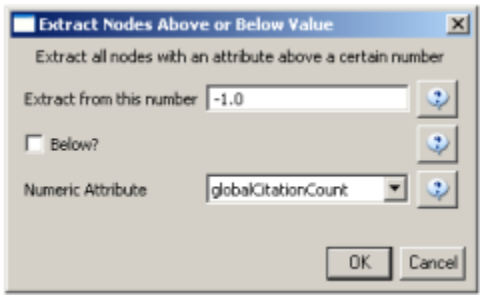
And then Save the file.

The ISI file should now load properly. More information on the ISI file format is available here (http://wiki.cns.iu.edu/display/CISHELL/ISI+%28*.isi%29).

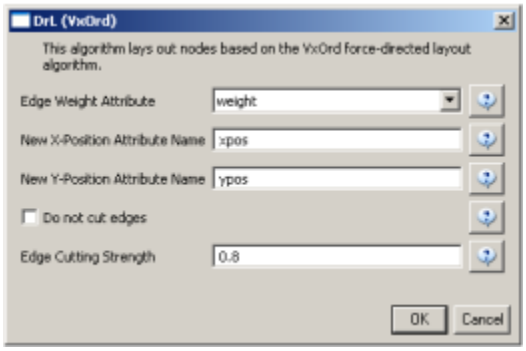
Select the dataset "2126 Unique ISI Records" in the Data Manager window and run *'Data Preparation > Extract Paper Citation Network'*.

Two files will appear in the Data Manager window: the paper-citation network and the paper information table.

Select the "Extracted paper-citation network" and run *'Preprocessing > Networks >* Extract Nodes Above or Below Value*'* with the following parameters:



The produced network contains only the original ISI records. Select the resulting file and run *'Data Preparation >* Extract Document Co-Citation Network*'*. Then, select the network and run *'Analysis > Networks >* Network Analysis Toolkit (NAT)*'*. There are 2056 nodes, 26070 edges and 775 isolates in the network. Run *'Preprocessing >* Delete Isolates*'* to remove all the isolates. Because this network is too dense to lay out in GUESS, run '*Visualization >* DrL (VxOrd)*'* with the parameters:



Next, select "Laid out with DrL" in the Data Manager and run '*Visualization > Network >* GUESS'. Run the following commands in the GUESS "Interpreter":

```
>for n in g.nodes:
    if n.xpos is not None and n.ypos is not None:
        n.x = n.xpos * 10
        n.y = n.ypos * 10
>resizeLinear(localcitationcount,1,50)
>colorize(localcitationcount, gray, black)
>resizeLinear(weight, .25, 8)
>colorize(weight, "127,193,65,255", black)
```

Go to "Graph Modifier" and choose *'Object: nodes based on -> > Property: localcitationcount > Operators: >= > Value: 20 > Show Label'*. See Figure 5.21.

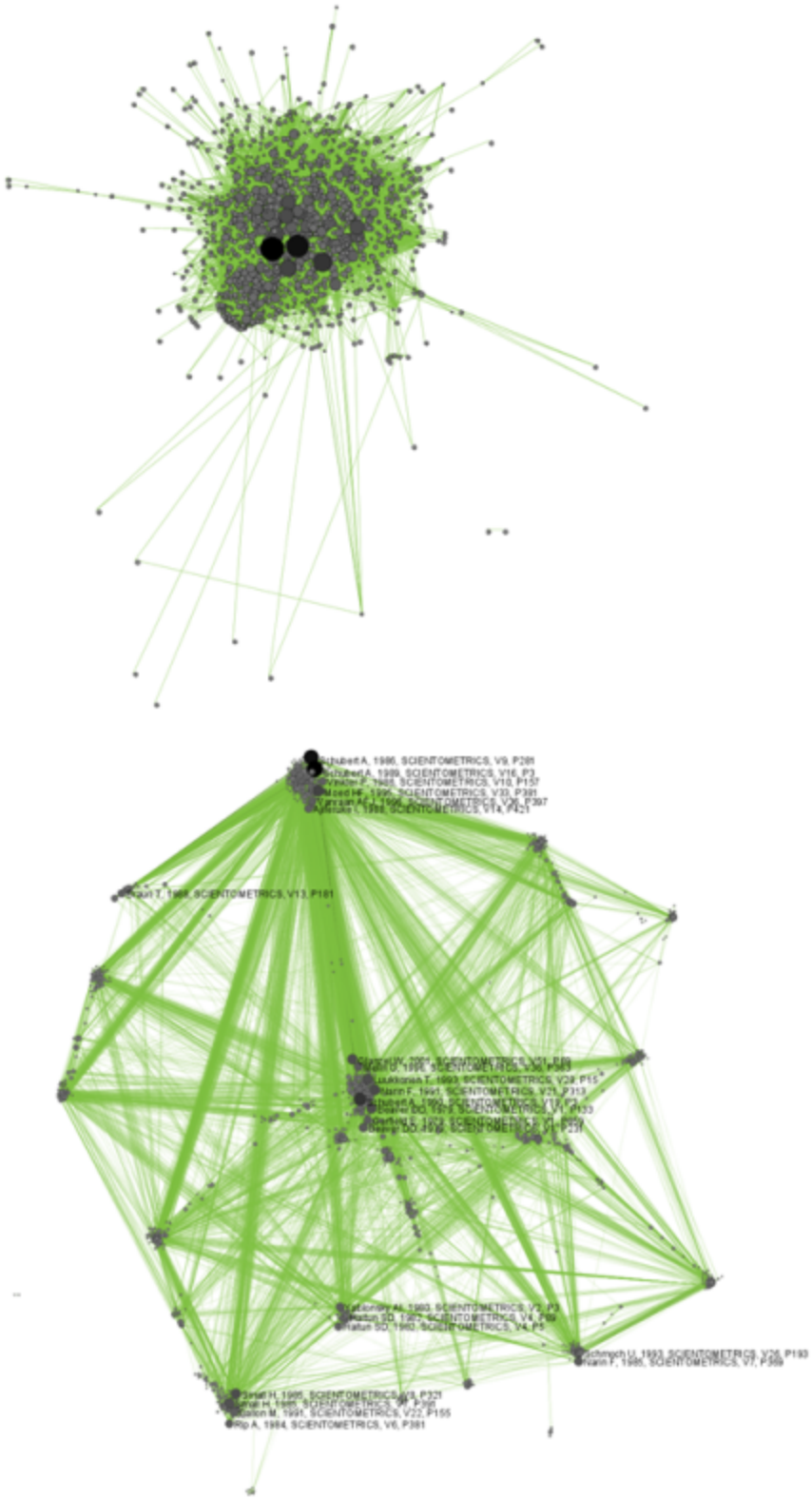


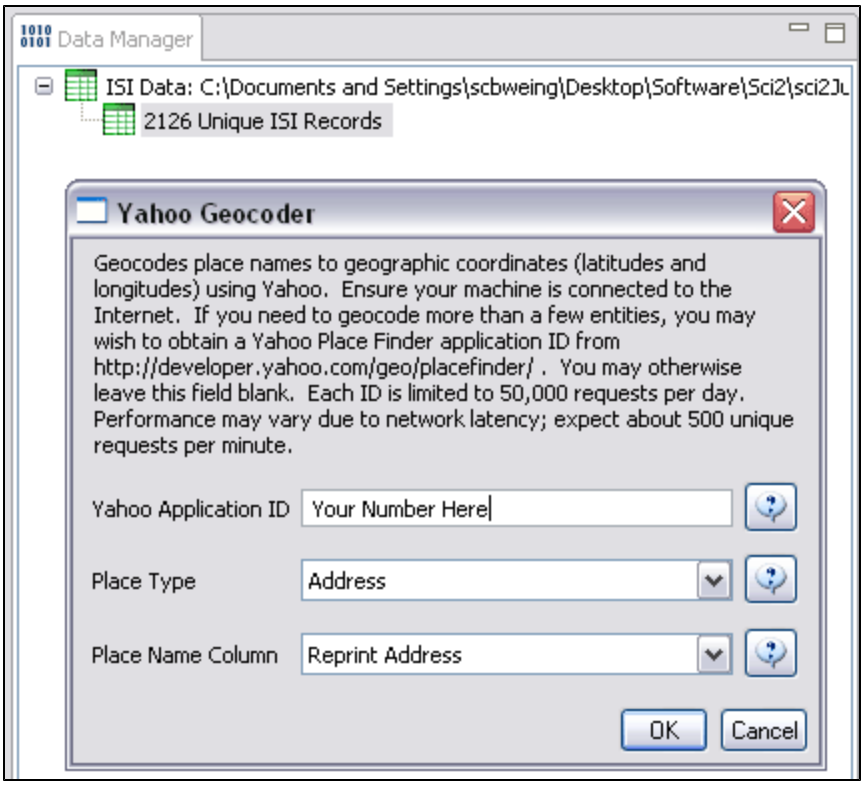
Figure 5.30: Document co-citation network for Scientometric.isi in GUESS without DrL edge cutting (top) and with DrL (VxOrd) (bottom).

To see the log file from this workflow save the 5.2.4.1 Document Co-Citation log file.

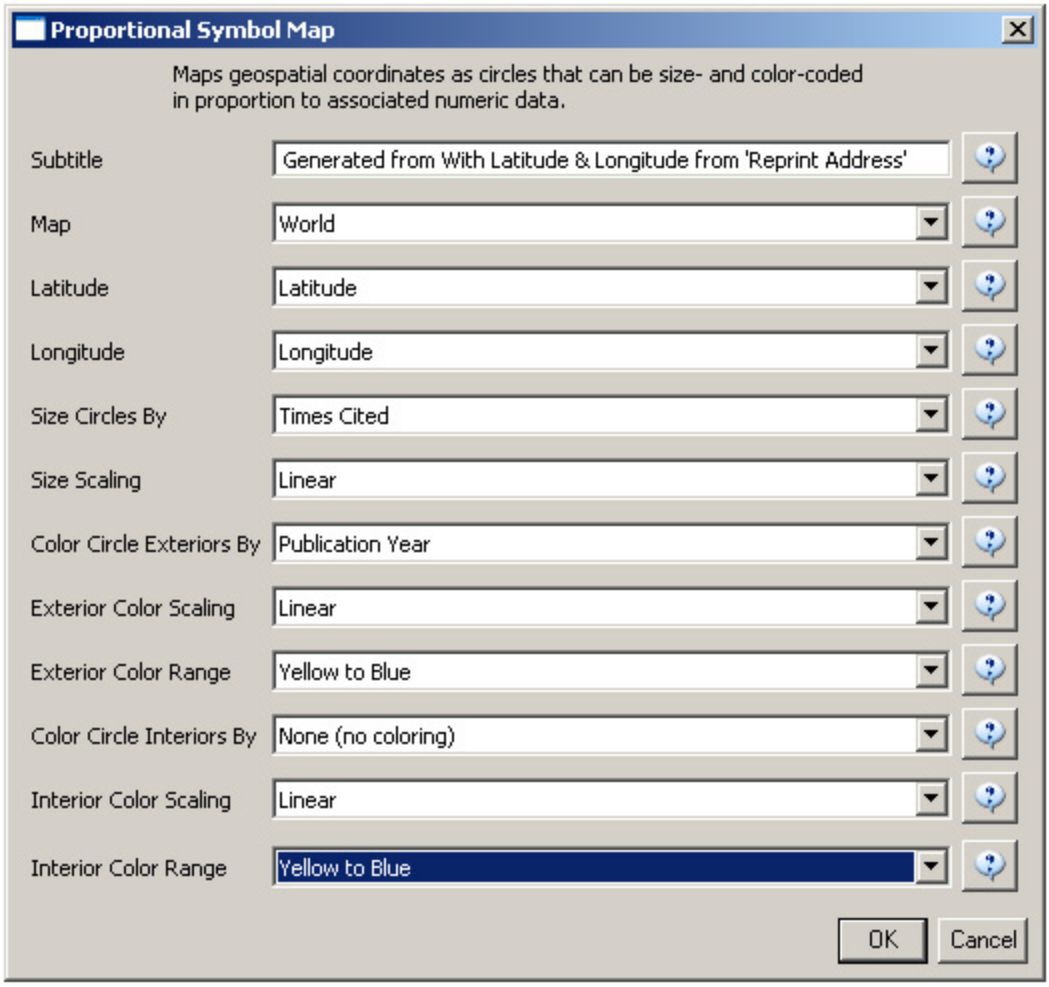
5.2.4.2 Geographic Visualization

Scientometrics.isi	
Time frame:	1978-2008
Region(s):	Miscellaneous
Topical Area(s):	Scientometrics
Analysis Type(s):	Geospatial Analysis

Using the dataset loaded from section 5.2.4.1, select '*2126 Unique ISI Records*' in the Data Manager. To find the latitude and longitudes of the locations of researches publishing in *Scientometrics*, run *Analysis* > *Geospatial* > *Yahoo Geocoder*. Note that you will need a Yahoo Application ID in order to run this query. Enter your Yahoo Application ID (sign up for one here), "Address" for the Place Type, and the "Reprint Address" for the Place Column Name. Press 'OK'. This step may take several minutes to complete.



A new table will appear in the Data Manager labeled '*With Latitude & Longitude from 'Reprint Address''* containing all data from the initial table, plus columns with the Yahoo Geocoded Latitude and Longitude Coordinates. Select this file and run *Visualization* > *Geospatial* > *Proportional Symbol Map* using the parameters listed below.



Save and view the resulting visualization using the directions described at 2.4 Saving Visualizations for Publication. This will be the result:

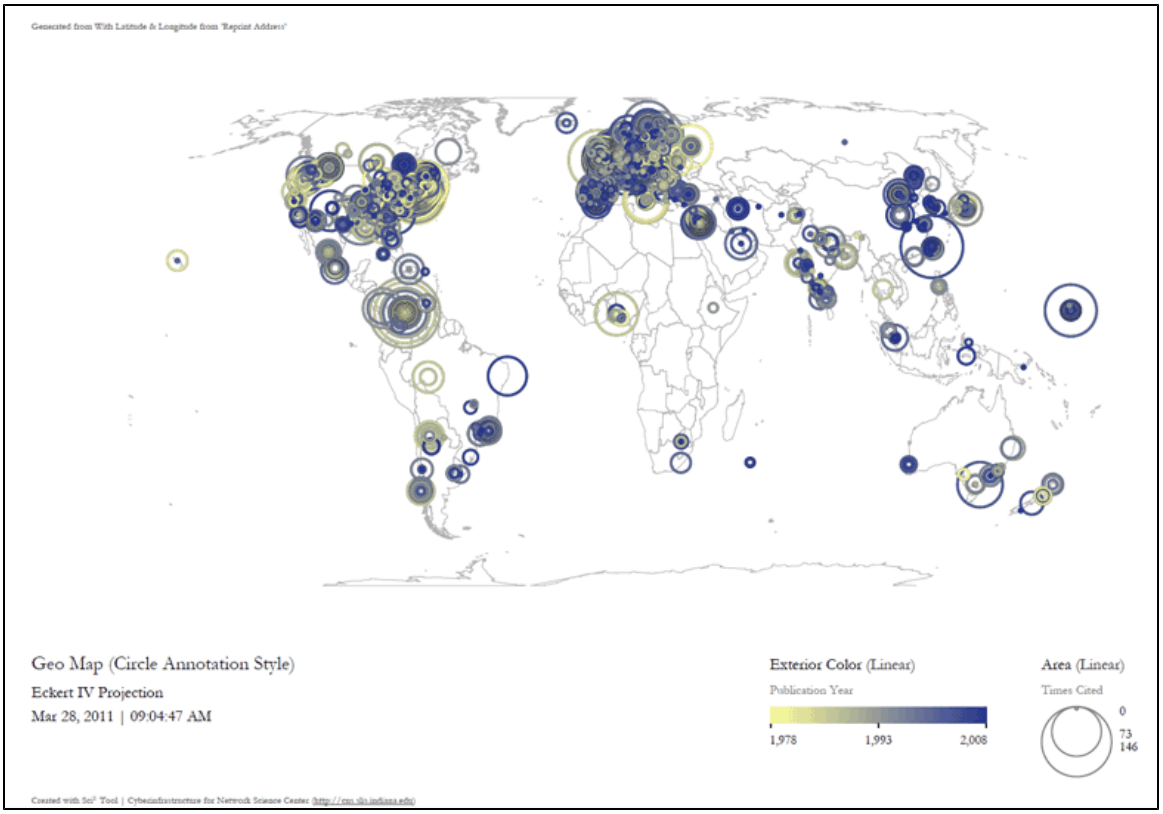


Figure 5.31: Scientometrics.isi geospatial visualization. Circle sizes represent times cited, colors represent publication year.

5.2.5 Burst Detection in Physics and Complex Networks (ISI Data)

AlessandroVespignani.isi	
Time frame:	1990-2006
Region(s):	Indiana University, University of Rome, Yale University, Leiden University, International Center for Theoretical Physics, University of Paris-Sud
Topical Area(s):	Informatics, Complex Network Science and System Research, Physics, Statistics, Epidemics
Analysis Type(s):	Burst Detection

Burst Detection

A scholarly dataset can be understood as a discrete time series: in other words, a sequence of events/ observations which are ordered in one dimension – time. Observations exist for regularly spaced intervals, e.g., each month or year.

The burst detection algorithm (see Section 4.6.1 Burst Detection) identifies sudden increases or "bursts" in the frequency-of-use of character strings over time. This algorithm identifies topics, terms, or concepts important to the events being studied that increased in usage, were more active for a period of time, and then faded away.

An analysis of publications authored or co-authored by Alessandro Vespignani from 1990 to 2006 will be used to illustrate the "burst" concept. Alessandro Vespignani is an Italian physicist and Professor of Informatics and Cognitive Science at Indiana University, Bloomington. In his publications, it is possible to see a change in research focus - from Physics to Complex Networks - beginning in 2001.

Load Alessandro Vespignani's ISI publication history using '*File* > *Load*' and following this path: '***yoursci2directory****/sampledata/scientometrics/isi/AlessandroVespignani.isi*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets).

New ISI File Format

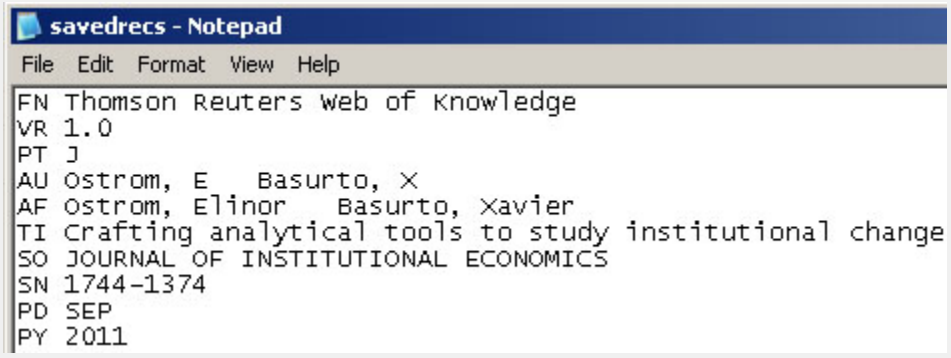
Web of Science made a change to their output format in September, 2011. Older versions of Sci2 tool (older than v0.5.2 alpha) may refuse to load these new files, with an error like "Invalid ISI format file selected."

Sci2 solution

If you are using an older version of the Sci2 tool, you can download the WOS-plugins.zip file and unzip the JAR files into your sci2/plugins/ directory. Restart Sci2 to activate the fixes. You can now load the downloaded ISI files into the Sci2 without any additional step. If you are using the old Sci2 tool you will need to follow the guidelines below before you can load the new WOS format file into the tool.

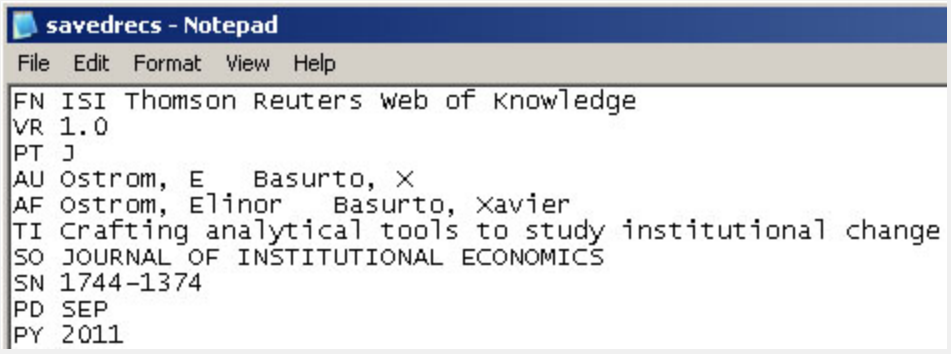
You can fix this problem for individual files by opening them in Notepad (or your favorite text editor). The file will start with the words:

Original file:



Just add the word ISI.

Original file:

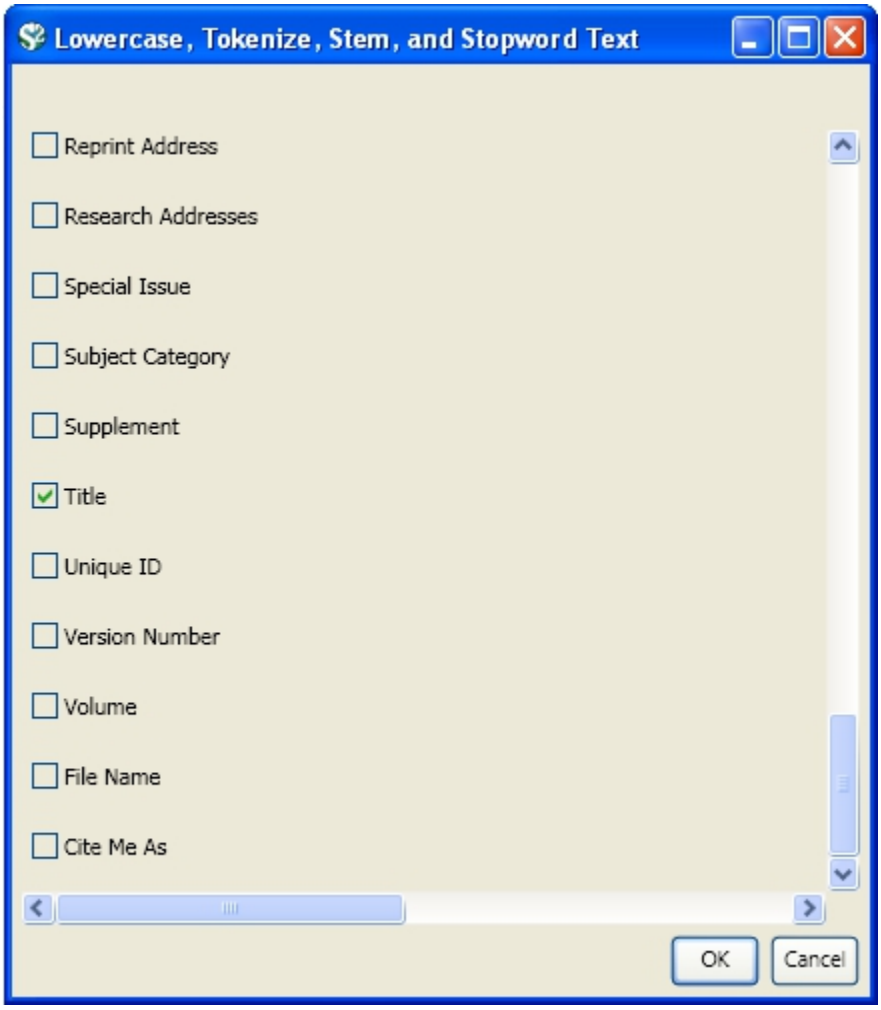


And then Save the file.

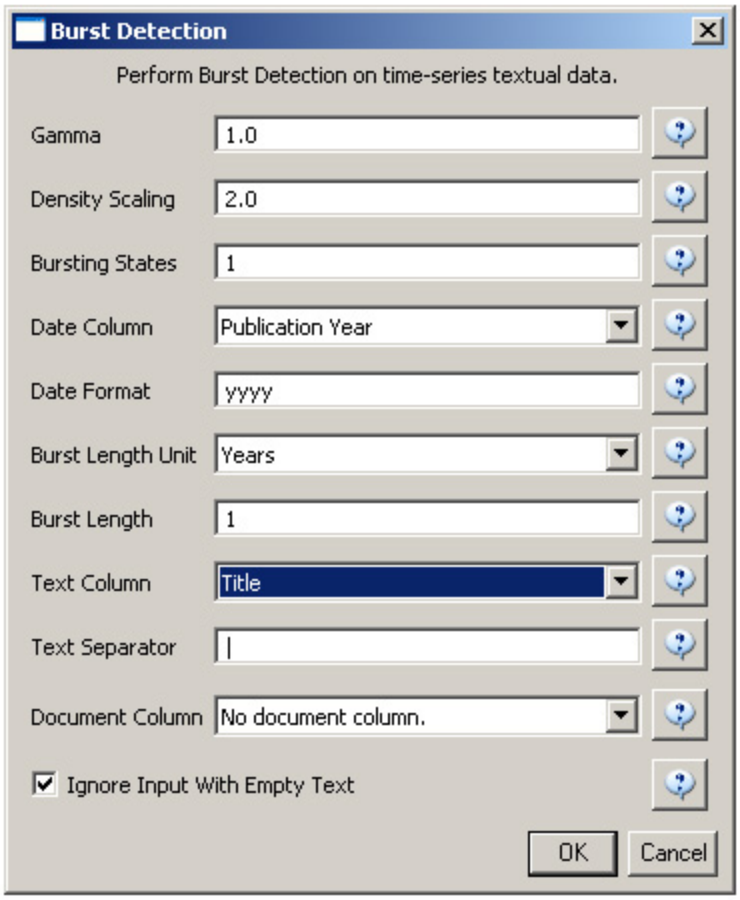
The ISI file should now load properly. More information on the ISI file format is available here (http://wiki.cns.iu.edu/display/CISHELL/ISI+%28*.isi%29).

This analysis will detect the "bursty" terms used in the title of papers in the dataset. Since the burst detection algorithm is case-sensitive, it is necessary to normalize the field to be analyzed before running the algorithm. Select the table "101 Unique ISI Records" and run *'Preprocessing* > *Topical* > *Lowercase, Tokenize, Stem, and Stopword*

Text.' Check the "Title" box to indicate that you want to normalize this field:



Select the resulting "with normalized Title" table in the Data Manager and run '*Analysis* > *Topical* > *Burst Detection*' with the following parameters:



The "Gamma" parameter is the value that state transition costs are proportional to. This parameter is used to control how ease the automaton can change states. The higher the "Gamma" value, the smaller the list of bursts generated.

The "Density Scaling" parameter determines how much 'more bursty' each level is beyond the previous one. The higher the scaling value, the more active (bursty) the event happens in each level.

The "Bursting States" parameter determines how many bursting states there will be, beyond the non-bursting state. An *i* value of bursting states is equals to *i + 1* automaton states.

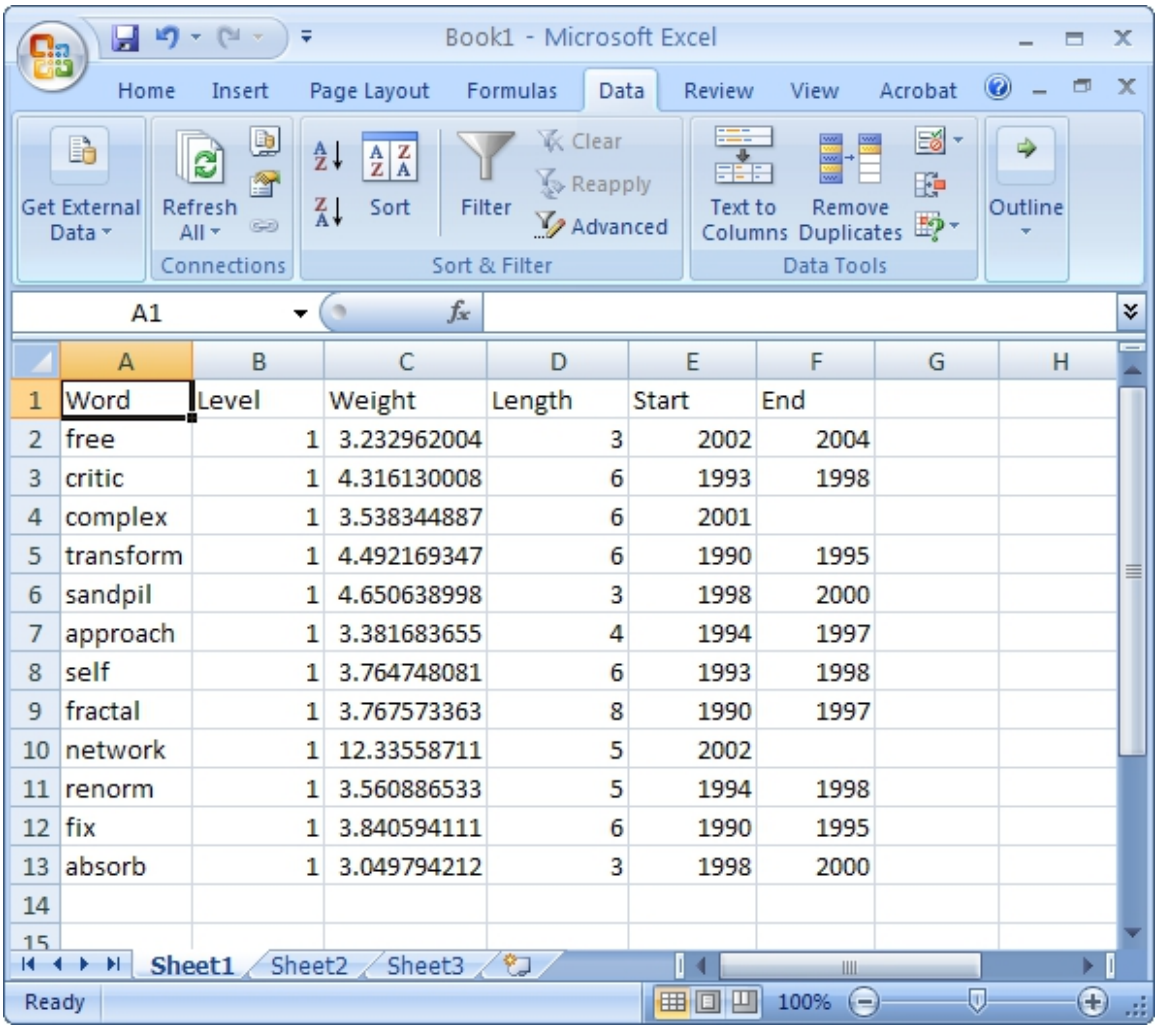
The "Date Column" parameter is the name of the column with date/time when the events / topics happens.

The "Date Format" specifies how the date column will be interpreted as a date/time. See http://java.sun.com/j2se/1.4.2/docs/api/java/text/SimpleDateFormat.html for details.

The "Text Column" parameter is the name of the column with values (delimiter and tokens) to be computed for bursting results.

The "Text Separator" parameters determines the separator that was used to delimit the tokens in the text column.

View the file "Burst detection analysis (Publication Year, Title): maximum burst level 1". On a PC running Windows, right click on this table and select view to see the data in Excel. On a Mac or a Linux system, right click and save the file, then open using the spreadsheet program of your choice.



In this table, there are six columns: "Word," "Level," "Weight," "Length," "Start," and "End."

The "Word" field identifies the specific character string which was detected as a "burst." The "Length" field indicates how long the burst lasted (over the selected time parameter).

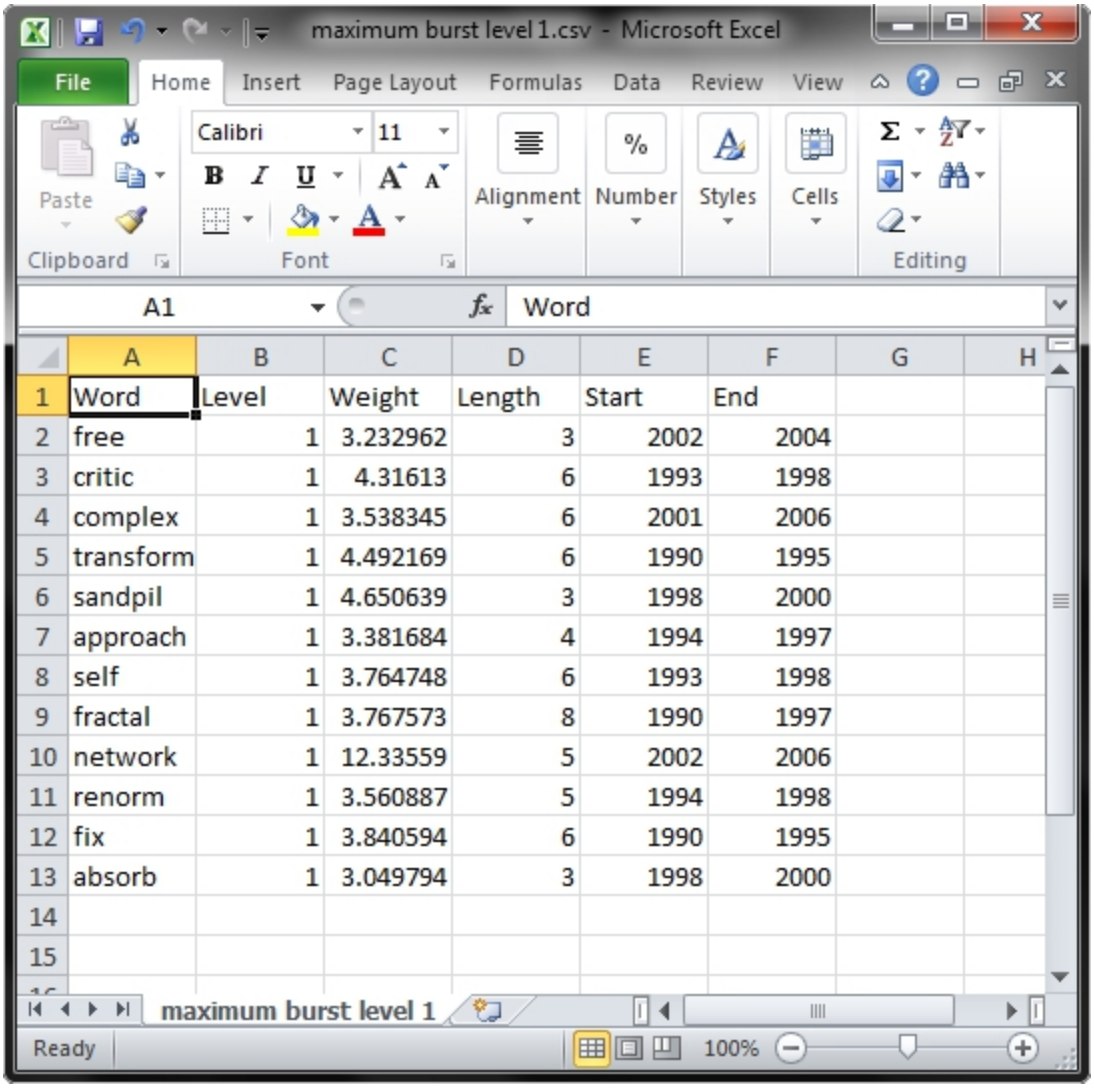
The "Level" is the burst level of this burst. The higher burst level, the more frequent the event / topic happens.

The "Weight" field is the weight of this burst between its "Length". A higher weight could be resulted by the longer "Length", the higher "Level" or both.

The "Length" is the period of the burst. It is generated based on *(Start - End + 1)*.

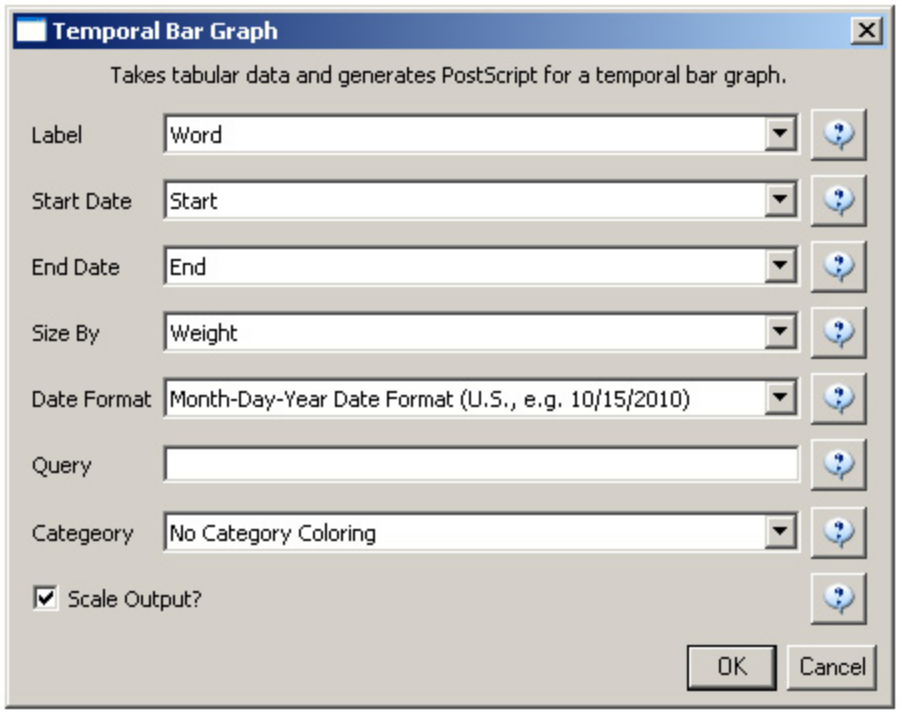
The "Start" field identifies when the burst began (again, according to the specified time parameter).

And the "End" field indicates when the burst stopped. A empty value in the "End" field indicates that the burst lasted until the last date present in the dataset. Where the "End" field is empty, put manually the last year present in the dataset. In this case, 2006.



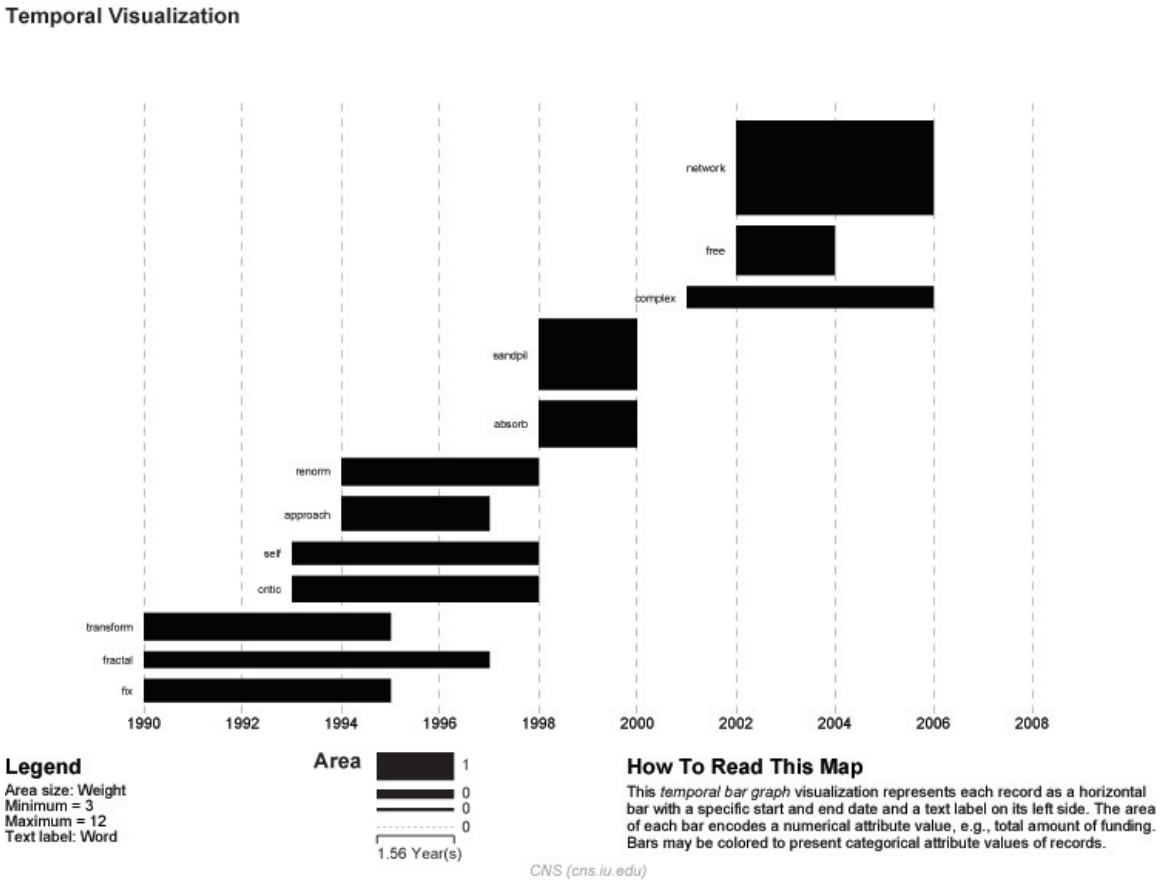
	A	B	C	D	E	F	G	H
1	Word	Level	Weight	Length	Start	End		
2	free	1	3.232962	3	2002	2004		
3	critic	1	4.31613	6	1993	1998		
4	complex	1	3.538345	6	2001	2006		
5	transform	1	4.492169	6	1990	1995		
6	sandpil	1	4.650639	3	1998	2000		
7	approach	1	3.381684	4	1994	1997		
8	self	1	3.764748	6	1993	1998		
9	fractal	1	3.767573	8	1990	1997		
10	network	1	12.33559	5	2002	2006		
11	renorm	1	3.560887	5	1994	1998		
12	fix	1	3.840594	6	1990	1995		
13	absorb	1	3.049794	3	1998	2000		
14								
15								

After you add manually this information, save this .csv file somewhere in your computer. Load back this .csv file into Sci2 using '*File* > *Load*'. Select 'Standart csv format' int the pop-up window. A new table will appear in the Data Manager. To visualize these table that contains the results of the Burst Detection algorithm, select the table you just loaded in the Data Manager and run '*Visualization* > *Temporal* > *Temporal Bar Graph*' with the following parameters:



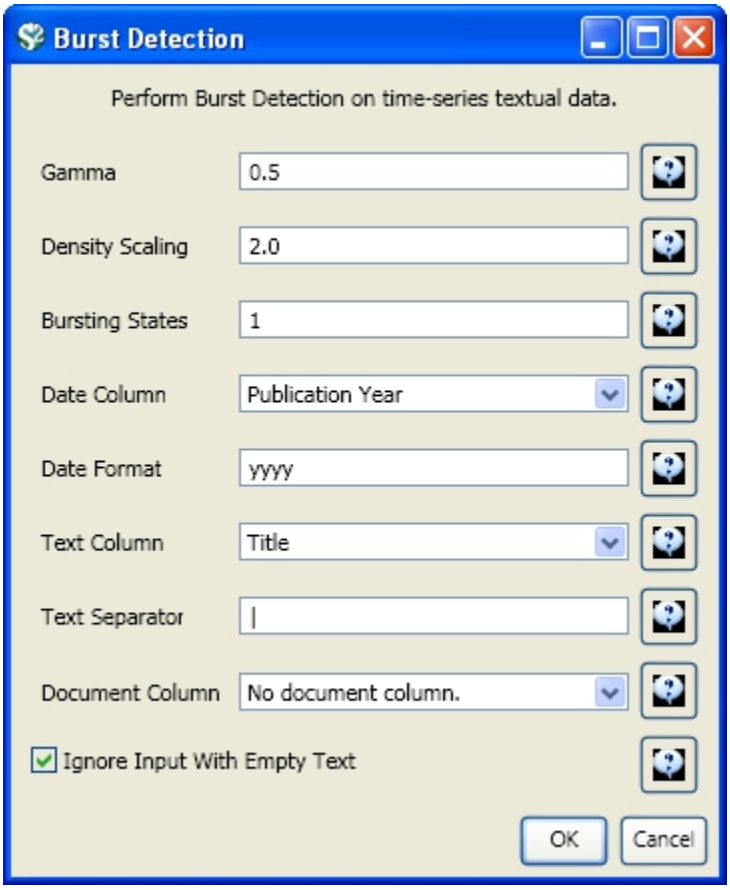
Horizontal bar graphs are used to visualize numeric data over time, generating labeled horizontal bars. A PostScript file containing the horizontal bar graph will appear in the Data Manager.

Open and view the file using the workflow from Section 2.4 Saving Visualizations for Publication.



The resulting analysis indicates a change in the research focus of Alessandro Vespignani for publications beginning in 2001. For example, the bursting terms "fractal," "growth," "transform," and "fix" starting at 1990 are related to Vespignani's Ph.D., entitled "Fractal Growth and Self-Organized Criticality" in Physics. Other bursts also related to Physics follow these, such as "sandipil." After 2001, bursting terms like "complex," "network," "free," and "weight" appear, signifying a change in Vespignani's research area from Physics to Complex Networks, with a larger number of publications on topics like "weighted networks" and "scale-free networks."

Now, let's run the Burst Detection algorithm again for the same dataset but for a different value for the 'Gamma' parameter. Select the table 'with normalized Title' in the Data Manager and run '*Analysis* > *Topical* > *Burst Detection* ' with the following parameters:



Notice that the value for the gamma parameter is now set to 0.5. The parameter gamma controls the ease with which the automaton can change states. With a smaller gamma value, more bursts will be generated. Running the algorithm with these parameters will generate a new table named "Burst detection analysis (Publication Year, Title): maximum burst level 1.2" in the Data Manager.

Microsoft Excel - maximum burst level 1.2.csv

	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	free	1	3.232962	3	2002	2004
3	epidem	1	2.532198	6	2001	
4	disloc	1	2.267251	2	2001	2002
5	system	1	1.45763	4	1996	1999
6	critic	1	4.31613	6	1993	1998
7	complex	1	3.538345	6	2001	
8	transform	1	4.492169	6	1990	1995
9	conserv	1	1.972471	3	1998	2000
10	field	1	1.834723	4	1997	2000
11	fractur	1	1.593632	6	1994	1999
12	dynam	1	2.404396	2	2001	2002
13	forest	1	1.809792	3	1995	1997
14	phase	1	2.446386	1	2000	2000
15	weight	1	2.716894	3	2004	
16	aggreg	1	2.50294	5	1991	1995
17	sandpil	1	4.650639	3	1998	2000
18	approach	1	3.381684	4	1994	1997
19	growth	1	2.839977	8	1990	1997
20	limit	1	1.45763	5	1991	1995
21	self	1	3.764748	6	1993	1998
22	similar	1	1.45763	5	1991	1995
23	fractal	1	3.767573	8	1990	1997
24	transit	1	2.783221	4	1997	2000
25	global	1	1.719802	4	2003	
26	scale	1	2.66792	4	1990	1993
27	state	1	1.875465	6	1996	2001
28	network	1	12.33559	5	2002	
29	evolut	1	1.804025	4	2003	
30	renorm	1	3.560887	5	1994	1998
31	fix	1	3.840594	6	1990	1995
32	absorb	1	3.049794	3	1998	2000
33	organ	1	2.751329	5	1994	1998
34	group	1	2.201667	5	1995	1999

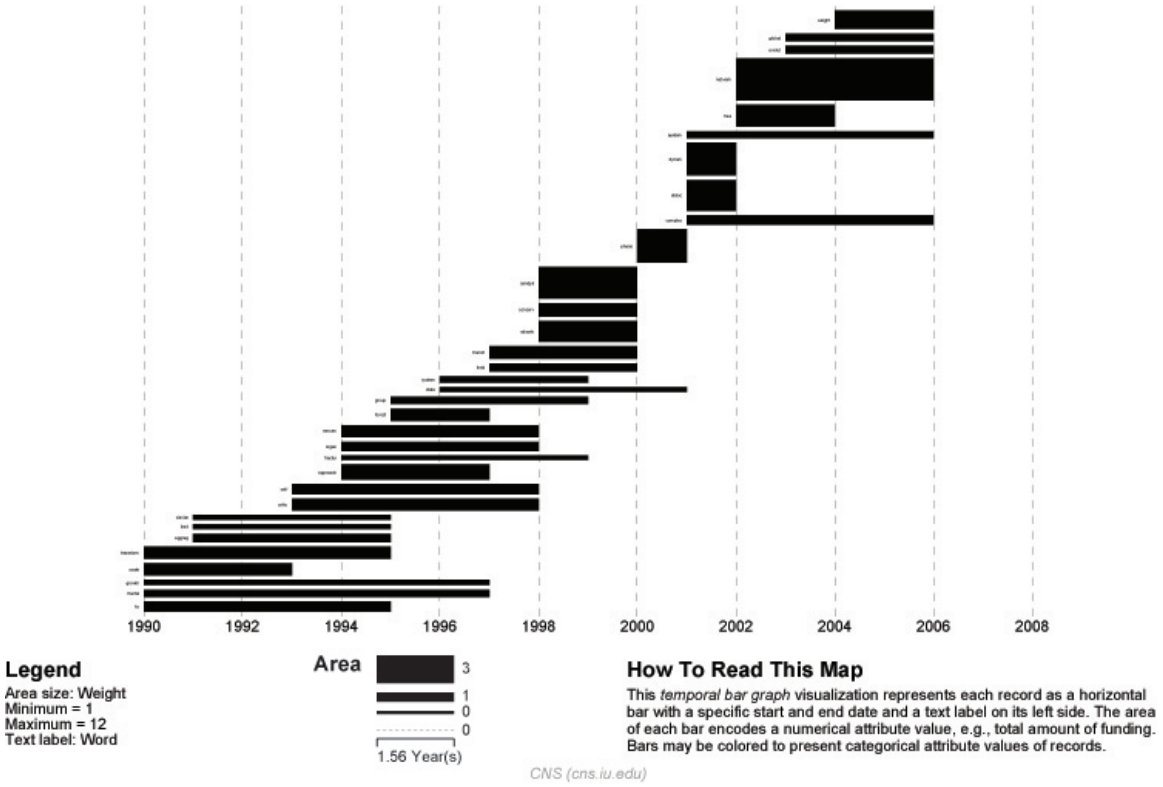
Again where the "End" field is empty, put manually the last year present in the dataset. In this case, 2006.

Microsoft Excel - maximum burst level 1.2.csv

	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	free	1	3.232962	3	2002	2004
3	epidem	1	2.532198	6	2001	2006
4	disloc	1	2.267251	2	2001	2002
5	system	1	1.45763	4	1996	1999
6	critic	1	4.31613	6	1993	1998
7	complex	1	3.538345	6	2001	2006
8	transform	1	4.492169	6	1990	1995
9	conserv	1	1.972471	3	1998	2000
10	field	1	1.834723	4	1997	2000
11	fractur	1	1.593632	6	1994	1999
12	dynam	1	2.404396	2	2001	2002
13	forest	1	1.809792	3	1995	1997
14	phase	1	2.446386	1	2000	2000
15	weight	1	2.716894	3	2004	2006
16	aggreg	1	2.50294	5	1991	1995
17	sandpil	1	4.650639	3	1998	2000
18	approach	1	3.381684	4	1994	1997
19	growth	1	2.839977	8	1990	1997
20	limit	1	1.45763	5	1991	1995
21	self	1	3.764748	6	1993	1998
22	similar	1	1.45763	5	1991	1995
23	fractal	1	3.767573	8	1990	1997
24	transit	1	2.783221	4	1997	2000
25	global	1	1.719802	4	2003	2006
26	scale	1	2.66792	4	1990	1993
27	state	1	1.875465	6	1996	2001
28	network	1	12.33559	5	2002	2006
29	evolut	1	1.804025	4	2003	2006
30	renorm	1	3.560887	5	1994	1998
31	fix	1	3.840594	6	1990	1995
32	absorb	1	3.049794	3	1998	2000
33	organ	1	2.751329	5	1994	1998
34	group	1	2.201667	5	1995	1999

After you add manually this information, save this .csv file somewhere in your computer. Load back this .csv file into Sci2 using '*File* > *Load*'. Select 'Standart csv format' int the pop-up window. A new table will appear in the Data Manager. To visualize these table that contains these new results for the Burst Detection algorithm, select the table you just loaded in the Data Manager and run '*Visualization* > *Temporal* > Horizontal Bar Graph (not included version)' with the same parameters.

A new PostScript file containing the horizontal bar graph will appear in the Data Manager. Once more, open and view the file using the workflow from Section 2.4 Saving Visualizations for Publication.



As expected, a larger number of bursts appear, and the new bursts have a smaller weight that those depicted in the first graph. These smaller, more numerous bursting terms permit a more detailed view of the dataset and allow the identification of trends. The "protein" burst starting in 2003, for example, indicates the year in which Alessandro Vespignani started to work with "protein-protein interaction networks," while the burst "epidem" - also from 2001 - is related to the application of complex networks to the analysis of epidemic phenomena in biological networks.

Visualizing Burst Detection in Excel

Its possible to generate a visualization for burst analysis in MS Excel. For this, open the results of the first burst analysis conducted ('Burst detection analysis (Publication Year, Title): maximum burst level 1') in MS Excel, by right clicking on this table in the Data Manager and selecting View.

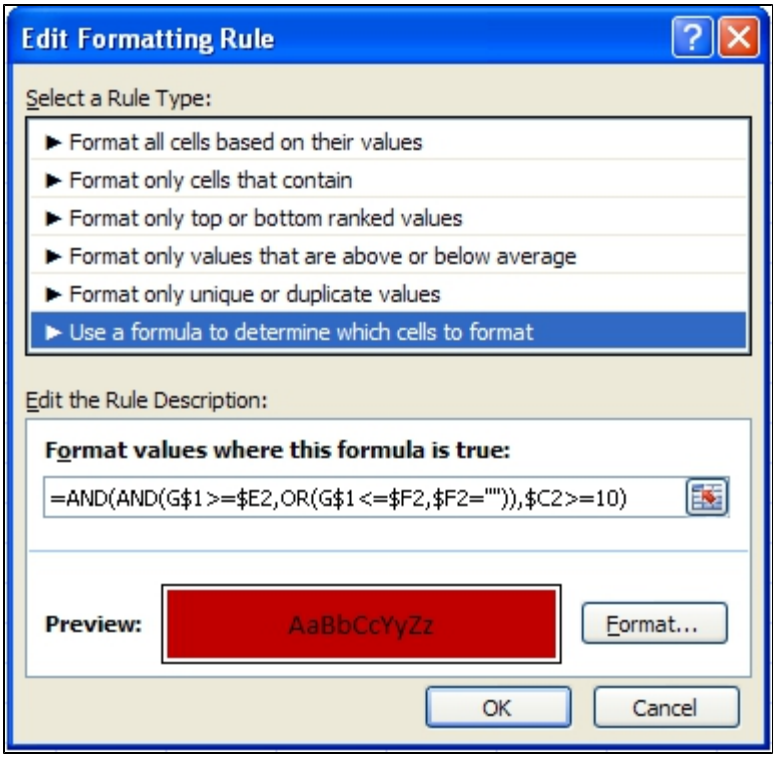
To generate a visual depiction of the bursts in MS Excel perform the following steps:

1. Sort the data ascending by burst start year.

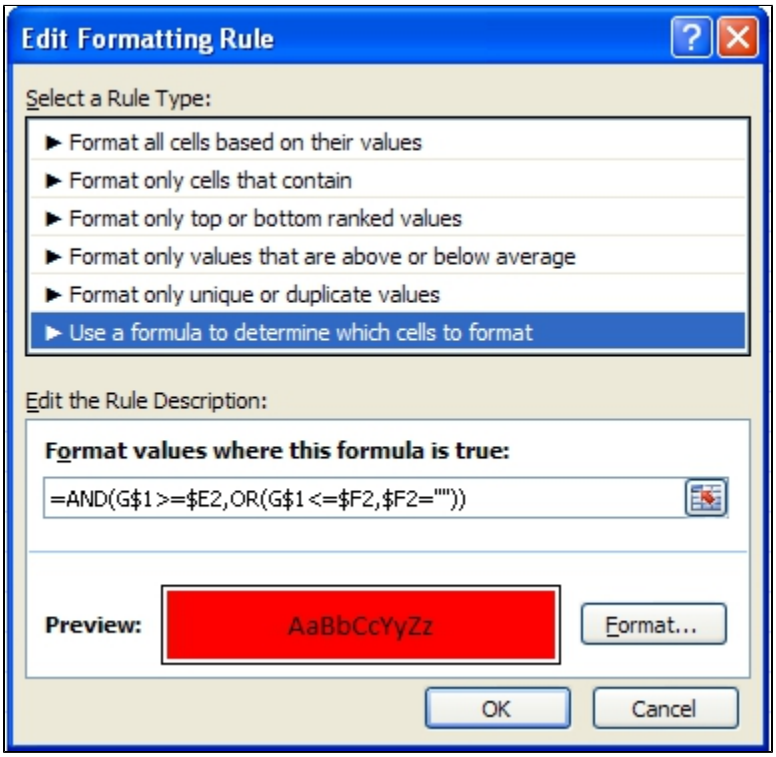
2. Add column headers for all years, i.e., enter first the start year in the cell of index G1, here 1990. As stated before, when there is no value in the "End" field that indicates that the burst lasted until the last date present in the dataset. So continue, e.g., using formula *'=G1+1'*, until highest burst end year, here 2006 in cell W1.

3. In the resulting word by burst year matrix, select the upper left blank cell (G2) and select 'Conditional Formatting' from the ribbon. Then select '*Data Bars > More Rules > Use a formula to determine which cells to format.'* To color cells for years with a burst weight value of more or equal 10 red and cells with a higher value dark red use the following formulas and format patterns:

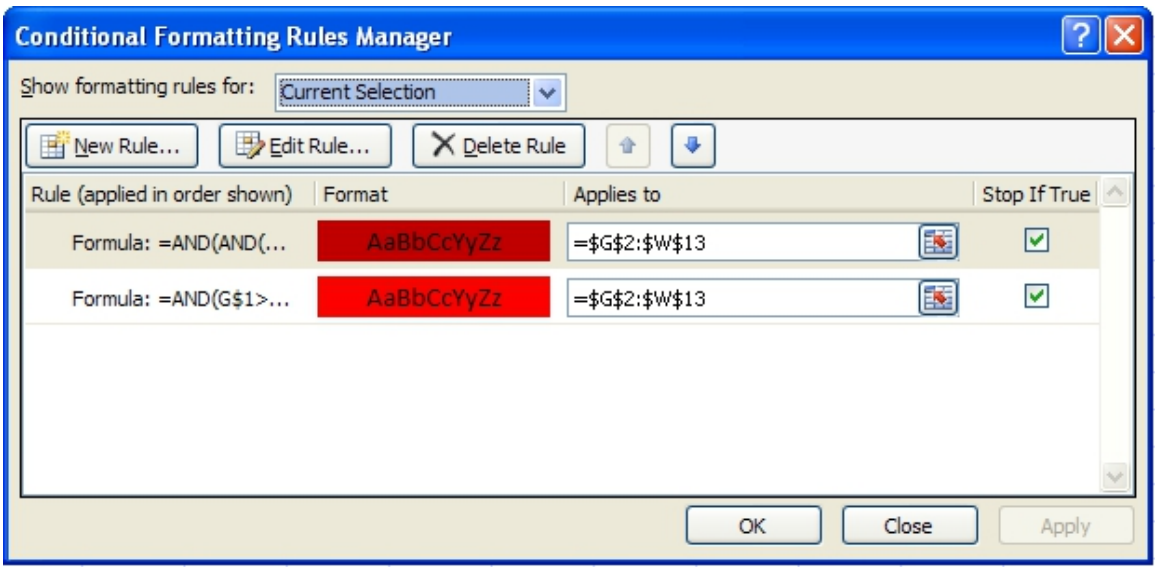
=AND(AND(G$1>=$E2,OR(G$1<=$F2,$F2="")),$C2>=10)



Select *'OK'* and then repeat step three, using the formula below:
=AND(G$1>=$E2,OR(G$1<=$F2,$F2=""))



4. Once both formatting rules have been established, select *'Conditional Formatting > Manage Rules'*, highlight the first formatting rule and move to the top of the list:



5. Make sure both formatting rules are selected and apply them to current selection. Apply the format to all cells in the word by year matrix by dragging the box around cell G2 to highlight all cells in the matrix. The result for the given example is shown in Figure 5.33

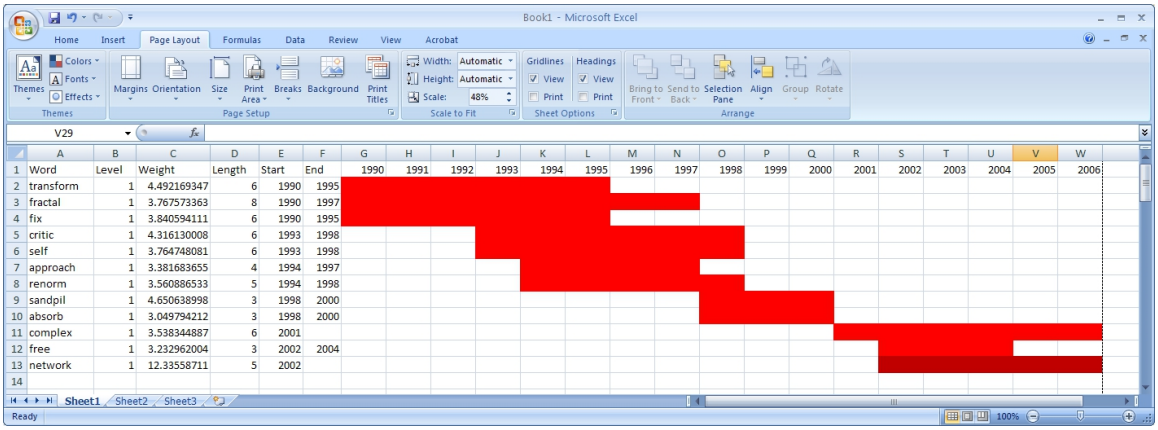


Figure 5.32.1: Visualizing burst results in MS Excel

5.2.6 Mapping the Field of RNAi Research (SDB Data)

RNAi	
Time frame:	1865-2008
Region(s):	Miscellaneous
Topical Area(s):	RNAi
Analysis Type(s):	Co-Author Network, Patent-Citation Network, Burst Detection

The data for this analysis comes from a search of the Scholarly Database (SDB) (http://sdb.cns.iu.edu/) for "RNAi" in

"All Text" from Medline, NSF, NIH and USPTO. A copy of this data is available in '***yoursci2directory****/sampledata/scientometrics/sdb/RNAi*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). The default export format is .csv, which can be loaded directly into the Sci^2 Tool.

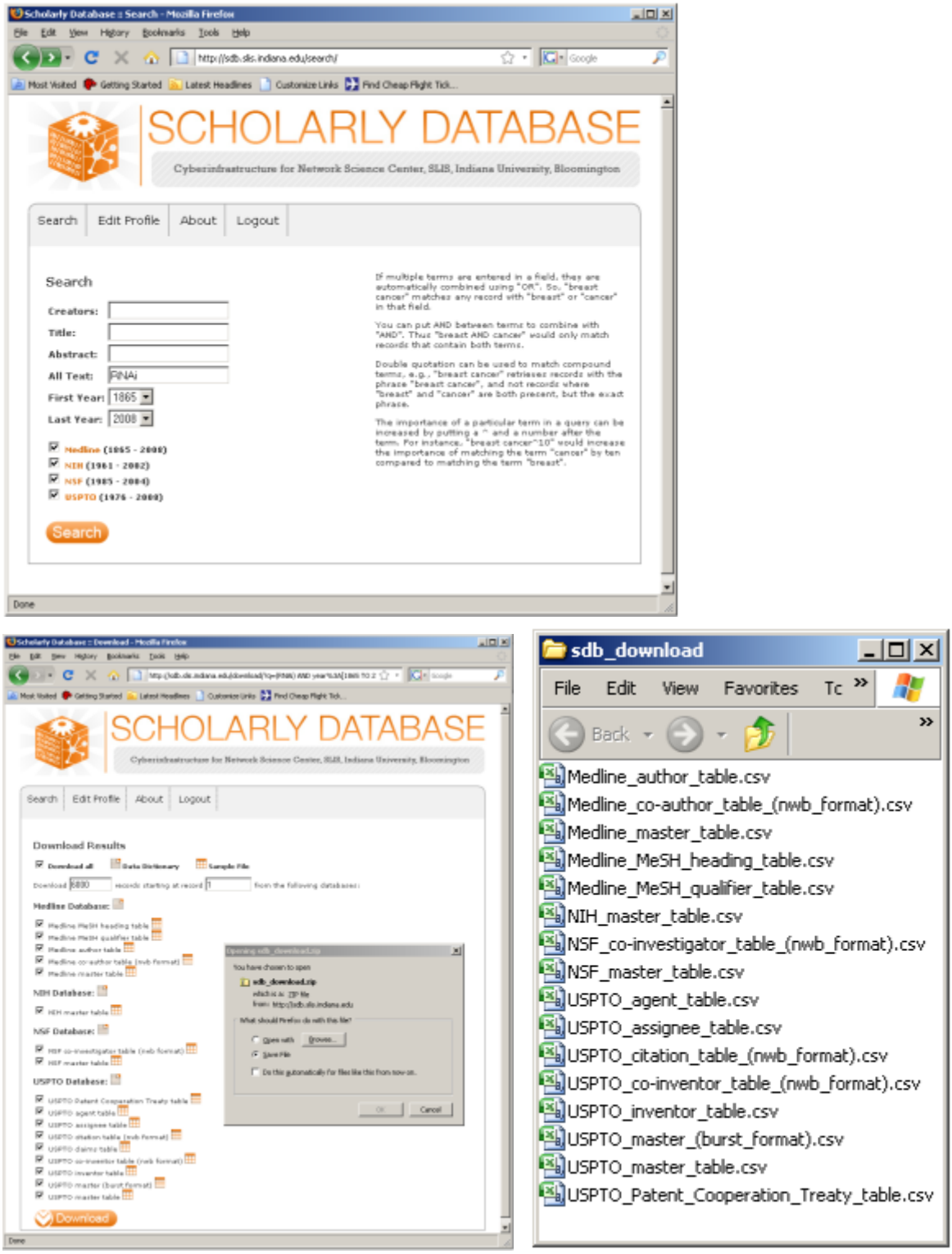
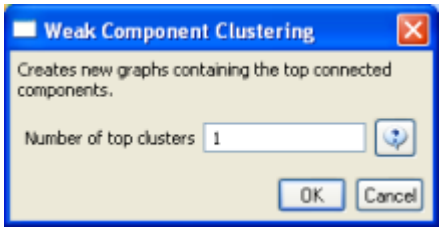


Figure 5.33: Downloading and saving RNAi data from the Scholarly Database.

To view the co-authorship network of Medline's RNAi records, go to '*File > Load*' and open '***yoursci2directory****/sampledata/scientometrics/sdb/RNAi/Medline_co-author_table*(nwb_format).csv'_ in Standard csv format (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). SDB tables are already normalized,

so simply run '*Data Preparation* > *Extract Co-Occurrence Network*' using the default parameters:

According to '*Analysis* > *Networks* > *Network Analysis Toolkit (NAT)*', the output network has 21,578 nodes with 131 isolates, and 77,739 edges. Visualizing such a large network is memory-intensive, so extract only the largest connected component by running '*Analysis* > *Networks* > *Unweighted and Undirected* > *Weak Component Clustering*' with the following parameters:

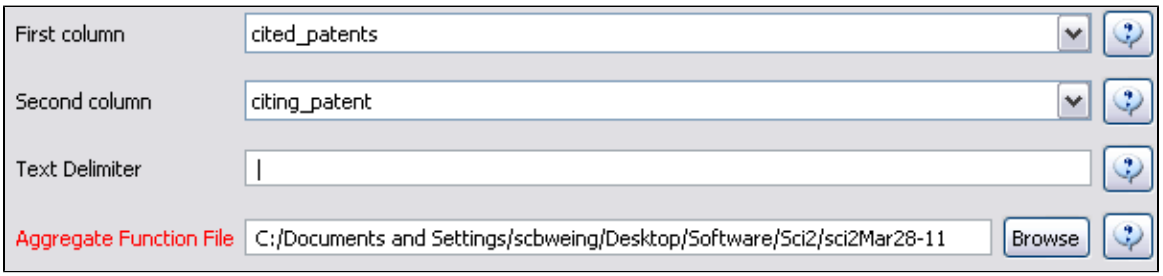


Make sure the newly extracted network ("Weak Component Cluster of 6446 nodes") is selected in the data manager, and run '*Visualization* > *Networks* > *GUESS*' followed by '*Layout* > *GEM*'. A custom python script has been used to color and size the network in Figure 5.34.

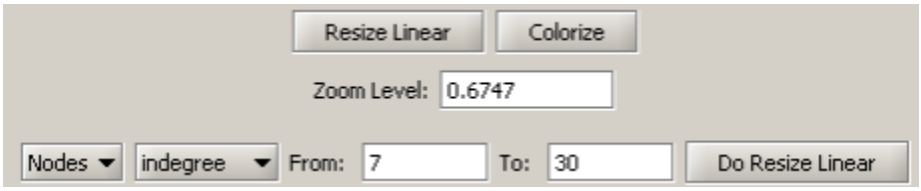
Figure 5.34: The Medline RNAi Co-authorship Network

To visualize the citation patterns of patents dealing with RNAi, load '***yoursci2directory****/sampledata/scientometrics/sdb/RNAi/USPTO_citation_table*(nwb_format).csv'_ in Standard csv format (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Then run '*Data Preparation* > *Extract Bipartite Network*' u

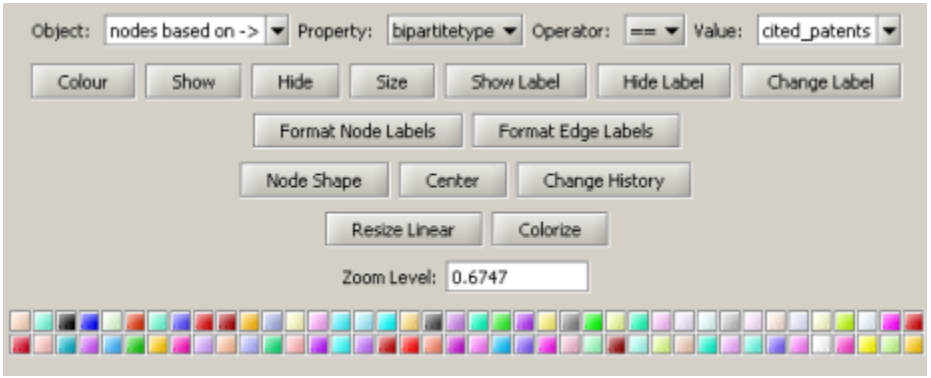
sing the following parameters:



Run '*Analysis* > *Networks* > *Unweighted & Directed* > *Node Indegree*' to append Indegree attributes to each node, and then visualize "Network with indegree attribute added to node list" using '*Visualization* > *Networks* > *GUESS*' followed by '*Layout* > *GEM*' and *'Layout* > *Bin Pack'*. In the graph modifier pane, use the following parameters and click "Do Resize Linear."



Then select "nodes based on ->" in the Object drop-down box, "bipartitetype" in the Property drop-down box, "==" in the Operator drop-down box, and "cited_patents" in the Value drop-down box. Press "Colour" and click on blue below.



Repeat the previous steps, but change the Value to "citing_patent" and select the color red. Now press "Show Label". The resulting graph should look like Figure 5.35.

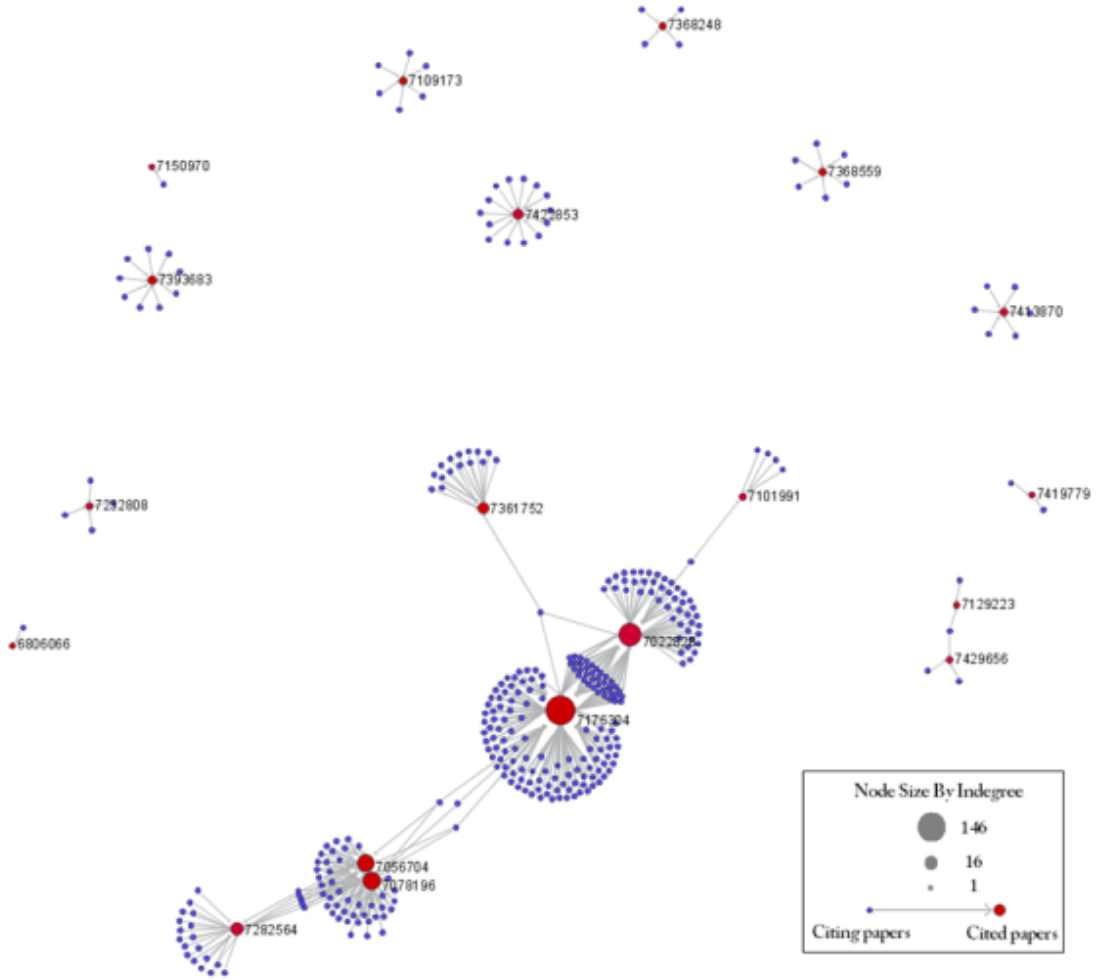
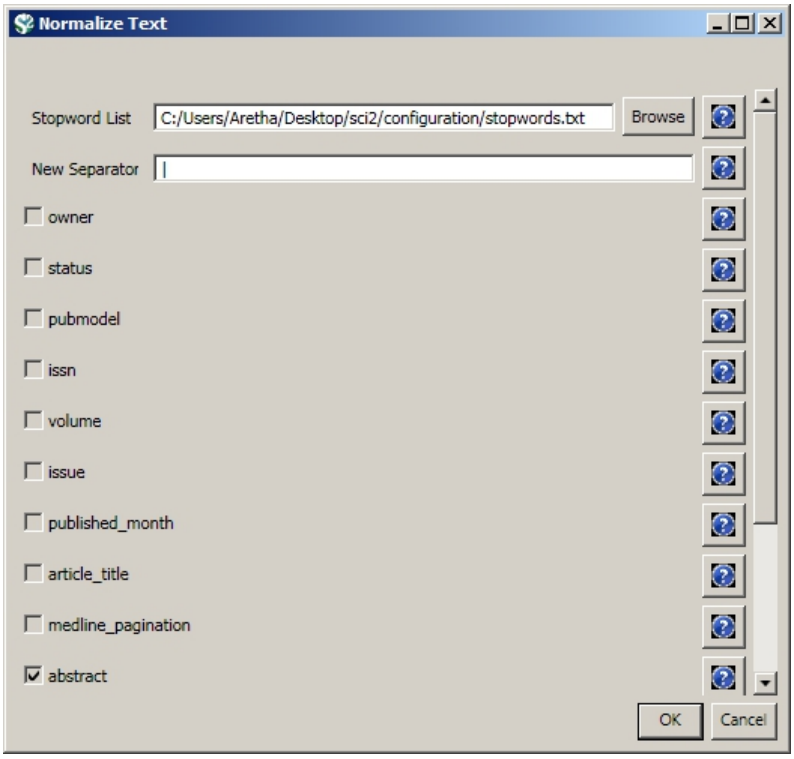


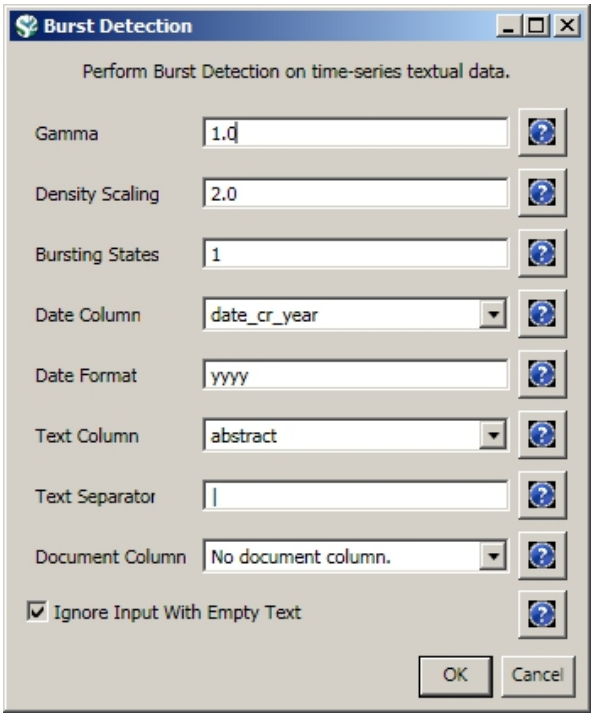
Figure 5.35: USPTO Patent citation network on RNAi

The SDB also outputs much more robust tables, for example '***yoursci2directory****/sampledata/scientometrics/sdb/RNAi/Medline_master_table.csv'.* This table includes full records of Medline papers, and will be used to find bursting terms from Medline abstracts dealing with RNAi. You can download the file by clicking here.

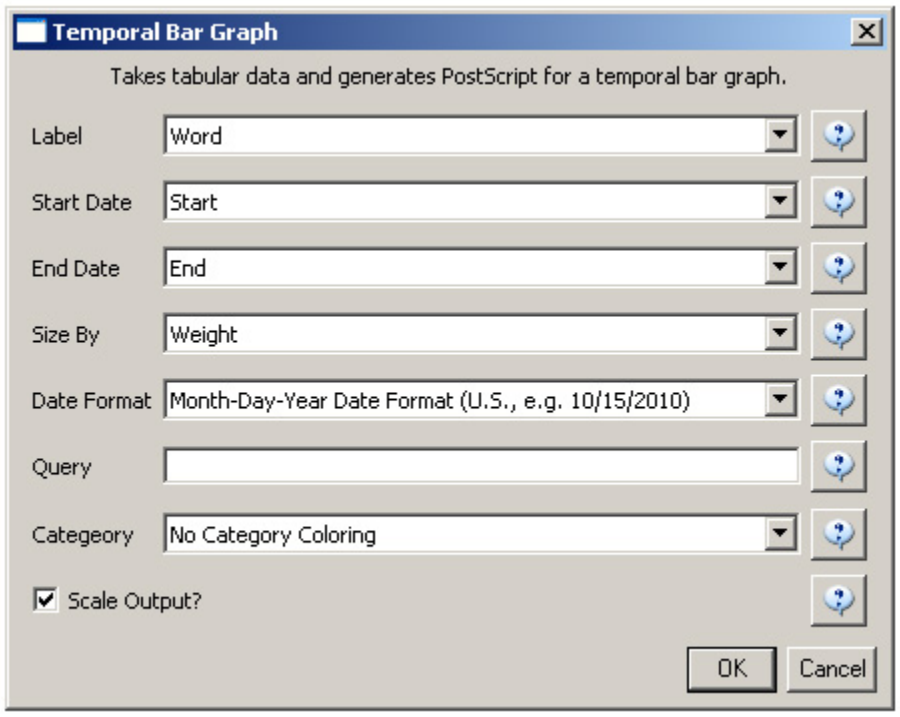
Load the file in Standard csv format and run '*Preprocessing > Topical > Lowercase, Tokenize, Stem, and Stopword Text*' with the following parameters:



Select the "with normalized abstract" table in the Data Manager and run '*Analysis* > *Topical* > *Burst Detection*' with the following parameters:



View the file "Burst detection analysis (date_cr_year, abstract): maximum burst level 1." There are more words than can easily be viewed with the horizontal bar graph, so sort the list by "Strength" and prune all but the strongest 10 words. In other words save the file from the data manager and sort the file by the weight column. Delete all but the top 10 rows. Save the file as a new .csv and load it into the Sci^2 Tool as a standard csv file.
Select the new table in the data manager and visualize it using '*Visualize* > *Temporal* > *Horizontal Bar Graph (not included version)Temporal Bar Graph*' with the following parameters:



Save and view the resulting PostScript file using the workflow described in section [2.4 Saving Visualizations for Publication](#).

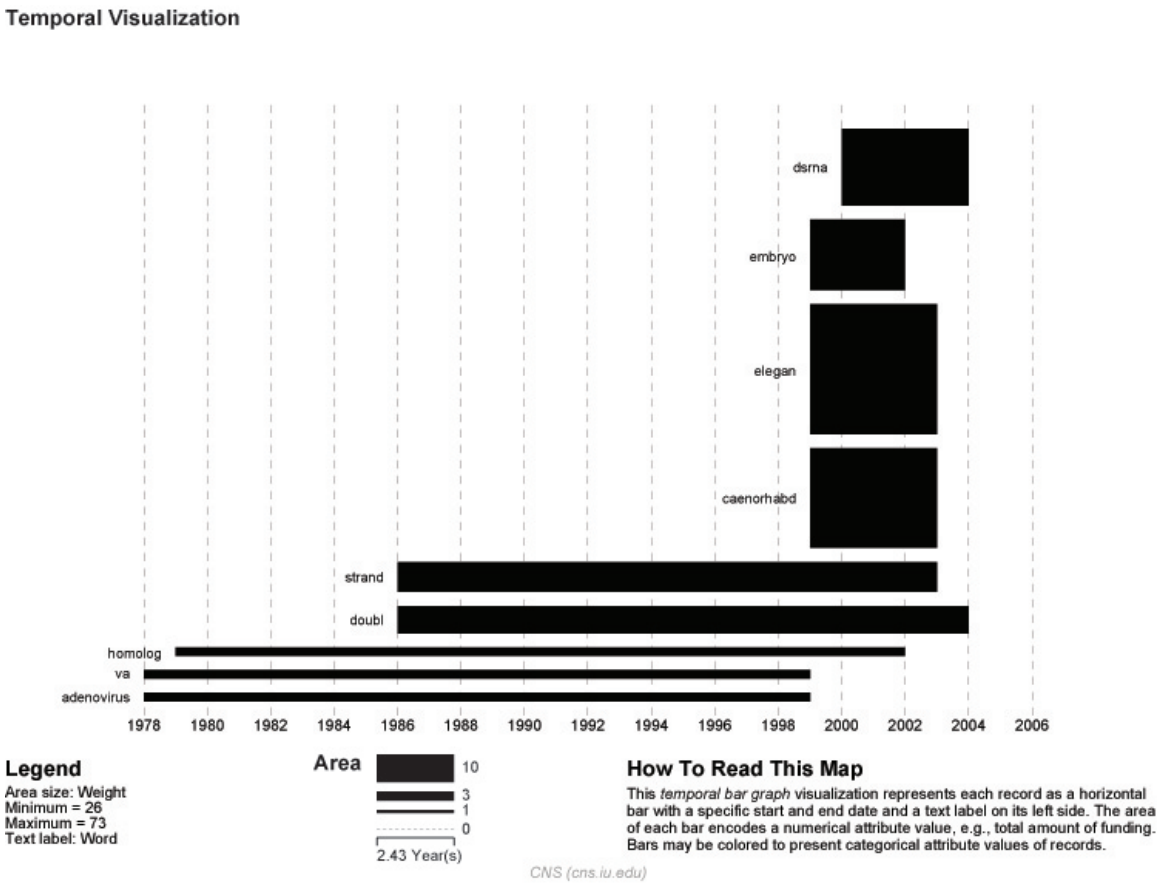


Figure 5.36: Top ten burst terms from Medline abstracts on RNAi Global Level Studies – Meso

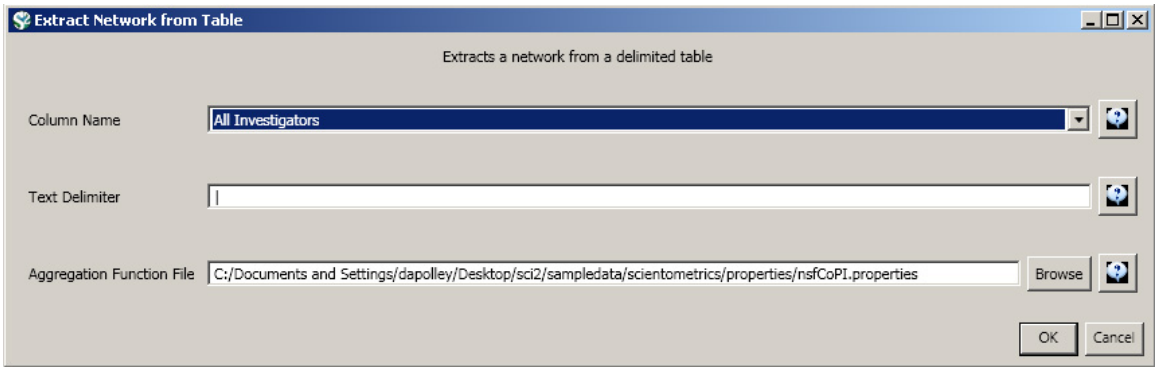
To see the log file from this workflow save the 5.2.6 Mapping the Field of RNAi Research (SDB Data) log file.

5.2.1 Funding Profiles of Three Universities (NSF Data)

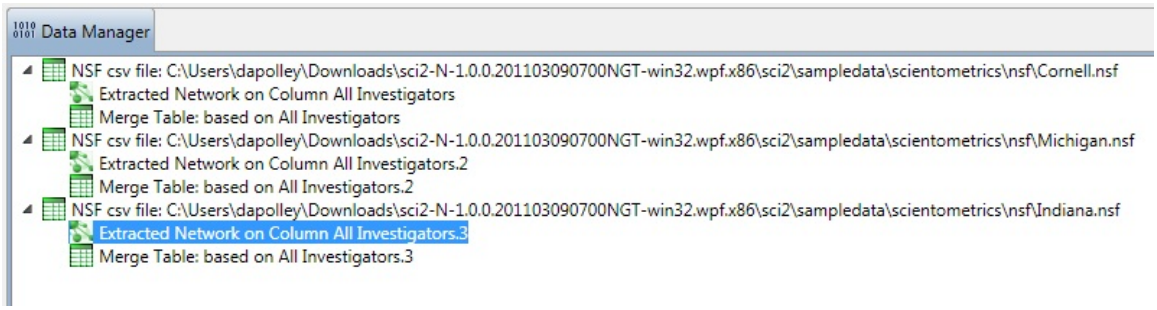
Cornell.nsfIndiana.nsfMichigan.nsf	
Time frame:	2000-2009
Region(s):	Cornell University, Indiana University, Michigan University
Topical Area(s):	Miscellaneous
Analysis Type(s):	Co-PI Network

Load '*Cornell.nsf*', '*Michigan.nsf*', and '*Indiana.nsf*' using '*File > Load*'' and following this path: '***yoursci2directory****/sampledata/scientometrics/nsf*' (if these files are not in the sample data directory they can be downloaded from 2.5 Sample Datasets). Use the following workflow for each of the three nsf files loaded. Select each of the datasets in the Data Manager window and run *'Data Preparation > Extract Co-Occurrence Network'* using the following parameters (Note that the Aggregation Function File is '***yoursci2directory****/sampledata/scientometrics/properties/nsfCoPI.properties'*):

> **Aggregate Function File**
>
> Make sure to use the aggregate function file indicated in the image below. Aggregate function files can be found in **sci2/sampledata/scientometrics/properties**.



Two derived files per dataset will appear in the Data Manager window: the co-PI network and a merge table.



In the network, nodes represent investigators and edges denote their co-PI relationships. (To learn how the merge table can be used to further clean PI names, see section 5.1.4.2 Author Co-Occurrence (Bibliographic Coupling)

Network.)

Choose the "Extracted Network on Column All Investigators" and run *'Analysis > Networks >* Network Analysis Toolkit (NAT)' for each dataset. This will display the amount of nodes and edges, as well as the amount of isolate nodes that can be removed by running *'Preprocessing > Networks >* Delete Isolates'.
Select *'Visualization > Networks >* GUESS' and run '*Layout > GEM'* followed by '*Layout > Bin Pack'* to visualize the network. Run the '***yoursci2directory****/scripts/GUESS/co-PI-nw.py'* script. Visualizations of the three university's co-PI networks are shown in Figure 5.23.

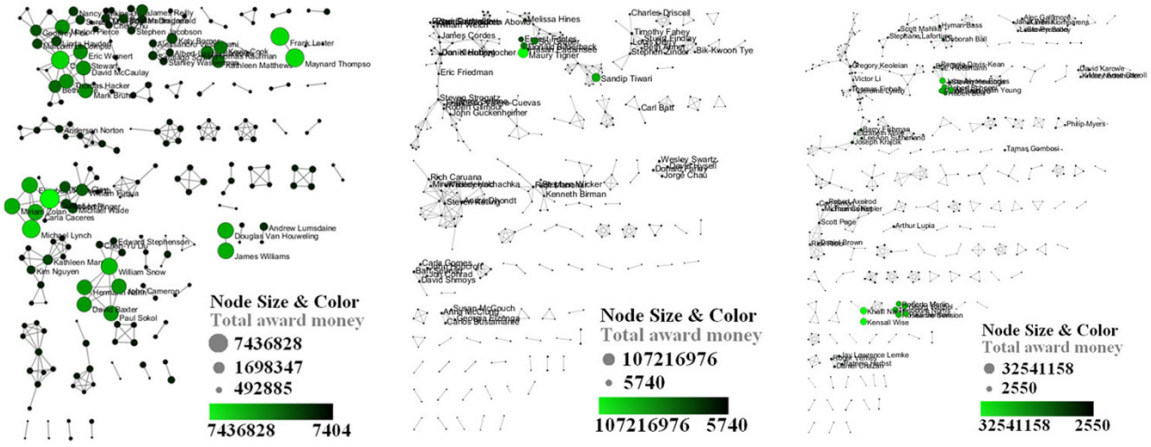
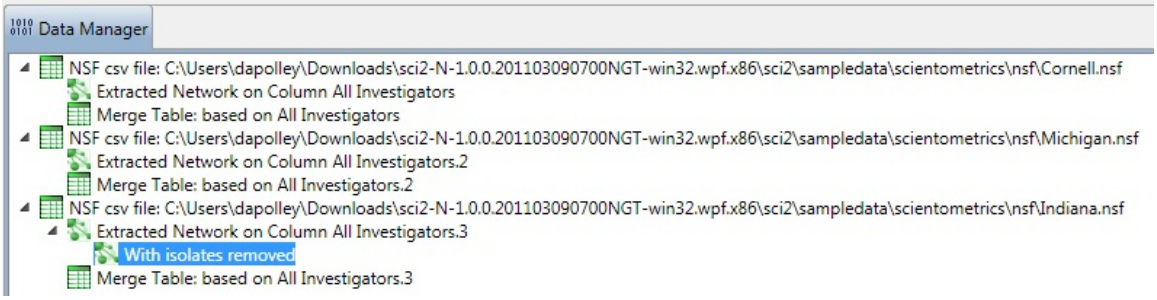
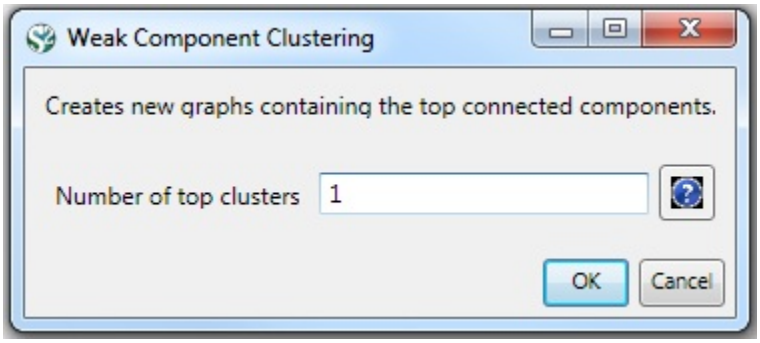


Figure 5.23: Co-PI network of Indiana University (top, left), Cornell University (top, right), University of Michigan (middle).

To see a more detailed view of any of the components in the network (e.g. the largest Indiana component) select the network with deleted isolates in the Data Manager:



Then, run *'Analysis > Networks > Unweighted & Undirected >* Weak Component Clustering' with the parameter:



Indiana's largest component has 19 nodes, Cornell's has 67 nodes, and Michigan's has 55 nodes. Visualize Indiana's network in GUESS using the '***yoursci2directory****/scripts/GUESS/co-PI-nw.py'* script. Save the file as a jpg

by selecting *'File > Export Image'*. Use the "Browse..." option in the "Export Image – GUESS" popup window to select the folder in which you would like to save the image.

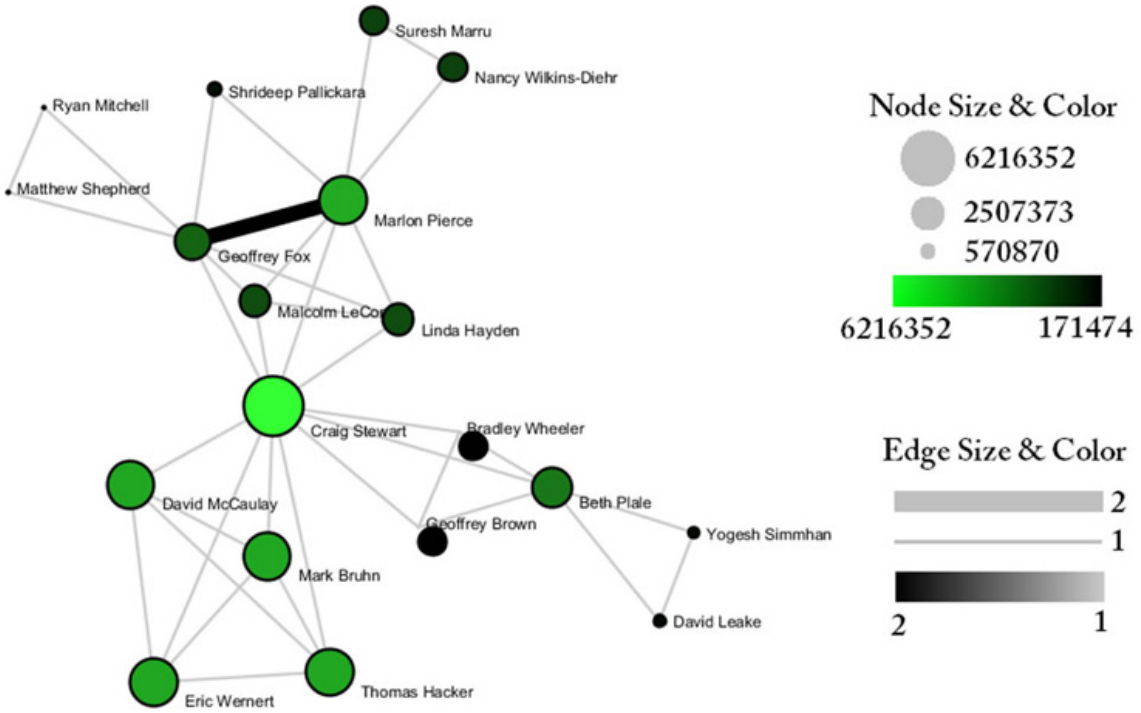


Figure 5.24: Largest component of Indiana University co-PI network. Node size and color display the total a ward amount.

To see the log file from this workflow save the 5.2.1 Funding Profiles of Three Universities (NSF Data) file.

5.2.1.1 Database Extractions

> **Extended Version**
>
> This workflow uses the extended version of the Sci^2 Tool. To know how to extend Sci^2 view Section 3.2 Additional Plugins.

The Sci^2 Tool supports the creation of databases for NSF files. Database loading improves the speed and functionality of data preparation and preprocessing. To load the Indiana NSF file as a database, go to *'File > Load >* and select ***yoursci2directory****/sampledata/scientometrics/nsf/Indiana.nsf.'* In the "Load" pop-up window, choose "NSF database." Cleaning should be performed before any other task using '*Data Preparation > Database > NSF >* Merge Identical NSF People'.
To view a breakdown of each investigator from Indiana, run '*Data Preparation > Database > NSF >* Extract Investigators'. View the table – notice that next to each investigator will be listed their total number of awards, total as the PI and as a Co-PI, the total amount awarded to date, as well as their earliest award start date and latest award expiration date.

To create Co-PI networks like those from the previous workflows, simply run '*Data Preparation > Databases > NSF >* Extract Co-PI Network' on the cleaned database. Delete the isolates by running *'Preprocessing > Networks >* Dele*te Isolates'*.

As before, to visualize the network, select *'Visualization > Networks > GUESS'* and run '*Layout > GEM'* followed by '*Layout > Bin Pack'*. Run the '***yoursci2directory****/scripts/GUESS/co-PI-nw_database.py'* script to apply the standard Co-PI network theme.

Figure 5.25: Indiana University Co-PI network using databases.

5.2.2 Mapping CTSA Centers (NIH RePORTER Data)

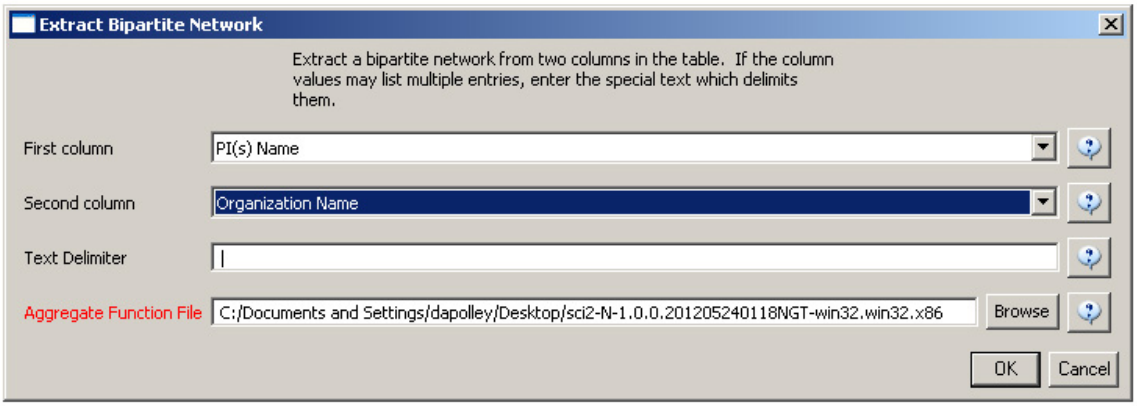
CTSA2005-2009.xls	
Time frame:	2005-2009
Region(s):	Miscellaneous
Topical Area(s):	Clinical and Translational Science
Analysis Type(s):	PI-Institution Network, Co-Authorship Network

Data doesn't always come in easily parsible formats. This study, a search for all recent grants to CTSA centers, requires some advanced searching and data manipulation to prepare the required data. The data comes from the union of NIH RePORTER downloads (see section 4.2.2.2 NIH RePORTER) and NIH ExPORTER data dumps (http://projectreporter.nih.gov/exporter/). Each CTSA Center grant was found and matched with its associated publications using a project-specific ID.

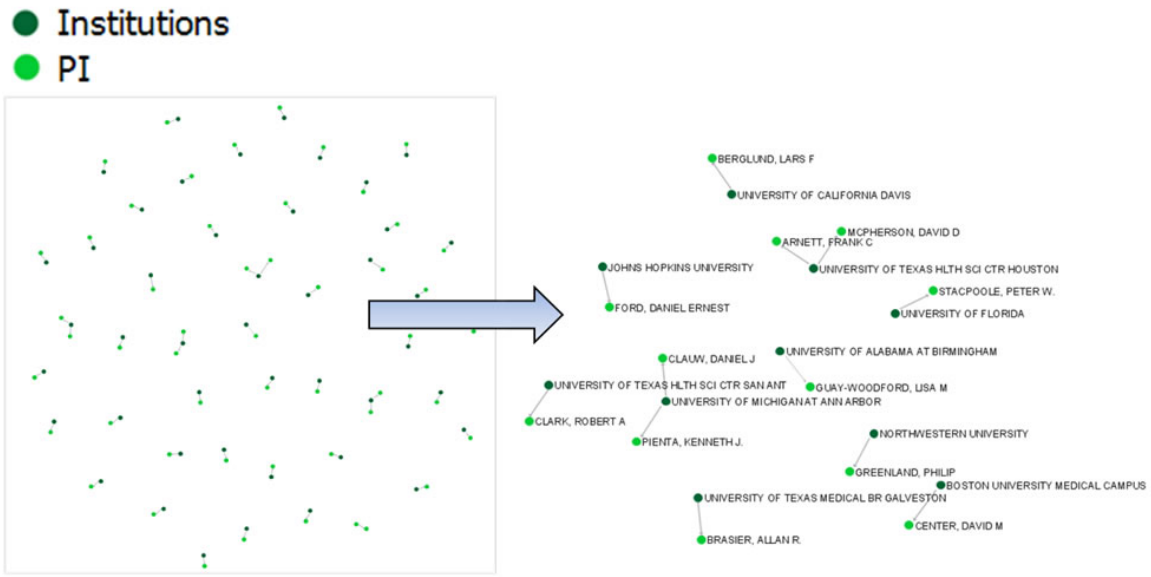
The resulting file, which contains all NIH Clinical and Translational Science Awards and their corresponding details

from 2005-2009, is saved in an Excel file in '***yoursci2directory***/*sampledata/scientometrics/nih/CTSA2005-2009*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). The file contains two spreadsheets, one with publications and one with grants. Save each spreadsheet out as *grants.csv* and *publications .csv*.

First load *grants.csv* in the Sci^2 Tool using '*File > Load*' and *'Standard csv format'* in the "Load" pop-up. To view a bimodal network visualizing which main PIs associate with which institution, run '*Data Preparation* > *Extract Bipartite Network* ' with the following parameters:

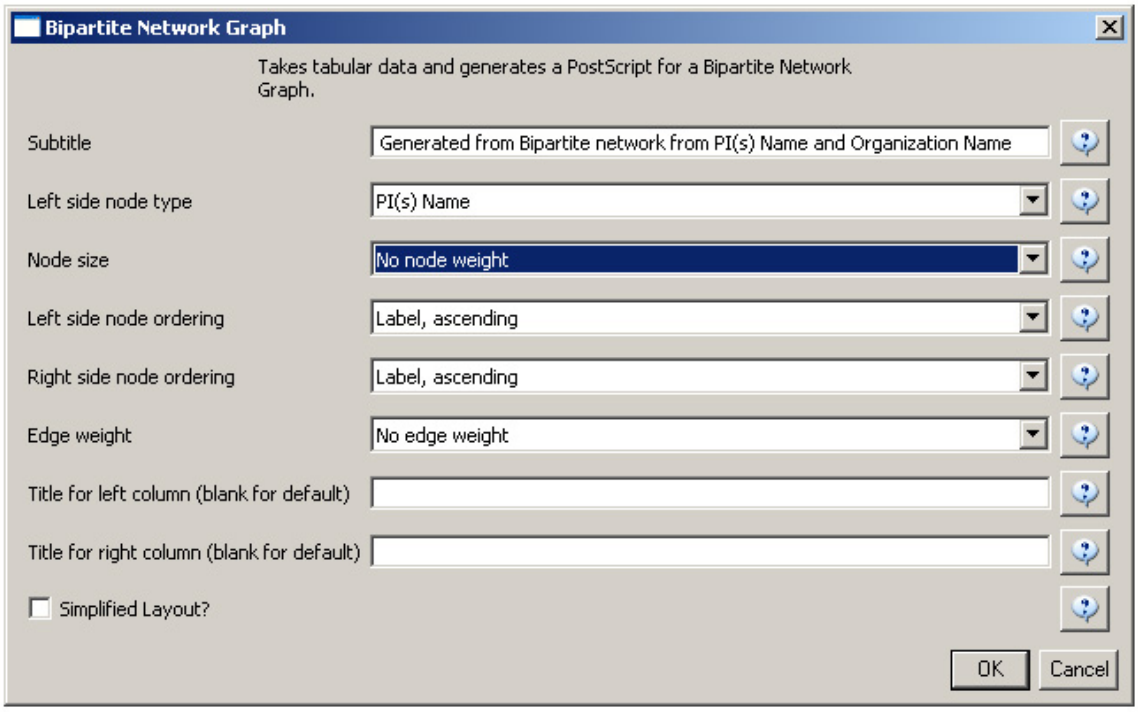


The resulting network can be visualized in GUESS and laid out using GEM, see Figure 5.26



.

The network can also be visualized with the Bipartite-specific visualization. Run '*Visualization* > *Networks* > *Bipartite Network Graph'* with the following parameters:



Sci2 with generate a visualization titled "Bipartite Network Graph PS" in the Data Manager. This network visualization can be saved as a PostScript file and the resulting visualization will look like this:

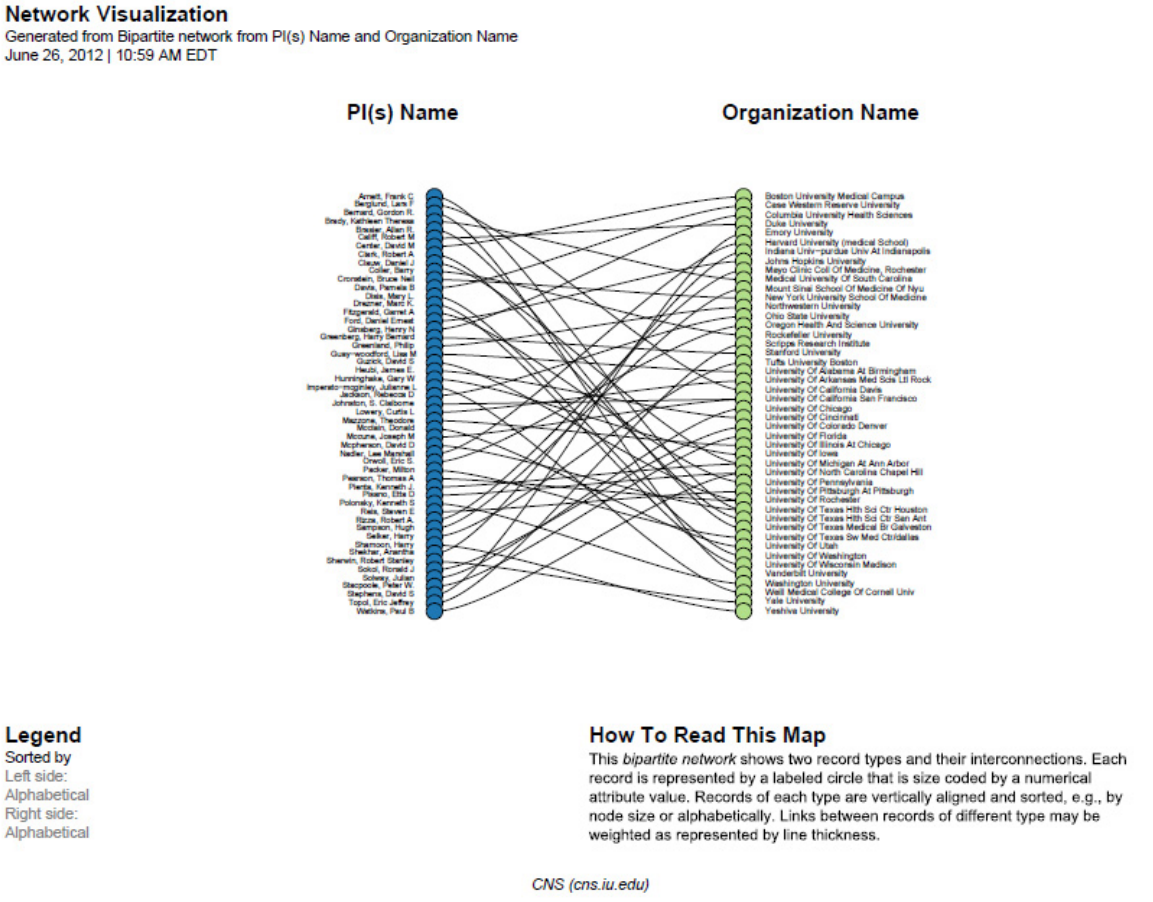


Figure 5.26: Bimodal institution-PI network for CTSA Centers.

Now load *publications.csv* as a standard csv and create a co-authorship network by running '*Data Preparation* > *Extract Co-Occurrence Network*' with text delimiter set to " **;** ". The parameter for Column Name should be set to "author." The resulting co-authorship network has 8,668 nodes, 26 isolates, and 50,129 edges (see Figures 5.27 and 5.28).

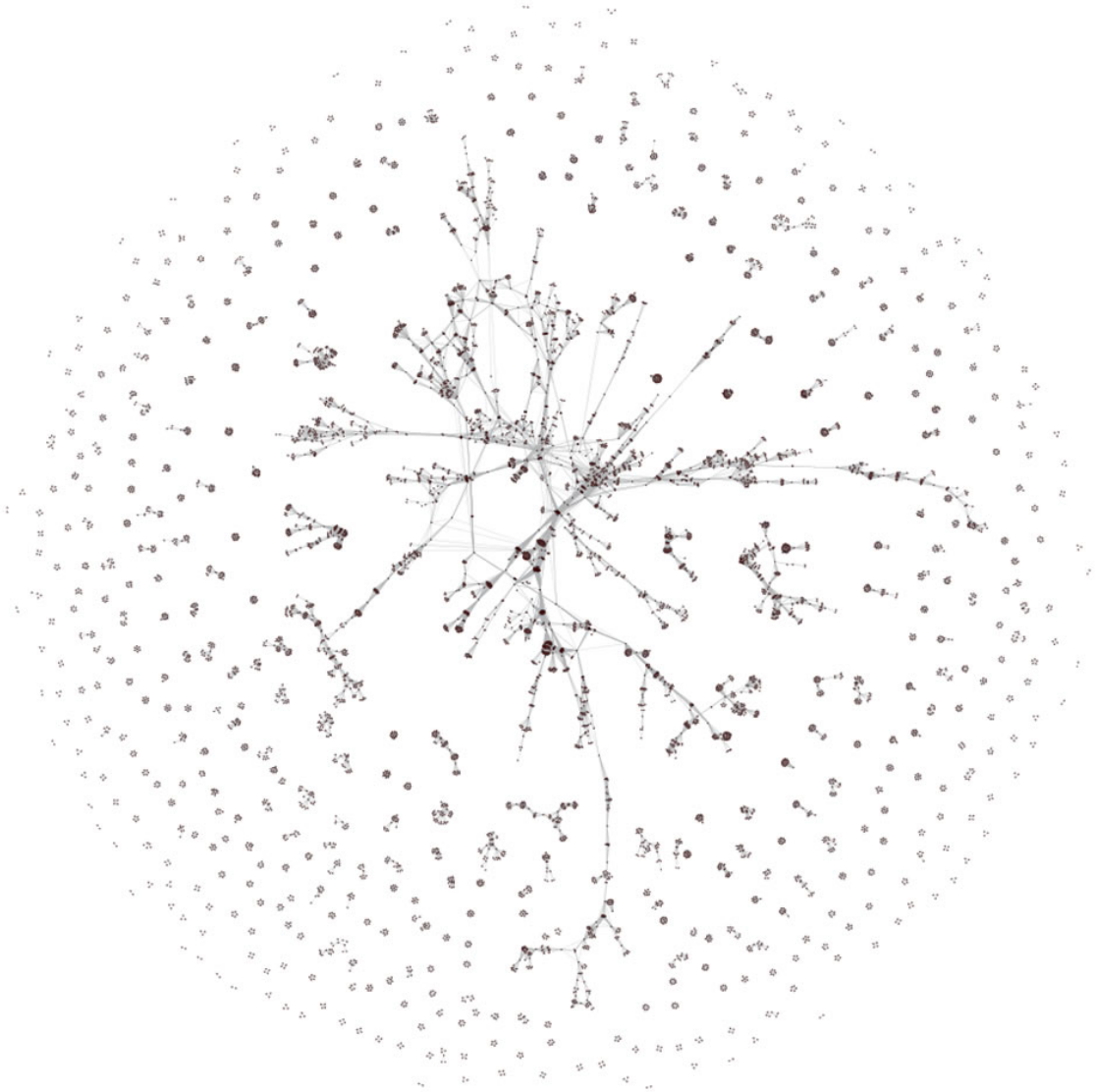


Figure 5.27: Co-authorship network of CTSA Center publications

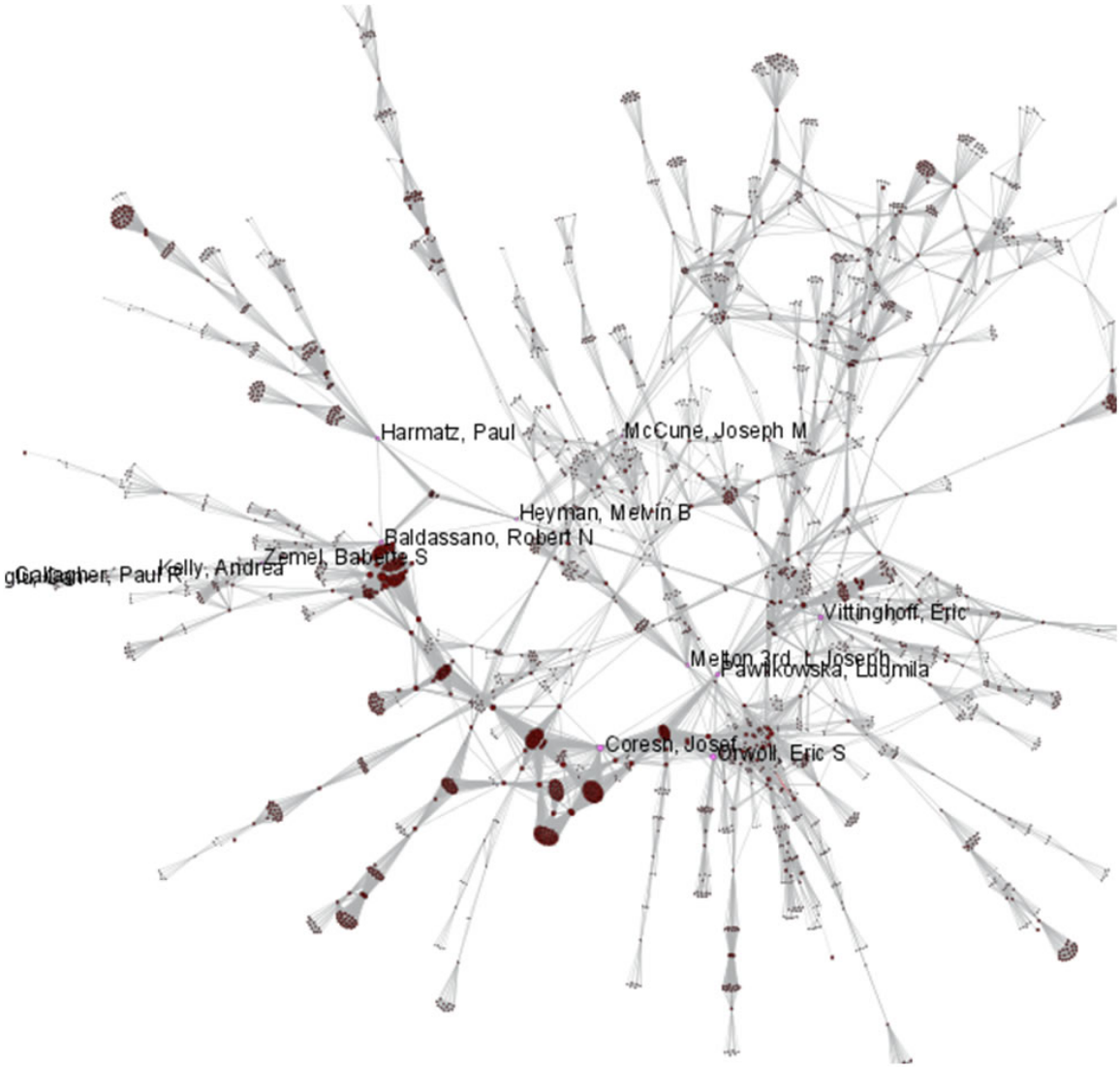


Figure 5.28: Largest connected component of CTSA Center publication co-authorship network

To see the log file from this workflow save the 5.2.2 Mapping CTSA Centers (NIH RePORTER Data) log file.

5.2.3 Biomedical Funding Profile of NSF (NSF Data)

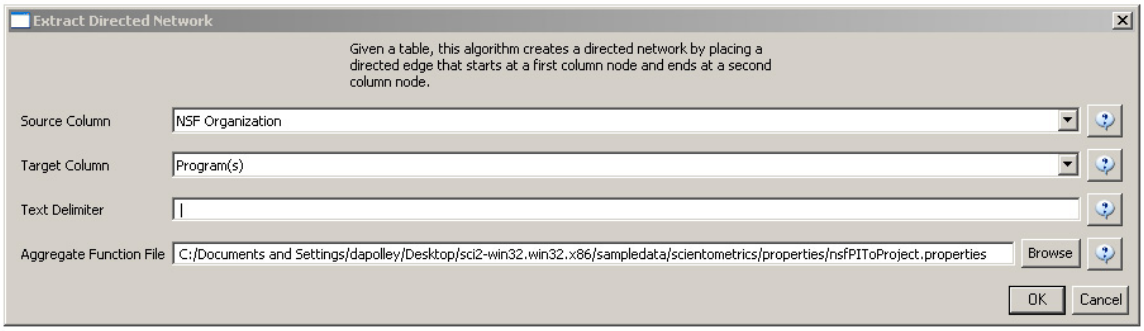
MedicalAndHealth.nsf	
Time frame:	2003-2010
Region(s):	Miscellaneous
Topical Area(s):	Biomedical
Analysis Type(s):	NSF Organization-Program Network

A NSF Organization-to-Program(s) bimodal network is exhibited here using data downloaded from the NSF Awards Search at (http://www.nsf.gov/awardsearch) on Nov 23th, 2009, using the query "medical AND health" in the title,

abstract, and awards field, with "Active awards only" checked. This query yielded 286 awards (see section 4.2.2.1 NSF Award Search for data retrieval details).

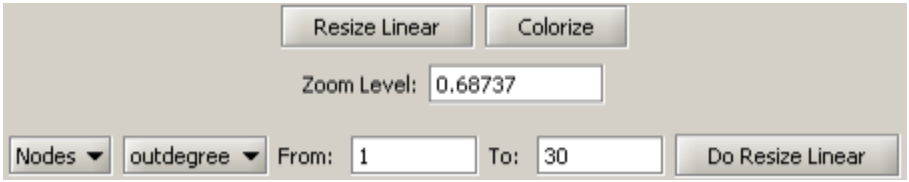
Load '***yoursci2directory***/*sampledata/scientometrics/nsf/MedicalAndHealth.nsf*' in NSF csv format (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Then run *'Data Preparation > Extract Directed Network'* with parameters:

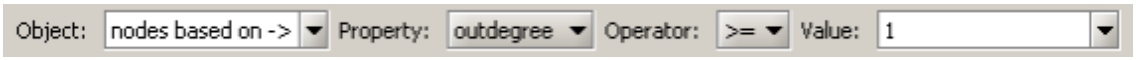
Make sure to select the nsfPIToProject.properties for the aggregate function file.



Select "Network with directed edges from NSF Organization to Program(s)" in the Data Manager and run *'Analysis > Networks > Unweighted and Directed > Node Indgree*. Then select "Network with indegree attribute added to node list" in the Data Manager and run '*Analysis > Networks > Unweighted and Directed > Node Outdegree'*. Select the resulting "Network with outdegree attribute added to node list" and run *'Visualization > Networks > GUESS'* followed by *' Layout > GEM'* and *'Layout > Bin Pack'*.

In graph modifier interface:





Select "Colour" from the options directly below and choose desired color.

Then select "Show Label".

Select "Colour" and choose desired color.

The resulting network is visualized in Figure 5.29.

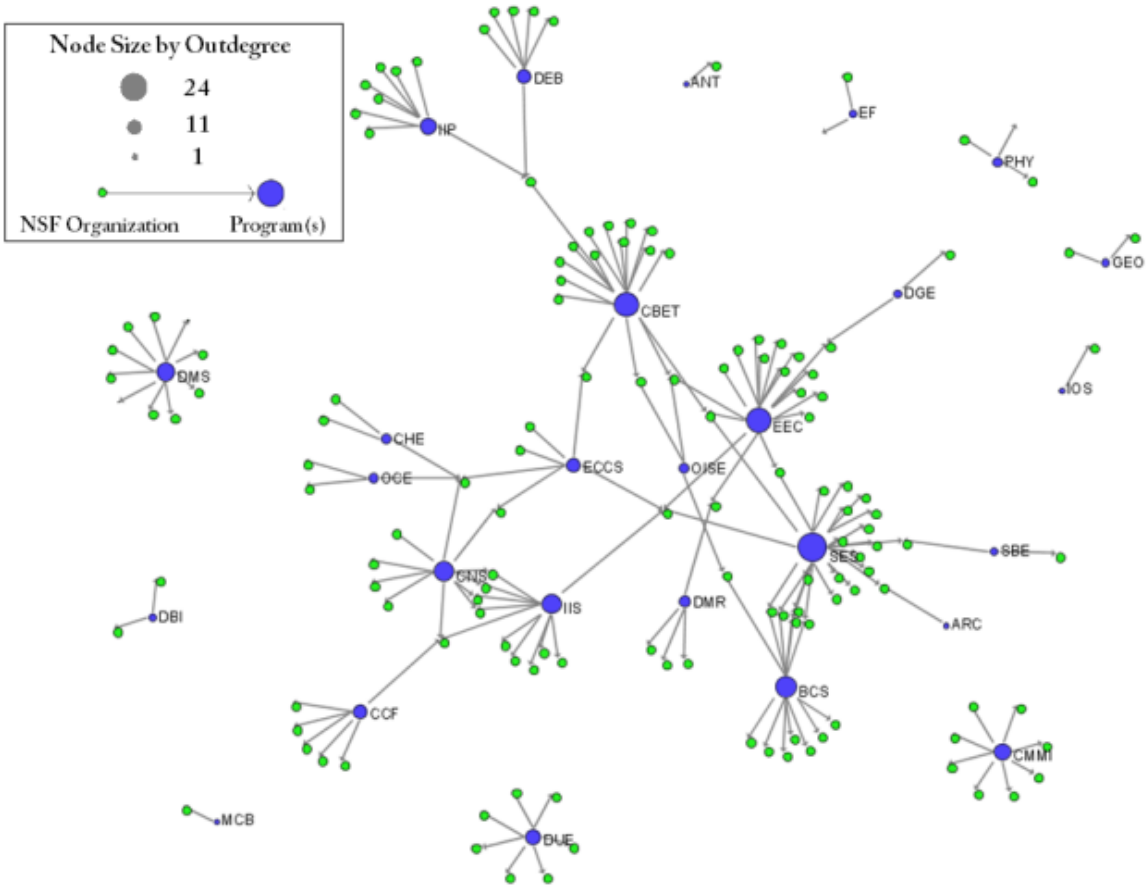


Figure 5.29: Bimodal Network of NSF Organization to Program(s) of NSF Medical + Health Funding

To see the log file from this workflow save the 5.2.3 Biomedical Funding Profile of NSF (NSF Data) log file.

5.2.4 Mapping Scientometrics (ISI Data)

5.2.4.1 Document Co-Citation

Scientometrics.isi	
Time frame:	1978-2008
Region(s):	Miscellaneous
Topical Area(s):	Scientometrics
Analysis Type(s):	Document Co-Citation Network

Scientometrics is a discipline which uses statistical and computational techniques in order to understand the structure and dynamics of science. Here we use ISI data from the journal "Scientometrics" and Science of Science and Innovation Policy (SciSIP) data from NSF Awards Search.

Download Scientometrics.isi. Load the file using '*File > Load*' and locating the downloaded file. This domain level dataset is ideal for document co-citation analysis, as the scale is large enough that the resulting network will paint a fairly accurate picture of document similarity within the domain of scientometrics.

New ISI File Format

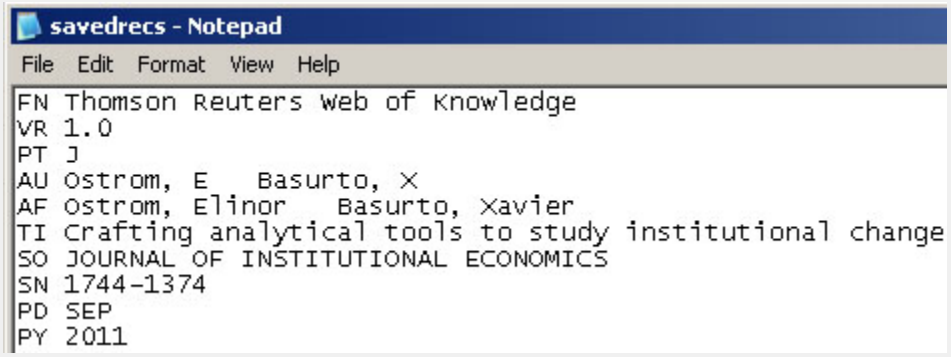
Web of Science made a change to their output format in September, 2011. Older versions of Sci2 tool (older than v0.5.2 alpha) may refuse to load these new files, with an error like "Invalid ISI format file selected."

Sci2 solution

If you are using an older version of the Sci2 tool, you can download the WOS-plugins.zip file and unzip the JAR files into your sci2/plugins/ directory. Restart Sci2 to activate the fixes. You can now load the downloaded ISI files into the Sci2 without any additional step. If you are using the old Sci2 tool you will need to follow the guidelines below before you can load the new WOS format file into the tool.

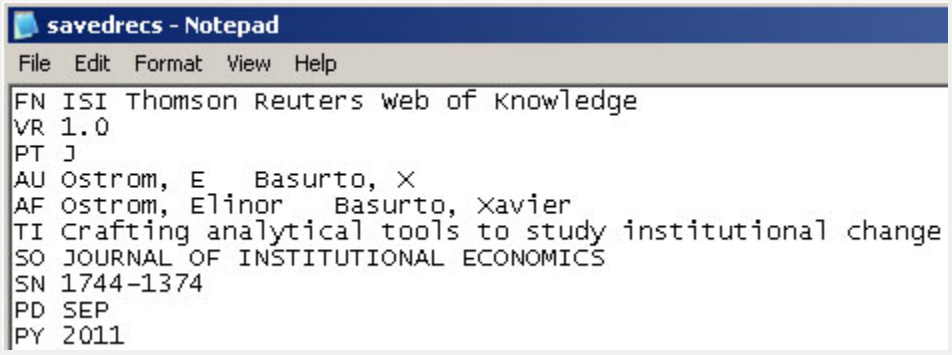
You can fix this problem for individual files by opening them in Notepad (or your favorite text editor). The file will start with the words:

Original file:



Just add the word ISI.

Updated file:



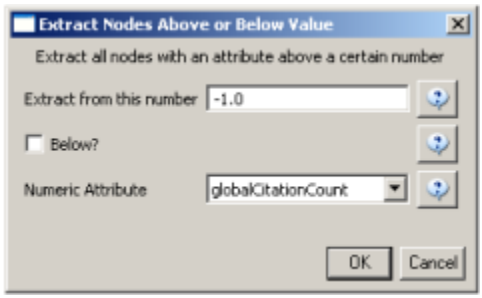
And then Save the file.

The ISI file should now load properly. More information on the ISI file format is available here (http://wiki.cns.iu.edu/display/CISHELL/ISI+%28*.isi%29).

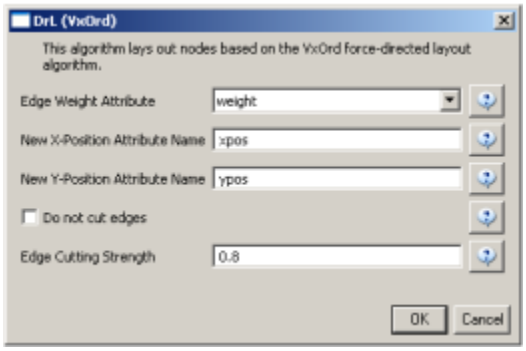
Select the dataset "2126 Unique ISI Records" in the Data Manager window and run *'Data Preparation* > *Extract Paper Citation Network*'.

Two files will appear in the Data Manager window: the paper-citation network and the paper information table.

Select the "Extracted paper-citation network" and run *'Preprocessing > Networks >* Extract Nodes Above or Below Value*'* with the following parameters:



The produced network contains only the original ISI records. Select the resulting file and run *'Data Preparation >* Extract Document Co-Citation Network*'*. Then, select the network and run *'Analysis > Networks >* Network Analysis Toolkit (NAT)*'*. There are 2056 nodes, 26070 edges and 775 isolates in the network. Run *'Preprocessing >* Delete Isolates*'* to remove all the isolates. Because this network is too dense to lay out in GUESS, run '*Visualization >* DrL (VxOrd)*'* with the parameters:



Next, select "Laid out with DrL" in the Data Manager and run '*Visualization > Network >* GUESS'. Run the following commands in the GUESS "Interpreter":

```
>for n in g.nodes:
    if n.xpos is not None and n.ypos is not None:
        n.x = n.xpos * 10
        n.y = n.ypos * 10
>resizeLinear(localcitationcount,1,50)
>colorize(localcitationcount, gray, black)
>resizeLinear(weight, .25, 8)
>colorize(weight, "127,193,65,255", black)
```

Go to "Graph Modifier" and choose *'Object: nodes based on -> > Property: localcitationcount > Operators: >= > Value: 20 > Show Label'*. See Figure 5.21.

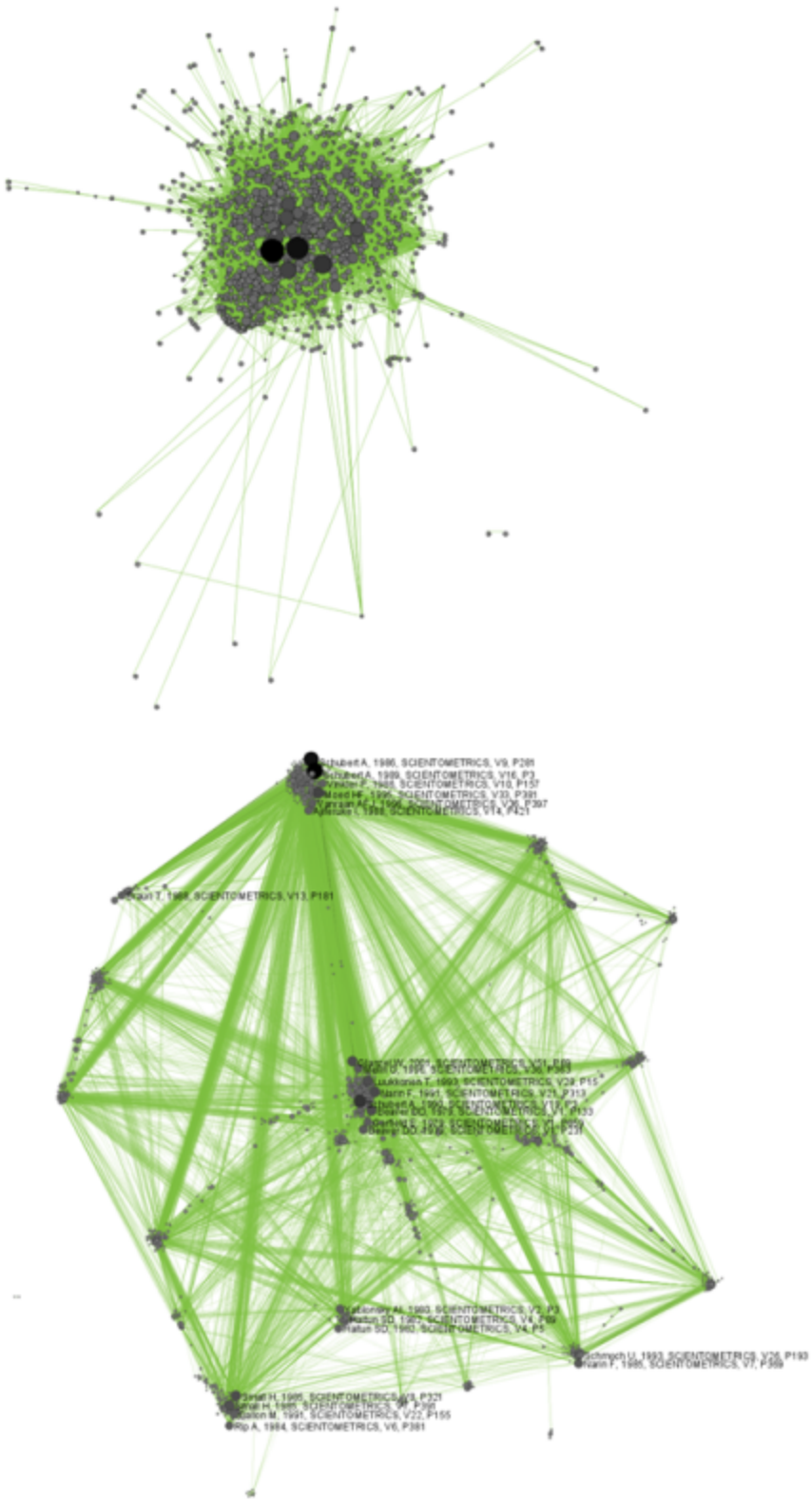


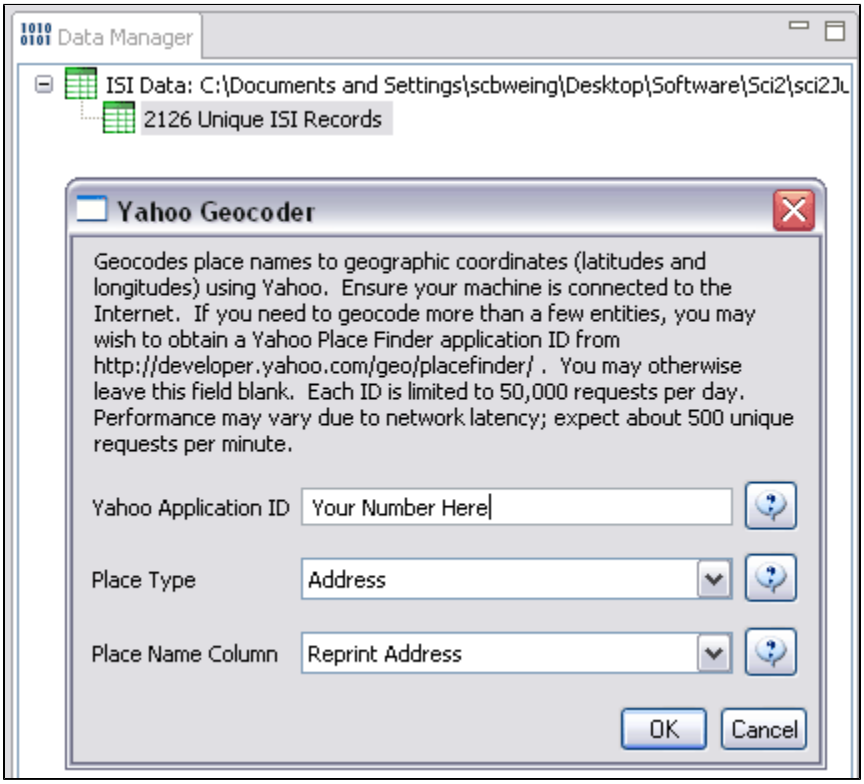
Figure 5.30: Document co-citation network for Scientometric.isi in GUESS without DrL edge cutting (top) and with DrL (VxOrd) (bottom).

To see the log file from this workflow save the 5.2.4.1 Document Co-Citation log file.

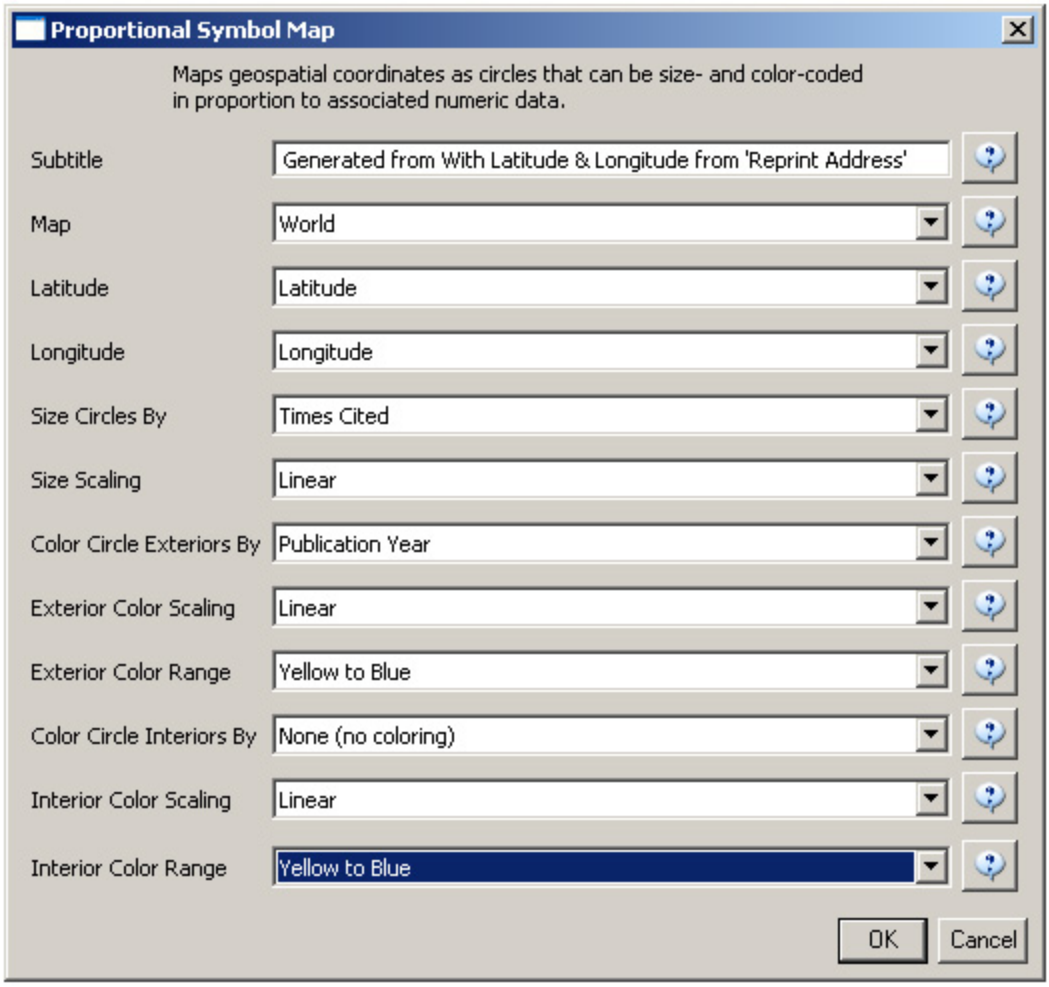
5.2.4.2 Geographic Visualization

Scientometrics.isi	
Time frame:	1978-2008
Region(s):	Miscellaneous
Topical Area(s):	Scientometrics
Analysis Type(s):	Geospatial Analysis

Using the dataset loaded from section 5.2.4.1, select '*2126 Unique ISI Records*' in the Data Manager. To find the latitude and longitudes of the locations of researches publishing in *Scientometrics*, run *Analysis* > *Geospatial* > *Yahoo Geocoder*. Note that you will need a Yahoo Application ID in order to run this query. Enter your Yahoo Application ID (sign up for one here), "Address" for the Place Type, and the "Reprint Address" for the Place Column Name. Press 'OK'. This step may take several minutes to complete.



A new table will appear in the Data Manager labeled '*With Latitude & Longitude from 'Reprint Address''* containing all data from the initial table, plus columns with the Yahoo Geocoded Latitude and Longitude Coordinates. Select this file and run *Visualization* > *Geospatial* > *Proportional Symbol Map* using the parameters listed below.



Save and view the resulting visualization using the directions described at 2.4 Saving Visualizations for Publication. This will be the result:

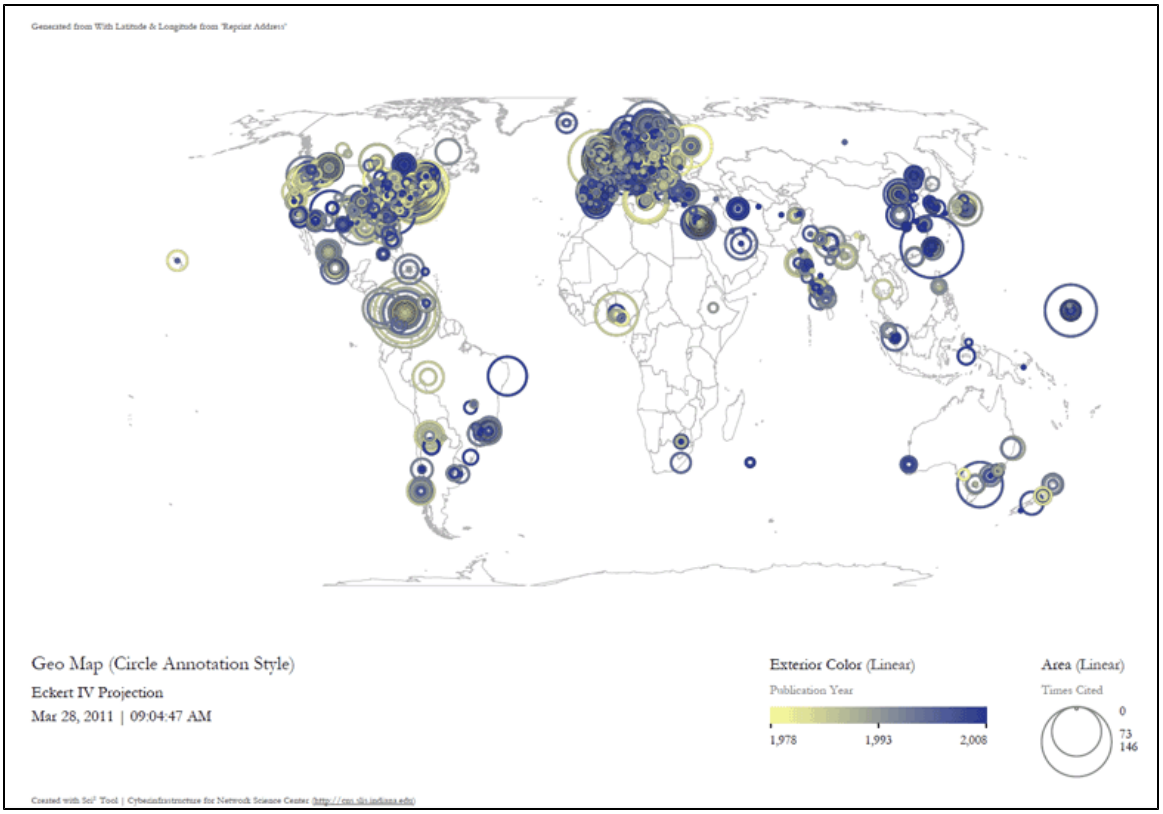


Figure 5.31: Scientometrics.isi geospatial visualization. Circle sizes represent times cited, colors represent publication year.

5.2.5 Burst Detection in Physics and Complex Networks (ISI Data)

AlessandroVespignani.isi	
Time frame:	1990-2006
Region(s):	Indiana University, University of Rome, Yale University, Leiden University, International Center for Theoretical Physics, University of Paris-Sud
Topical Area(s):	Informatics, Complex Network Science and System Research, Physics, Statistics, Epidemics
Analysis Type(s):	Burst Detection

Burst Detection

A scholarly dataset can be understood as a discrete time series: in other words, a sequence of events/ observations which are ordered in one dimension – time. Observations exist for regularly spaced intervals, e.g., each month or year.

The burst detection algorithm (see Section 4.6.1 Burst Detection) identifies sudden increases or "bursts" in the frequency-of-use of character strings over time. This algorithm identifies topics, terms, or concepts important to the events being studied that increased in usage, were more active for a period of time, and then faded away.

An analysis of publications authored or co-authored by Alessandro Vespignani from 1990 to 2006 will be used to illustrate the "burst" concept. Alessandro Vespignani is an Italian physicist and Professor of Informatics and Cognitive Science at Indiana University, Bloomington. In his publications, it is possible to see a change in research focus - from Physics to Complex Networks - beginning in 2001.

Load Alessandro Vespignani's ISI publication history using '*File* > *Load*' and following this path: '***yoursci2directory****/sampledata/scientometrics/isi/AlessandroVespignani.isi*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets).

New ISI File Format

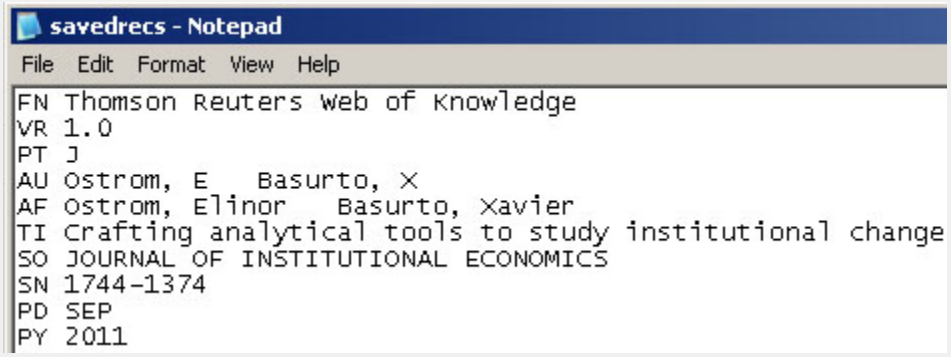
Web of Science made a change to their output format in September, 2011. Older versions of Sci2 tool (older than v0.5.2 alpha) may refuse to load these new files, with an error like "Invalid ISI format file selected."

Sci2 solution

If you are using an older version of the Sci2 tool, you can download the WOS-plugins.zip file and unzip the JAR files into your sci2/plugins/ directory. Restart Sci2 to activate the fixes. You can now load the downloaded ISI files into the Sci2 without any additional step. If you are using the old Sci2 tool you will need to follow the guidelines below before you can load the new WOS format file into the tool.

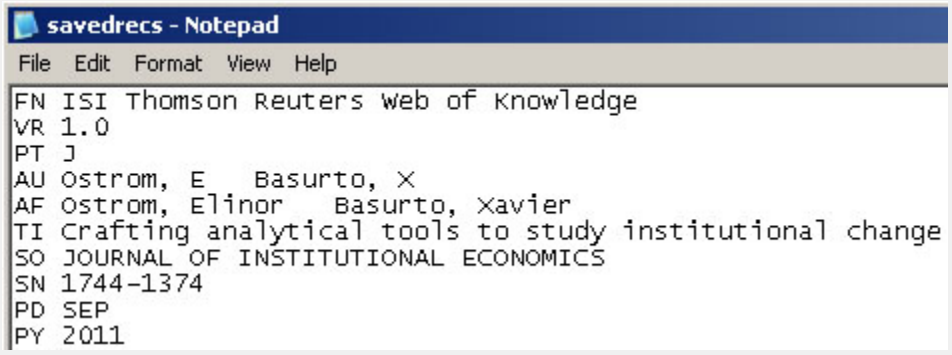
You can fix this problem for individual files by opening them in Notepad (or your favorite text editor). The file will start with the words:

Original file:



Just add the word ISI.

Original file:

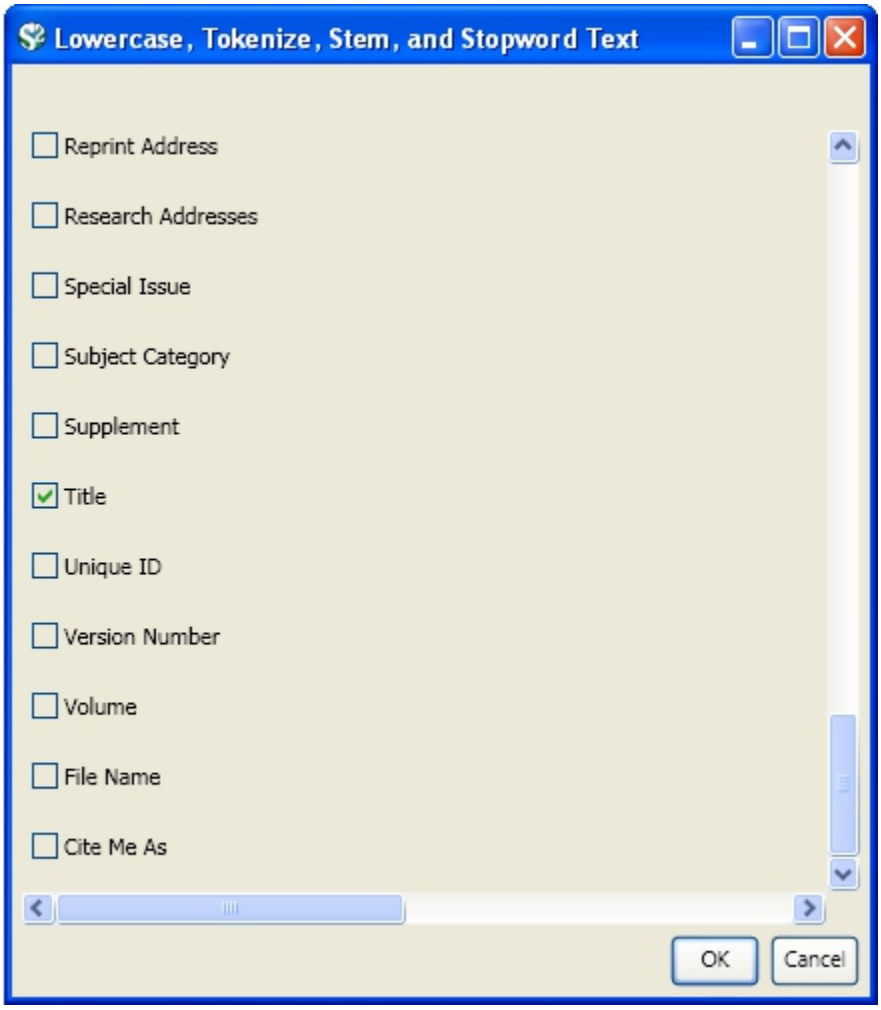


And then Save the file.

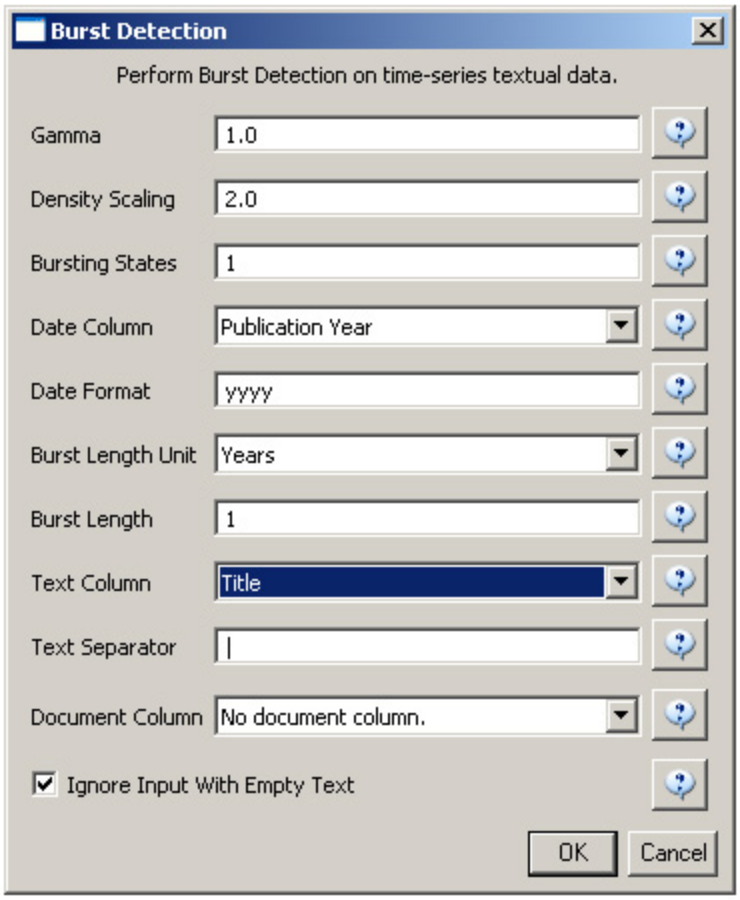
The ISI file should now load properly. More information on the ISI file format is available here (http://wiki.cns.iu.edu/display/CISHELL/ISI+%28*.isi%29).

This analysis will detect the "bursty" terms used in the title of papers in the dataset. Since the burst detection algorithm is case-sensitive, it is necessary to normalize the field to be analyzed before running the algorithm. Select the table "101 Unique ISI Records" and run *'Preprocessing > Topical >* *Lowercase, Tokenize, Stem, and Stopword*

Text.' Check the "Title" box to indicate that you want to normalize this field:



Select the resulting "with normalized Title" table in the Data Manager and run '*Analysis* > *Topical* > *Burst Detection*' with the following parameters:



The "Gamma" parameter is the value that state transition costs are proportional to. This parameter is used to control how ease the automaton can change states. The higher the "Gamma" value, the smaller the list of bursts generated.

The "Density Scaling" parameter determines how much 'more bursty' each level is beyond the previous one. The higher the scaling value, the more active (bursty) the event happens in each level.

The "Bursting States" parameter determines how many bursting states there will be, beyond the non-bursting state. An *i* value of bursting states is equals to *i + 1* automaton states.

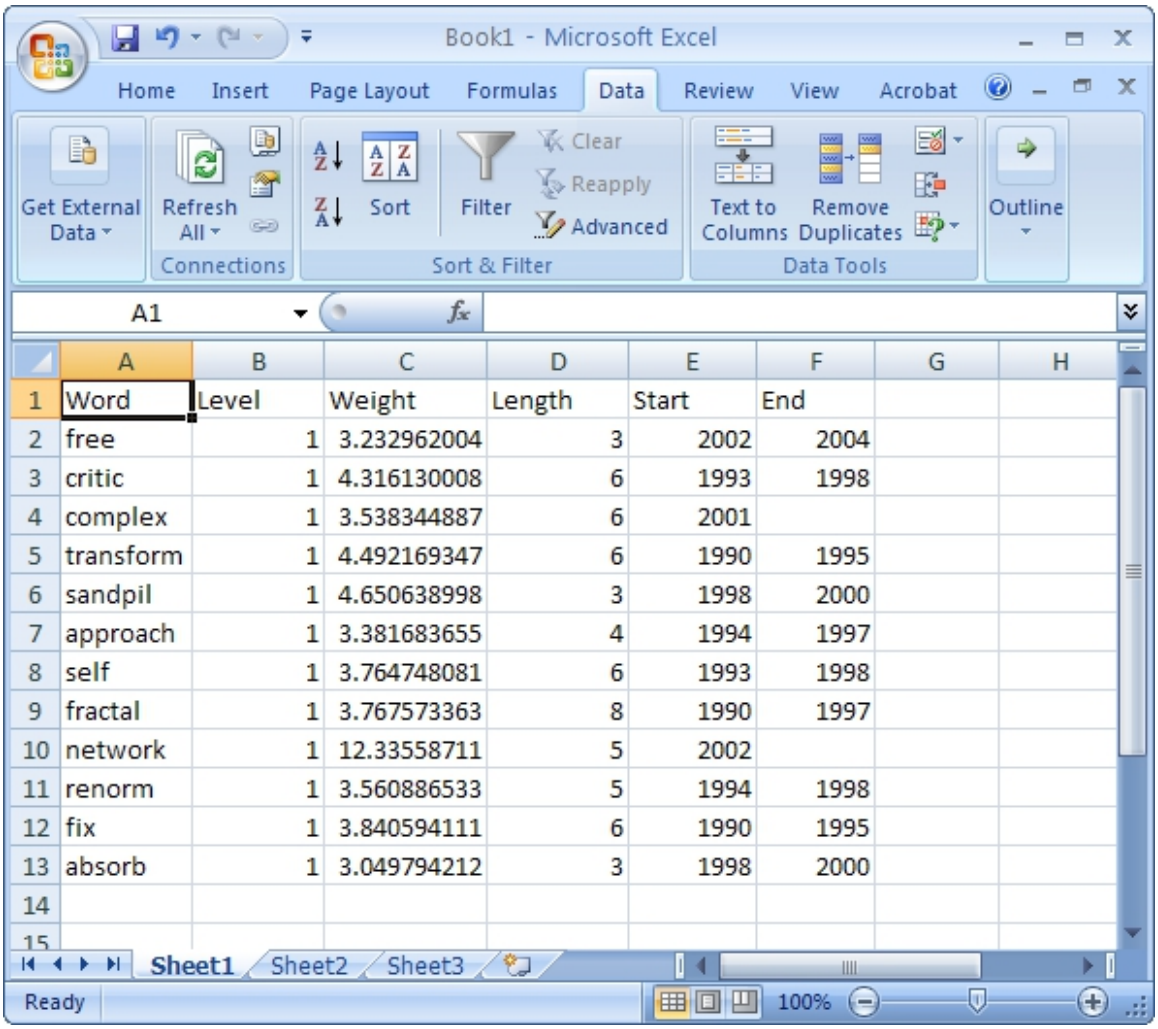
The "Date Column" parameter is the name of the column with date/time when the events / topics happens.

The "Date Format" specifies how the date column will be interpreted as a date/time. See http://java.sun.com/j2se/1.4.2/docs/api/java/text/SimpleDateFormat.html for details.

The "Text Column" parameter is the name of the column with values (delimiter and tokens) to be computed for bursting results.

The "Text Separator" parameters determines the separator that was used to delimit the tokens in the text column.

View the file "Burst detection analysis (Publication Year, Title): maximum burst level 1". On a PC running Windows, right click on this table and select view to see the data in Excel. On a Mac or a Linux system, right click and save the file, then open using the spreadsheet program of your choice.



In this table, there are six columns: "Word," "Level," "Weight," "Length," "Start," and "End."

The "Word" field identifies the specific character string which was detected as a "burst." The "Length" field indicates how long the burst lasted (over the selected time parameter).

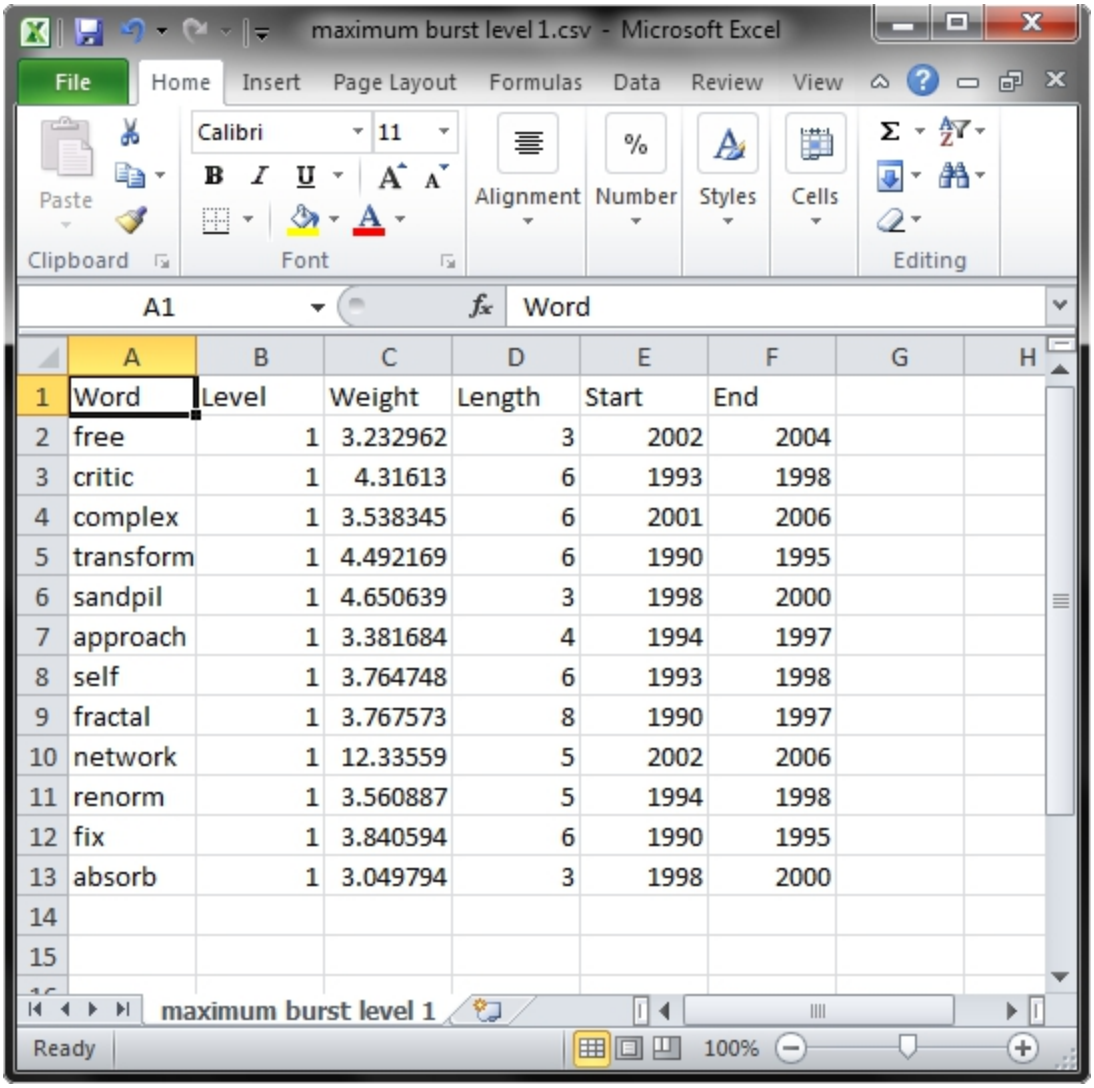
The "Level" is the burst level of this burst. The higher burst level, the more frequent the event / topic happens.

The "Weight" field is the weight of this burst between its "Length". A higher weight could be resulted by the longer "Length", the higher "Level" or both.

The "Length" is the period of the burst. It is generated based on *(Start - End + 1)*.

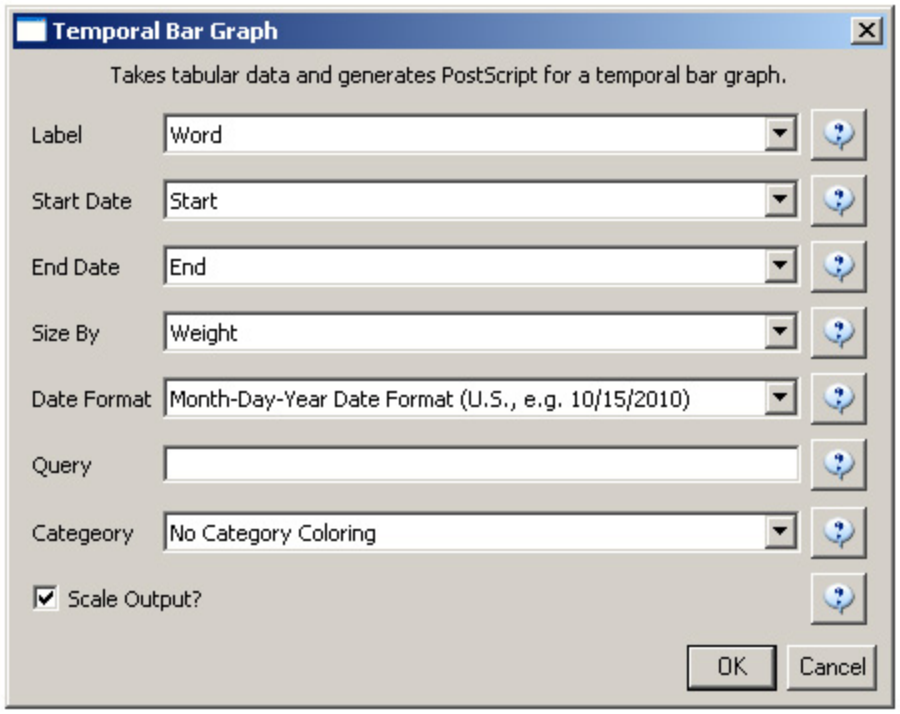
The "Start" field identifies when the burst began (again, according to the specified time parameter).

And the "End" field indicates when the burst stopped. A empty value in the "End" field indicates that the burst lasted until the last date present in the dataset. Where the "End" field is empty, put manually the last year present in the dataset. In this case, 2006.



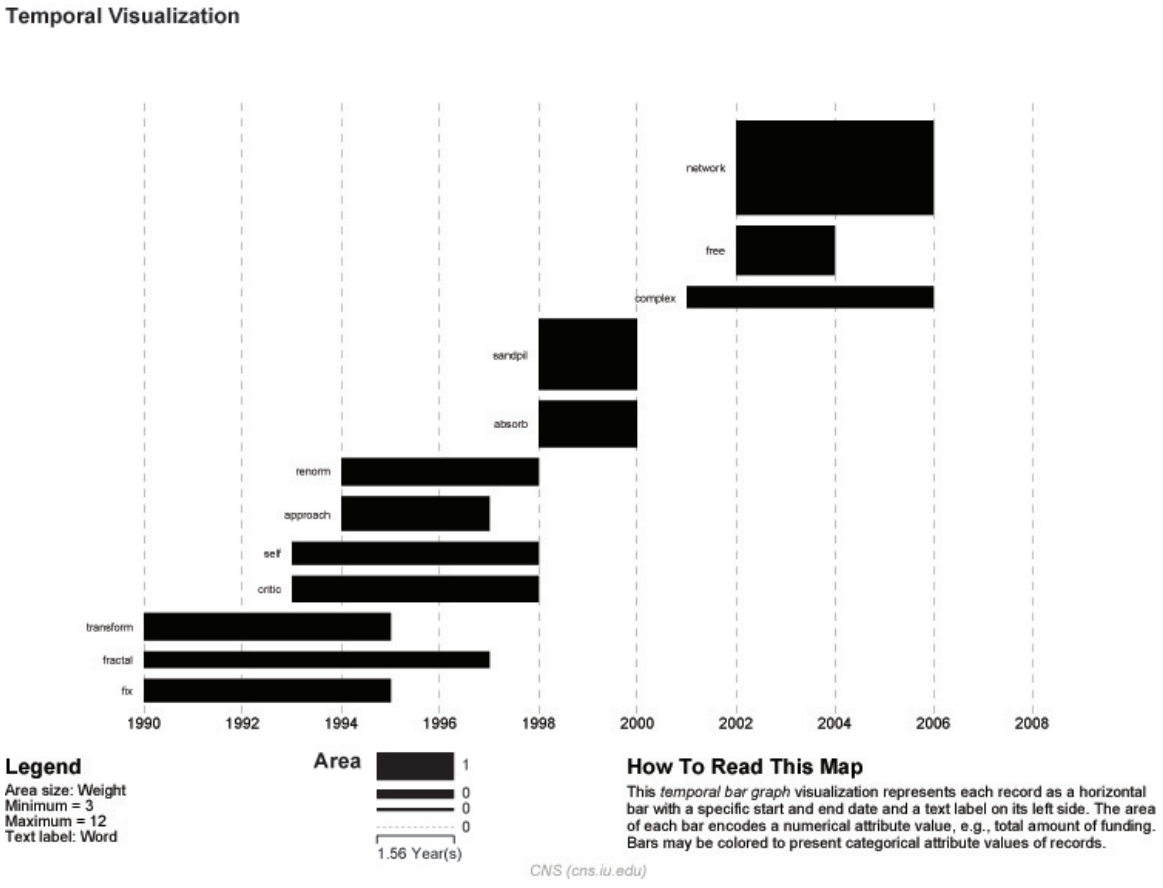
	A	B	C	D	E	F	G	H
1	Word	Level	Weight	Length	Start	End		
2	free	1	3.232962	3	2002	2004		
3	critic	1	4.31613	6	1993	1998		
4	complex	1	3.538345	6	2001	2006		
5	transform	1	4.492169	6	1990	1995		
6	sandpil	1	4.650639	3	1998	2000		
7	approach	1	3.381684	4	1994	1997		
8	self	1	3.764748	6	1993	1998		
9	fractal	1	3.767573	8	1990	1997		
10	network	1	12.33559	5	2002	2006		
11	renorm	1	3.560887	5	1994	1998		
12	fix	1	3.840594	6	1990	1995		
13	absorb	1	3.049794	3	1998	2000		
14								
15								

After you add manually this information, save this .csv file somewhere in your computer. Load back this .csv file into Sci2 using '*File* > *Load*'. Select 'Standart csv format' int the pop-up window. A new table will appear in the Data Manager. To visualize these table that contains the results of the Burst Detection algorithm, select the table you just loaded in the Data Manager and run '*Visualization* > *Temporal* > *Temporal Bar Graph*' with the following parameters:



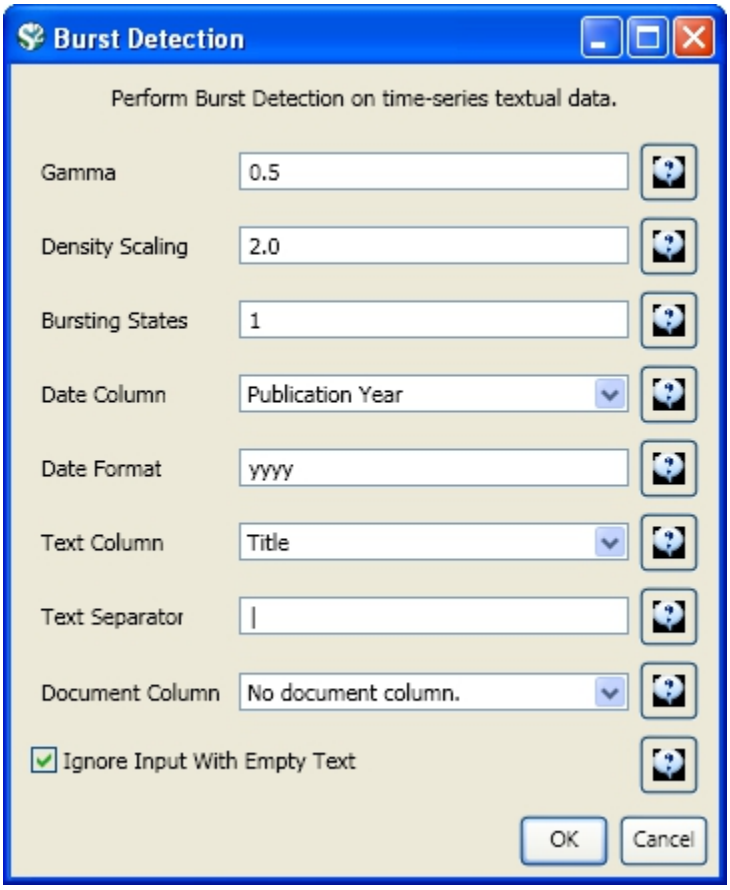
Horizontal bar graphs are used to visualize numeric data over time, generating labeled horizontal bars. A PostScript file containing the horizontal bar graph will appear in the Data Manager.

Open and view the file using the workflow from Section 2.4 Saving Visualizations for Publication.



The resulting analysis indicates a change in the research focus of Alessandro Vespignani for publications beginning in 2001. For example, the bursting terms "fractal," "growth," "transform," and "fix" starting at 1990 are related to Vespignani's Ph.D., entitled "Fractal Growth and Self-Organized Criticality" in Physics. Other bursts also related to Physics follow these, such as "sandipil." After 2001, bursting terms like "complex," "network," "free," and "weight" appear, signifying a change in Vespignani's research area from Physics to Complex Networks, with a larger number of publications on topics like "weighted networks" and "scale-free networks."

Now, let's run the Burst Detection algorithm again for the same dataset but for a different value for the 'Gamma' parameter. Select the table 'with normalized Title' in the Data Manager and run '*Analysis* > *Topical* > *Burst Detection* ' with the following parameters:



Notice that the value for the gamma parameter is now set to 0.5. The parameter gamma controls the ease with which the automaton can change states. With a smaller gamma value, more bursts will be generated. Running the algorithm with these parameters will generate a new table named "Burst detection analysis (Publication Year, Title): maximum burst level 1.2" in the Data Manager.

Microsoft Excel - maximum burst level 1.2.csv

	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	free	1	3.232962	3	2002	2004
3	epidem	1	2.532198	6	2001	
4	disloc	1	2.267251	2	2001	2002
5	system	1	1.45763	4	1996	1999
6	critic	1	4.31613	6	1993	1998
7	complex	1	3.538345	6	2001	
8	transform	1	4.492169	6	1990	1995
9	conserv	1	1.972471	3	1998	2000
10	field	1	1.834723	4	1997	2000
11	fractur	1	1.593632	6	1994	1999
12	dynam	1	2.404396	2	2001	2002
13	forest	1	1.809792	3	1995	1997
14	phase	1	2.446386	1	2000	2000
15	weight	1	2.716894	3	2004	
16	aggreg	1	2.50294	5	1991	1995
17	sandpil	1	4.650639	3	1998	2000
18	approach	1	3.381684	4	1994	1997
19	growth	1	2.839977	8	1990	1997
20	limit	1	1.45763	5	1991	1995
21	self	1	3.764748	6	1993	1998
22	similar	1	1.45763	5	1991	1995
23	fractal	1	3.767573	8	1990	1997
24	transit	1	2.783221	4	1997	2000
25	global	1	1.719802	4	2003	
26	scale	1	2.66792	4	1990	1993
27	state	1	1.875465	6	1996	2001
28	network	1	12.33559	5	2002	
29	evolut	1	1.804025	4	2003	
30	renorm	1	3.560887	5	1994	1998
31	fix	1	3.840594	6	1990	1995
32	absorb	1	3.049794	3	1998	2000
33	organ	1	2.751329	5	1994	1998
34	group	1	2.201667	5	1995	1999

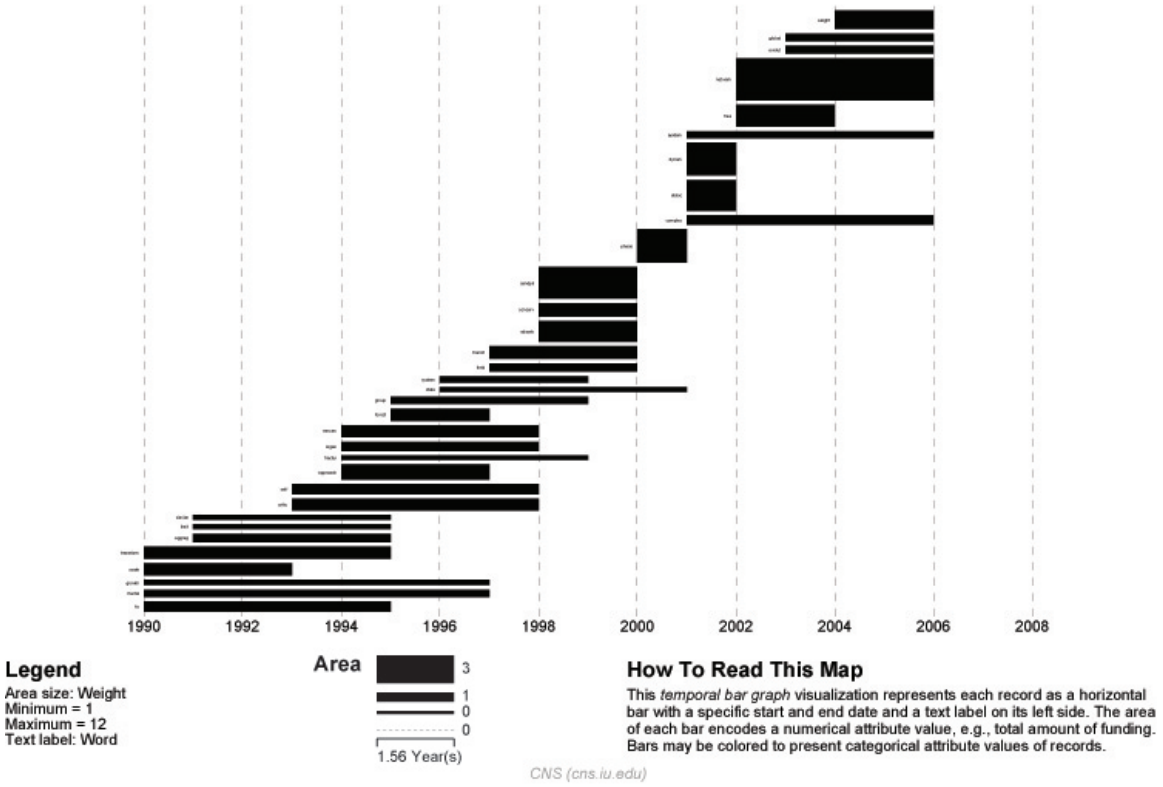
Again where the "End" field is empty, put manually the last year present in the dataset. In this case, 2006.

Microsoft Excel - maximum burst level 1.2.csv

	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	free	1	3.232962	3	2002	2004
3	epidem	1	2.532198	6	2001	2006
4	disloc	1	2.267251	2	2001	2002
5	system	1	1.45763	4	1996	1999
6	critic	1	4.31613	6	1993	1998
7	complex	1	3.538345	6	2001	2006
8	transform	1	4.492169	6	1990	1995
9	conserv	1	1.972471	3	1998	2000
10	field	1	1.834723	4	1997	2000
11	fractur	1	1.593632	6	1994	1999
12	dynam	1	2.404396	2	2001	2002
13	forest	1	1.809792	3	1995	1997
14	phase	1	2.446386	1	2000	2000
15	weight	1	2.716894	3	2004	2006
16	aggreg	1	2.50294	5	1991	1995
17	sandpil	1	4.650639	3	1998	2000
18	approach	1	3.381684	4	1994	1997
19	growth	1	2.839977	8	1990	1997
20	limit	1	1.45763	5	1991	1995
21	self	1	3.764748	6	1993	1998
22	similar	1	1.45763	5	1991	1995
23	fractal	1	3.767573	8	1990	1997
24	transit	1	2.783221	4	1997	2000
25	global	1	1.719802	4	2003	2006
26	scale	1	2.66792	4	1990	1993
27	state	1	1.875465	6	1996	2001
28	network	1	12.33559	5	2002	2006
29	evolut	1	1.804025	4	2003	2006
30	renorm	1	3.560887	5	1994	1998
31	fix	1	3.840594	6	1990	1995
32	absorb	1	3.049794	3	1998	2000
33	organ	1	2.751329	5	1994	1998
34	group	1	2.201667	5	1995	1999

After you add manually this information, save this .csv file somewhere in your computer. Load back this .csv file into Sci2 using '*File* > *Load*'. Select 'Standart csv format' int the pop-up window. A new table will appear in the Data Manager. To visualize these table that contains these new results for the Burst Detection algorithm, select the table you just loaded in the Data Manager and run '*Visualization* > *Temporal* > Horizontal Bar Graph (not included version)' with the same parameters.

A new PostScript file containing the horizontal bar graph will appear in the Data Manager. Once more, open and view the file using the workflow from Section 2.4 Saving Visualizations for Publication.



As expected, a larger number of bursts appear, and the new bursts have a smaller weight that those depicted in the first graph. These smaller, more numerous bursting terms permit a more detailed view of the dataset and allow the identification of trends. The "protein" burst starting in 2003, for example, indicates the year in which Alessandro Vespignani started to work with "protein-protein interaction networks," while the burst "epidem" - also from 2001 - is related to the application of complex networks to the analysis of epidemic phenomena in biological networks.

Visualizing Burst Detection in Excel

Its possible to generate a visualization for burst analysis in MS Excel. For this, open the results of the first burst analysis conducted ('Burst detection analysis (Publication Year, Title): maximum burst level 1') in MS Excel, by right clicking on this table in the Data Manager and selecting View.

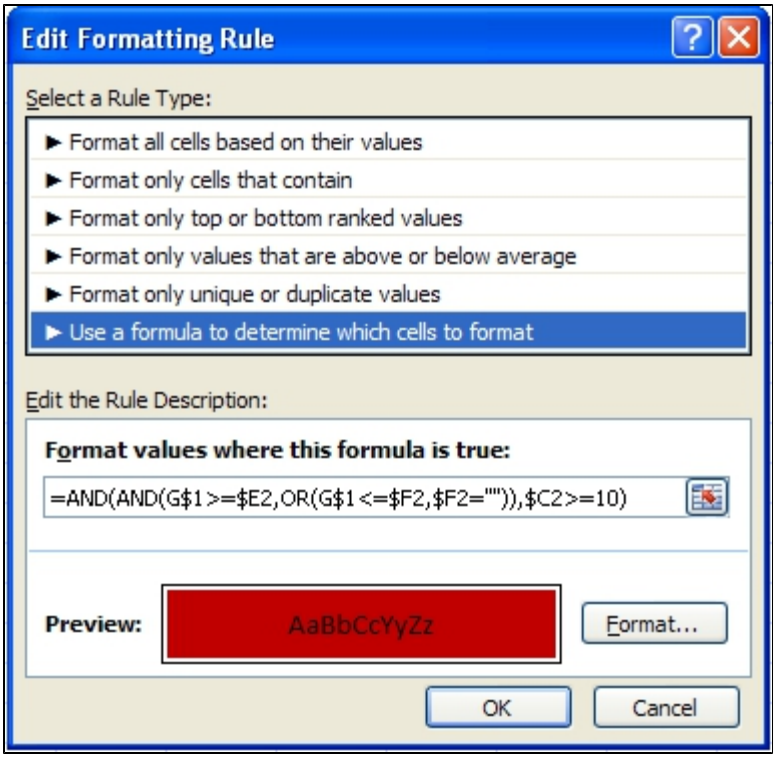
To generate a visual depiction of the bursts in MS Excel perform the following steps:

1. Sort the data ascending by burst start year.

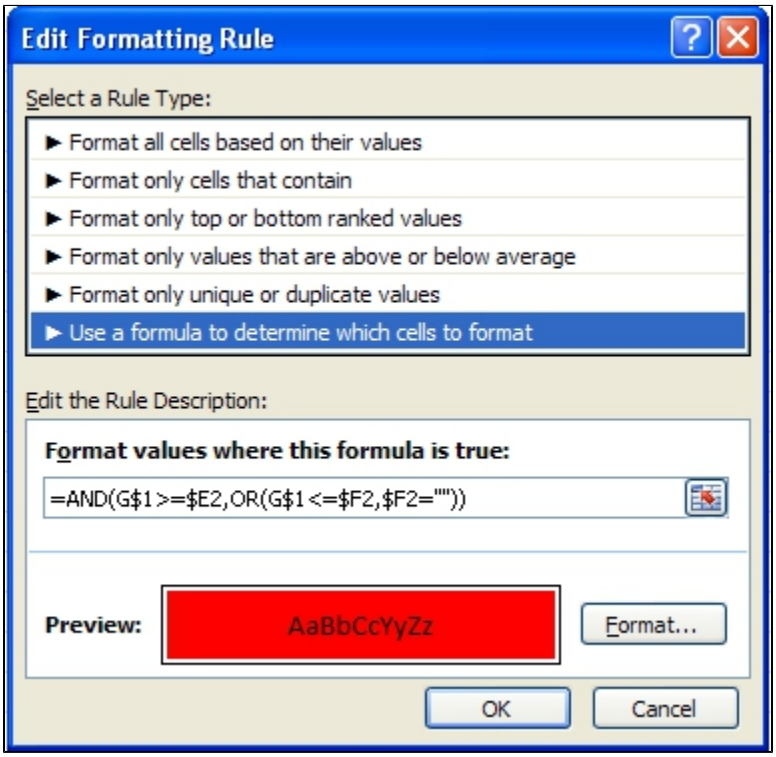
2. Add column headers for all years, i.e., enter first the start year in the cell of index G1, here 1990. As stated before, when there is no value in the "End" field that indicates that the burst lasted until the last date present in the dataset. So continue, e.g., using formula *'=G1+1'*, until highest burst end year, here 2006 in cell W1.

3. In the resulting word by burst year matrix, select the upper left blank cell (G2) and select 'Conditional Formatting' from the ribbon. Then select '*Data Bars > More Rules > Use a formula to determine which cells to format.'* To color cells for years with a burst weight value of more or equal 10 red and cells with a higher value dark red use the following formulas and format patterns:

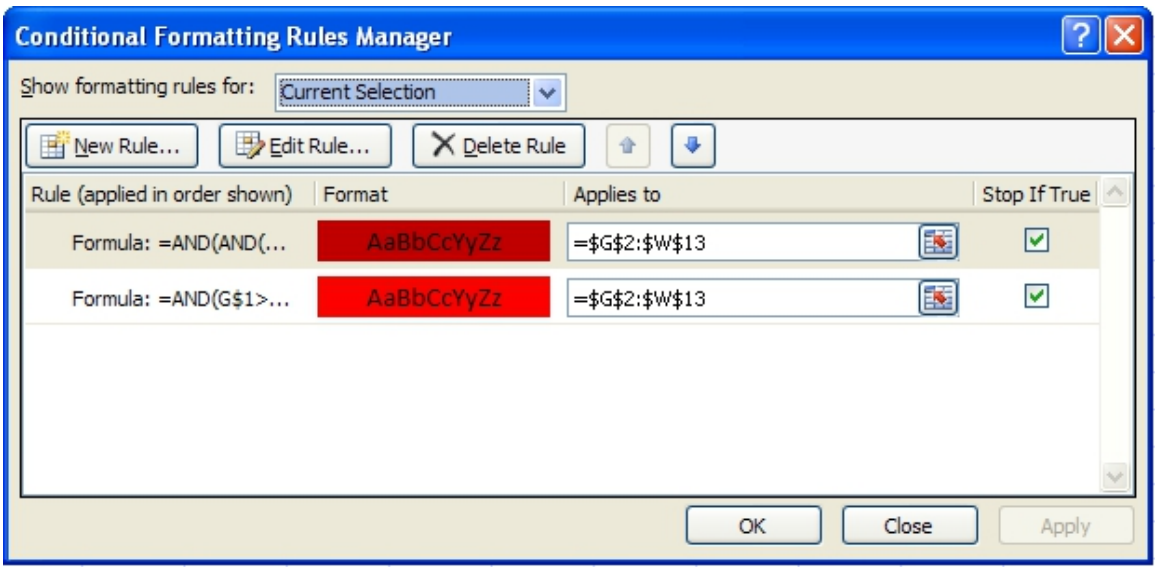
=AND(AND(G$1>=$E2,OR(G$1<=$F2,$F2="")),$C2>=10)



Select *'OK'* and then repeat step three, using the formula below:
=AND(G$1>=$E2,OR(G$1<=$F2,$F2=""))



4. Once both formatting rules have been established, select *'Conditional Formatting > Manage Rules'*, highlight the first formatting rule and move to the top of the list:



5. Make sure both formatting rules are selected and apply them to current selection. Apply the format to all cells in the word by year matrix by dragging the box around cell G2 to highlight all cells in the matrix. The result for the given example is shown in Figure 5.33

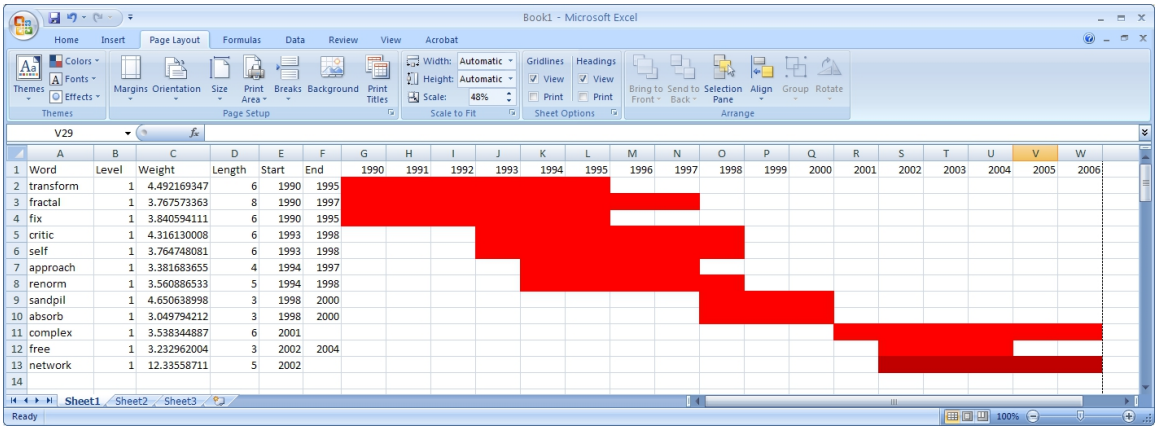


Figure 5.32.1: Visualizing burst results in MS Excel

5.2.6 Mapping the Field of RNAi Research (SDB Data)

RNAi	
Time frame:	1865-2008
Region(s):	Miscellaneous
Topical Area(s):	RNAi
Analysis Type(s):	Co-Author Network, Patent-Citation Network, Burst Detection

The data for this analysis comes from a search of the Scholarly Database (SDB) (http://sdb.cns.iu.edu/) for "RNAi" in "All Text" from Medline, NSF, NIH and USPTO. A copy of this data is available in '***yoursci2directory****/sampledata/sc*

ientometrics/sdb/RNAi' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). The default export format is .csv, which can be loaded directly into the Sci^2 Tool.

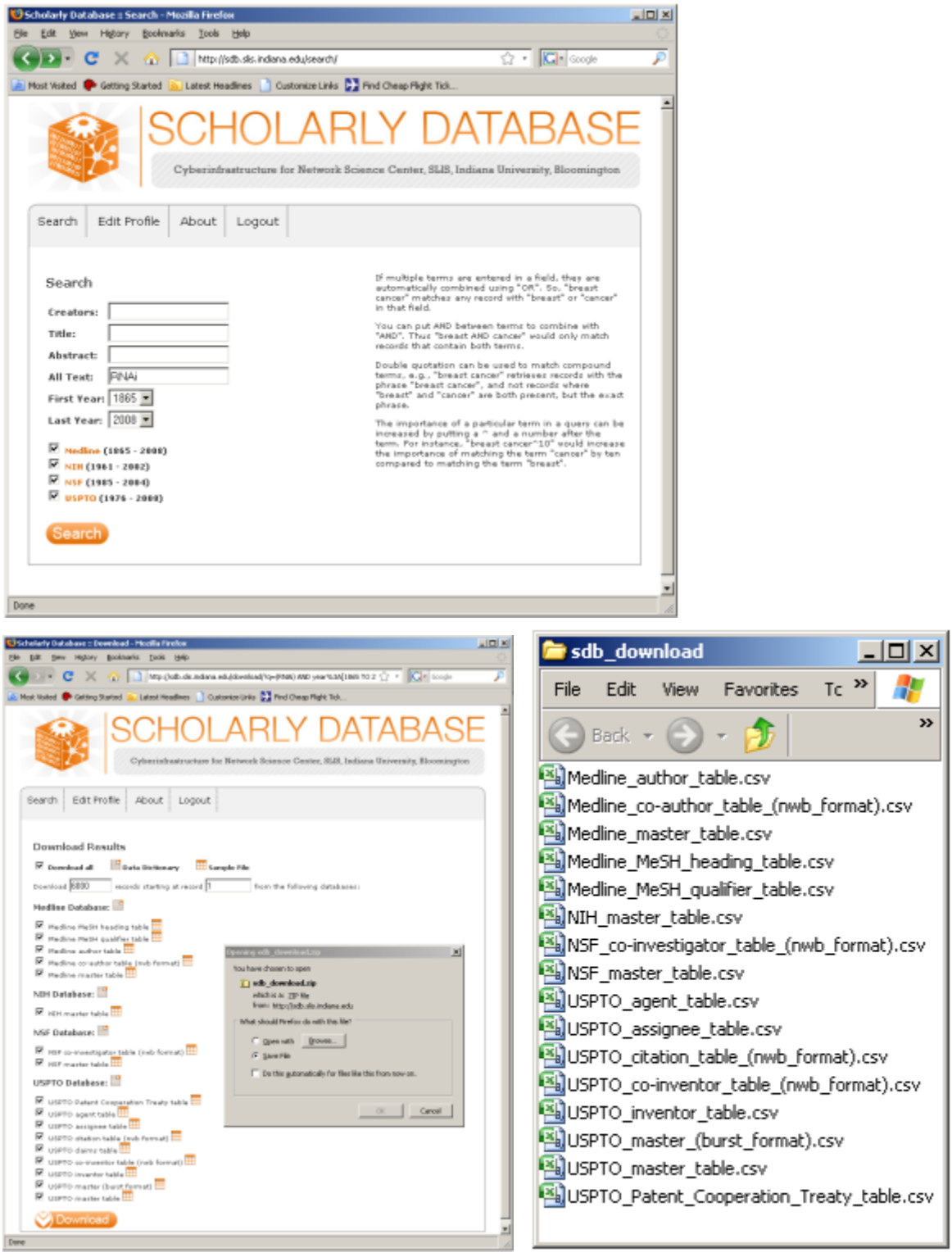
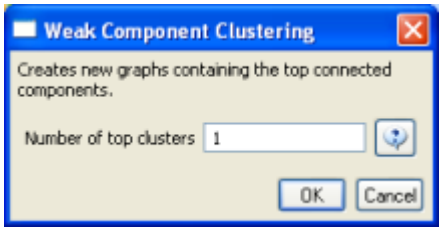


Figure 5.33: Downloading and saving RNAi data from the Scholarly Database.

To view the co-authorship network of Medline's RNAi records, go to '*File* > *Load*' and open '***yoursci2directory****/sampledata/scientometrics/sdb/RNAi/Medline_co-author_table*(nwb_format).csv'_ in Standard csv format (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). SDB tables are already normalized, so simply run '*Data Preparation* > *Extract Co-Occurrence Network*' using the default parameters:

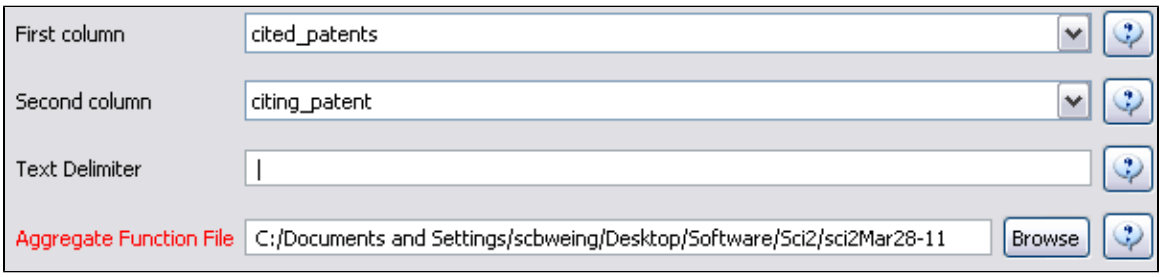
According to '*Analysis* > *Networks* > *Network Analysis Toolkit (NAT)*', the output network has 21,578 nodes with 131 isolates, and 77,739 edges. Visualizing such a large network is memory-intensive, so extract only the largest connected component by running '*Analysis* > *Networks* > *Unweighted and Undirected* > *Weak Component Clustering*' with the following parameters:



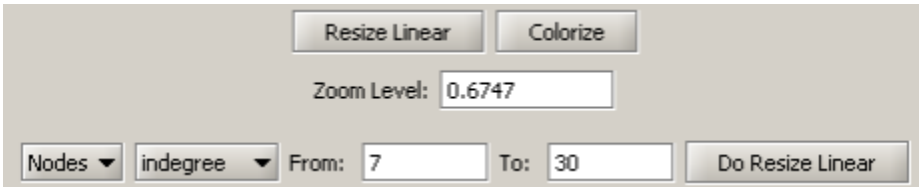
Make sure the newly extracted network ("Weak Component Cluster of 6446 nodes") is selected in the data manager, and run '*Visualization* > *Networks* > *GUESS*' followed by '*Layout* > *GEM*'. A custom python script has been used to color and size the network in Figure 5.34.

Figure 5.34: The Medline RNAi Co-authorship Network

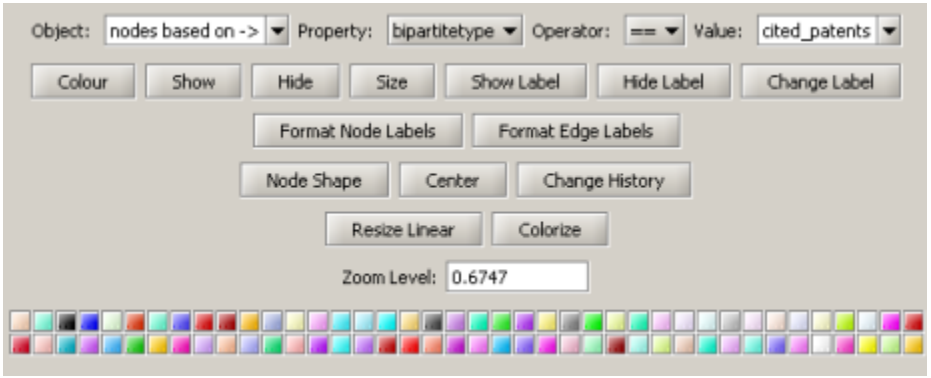
To visualize the citation patterns of patents dealing with RNAi, load '***yoursci2directory****/sampledata/scientometrics/sdb/RNAi/USPTO_citation_table*(nwb_format).csv'_ in Standard csv format (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Then run '*Data Preparation* > *Extract Bipartite Network*' using the following parameters:



Run '*Analysis* > *Networks* > *Unweighted & Directed* > *Node Indegree*' to append Indegree attributes to each node, and then visualize "Network with indegree attribute added to node list" using '*Visualization* > *Networks* > *GUESS*' followed by '*Layout* > *GEM*' and *'Layout* > *Bin Pack'*. In the graph modifier pane, use the following parameters and click "Do Resize Linear."



Then select "nodes based on ->" in the Object drop-down box, "bipartitetype" in the Property drop-down box, "==" in the Operator drop-down box, and "cited_patents" in the Value drop-down box. Press "Colour" and click on blue below.



Repeat the previous steps, but change the Value to "citing_patent" and select the color red. Now press "Show Label". The resulting graph should look like Figure 5.35.

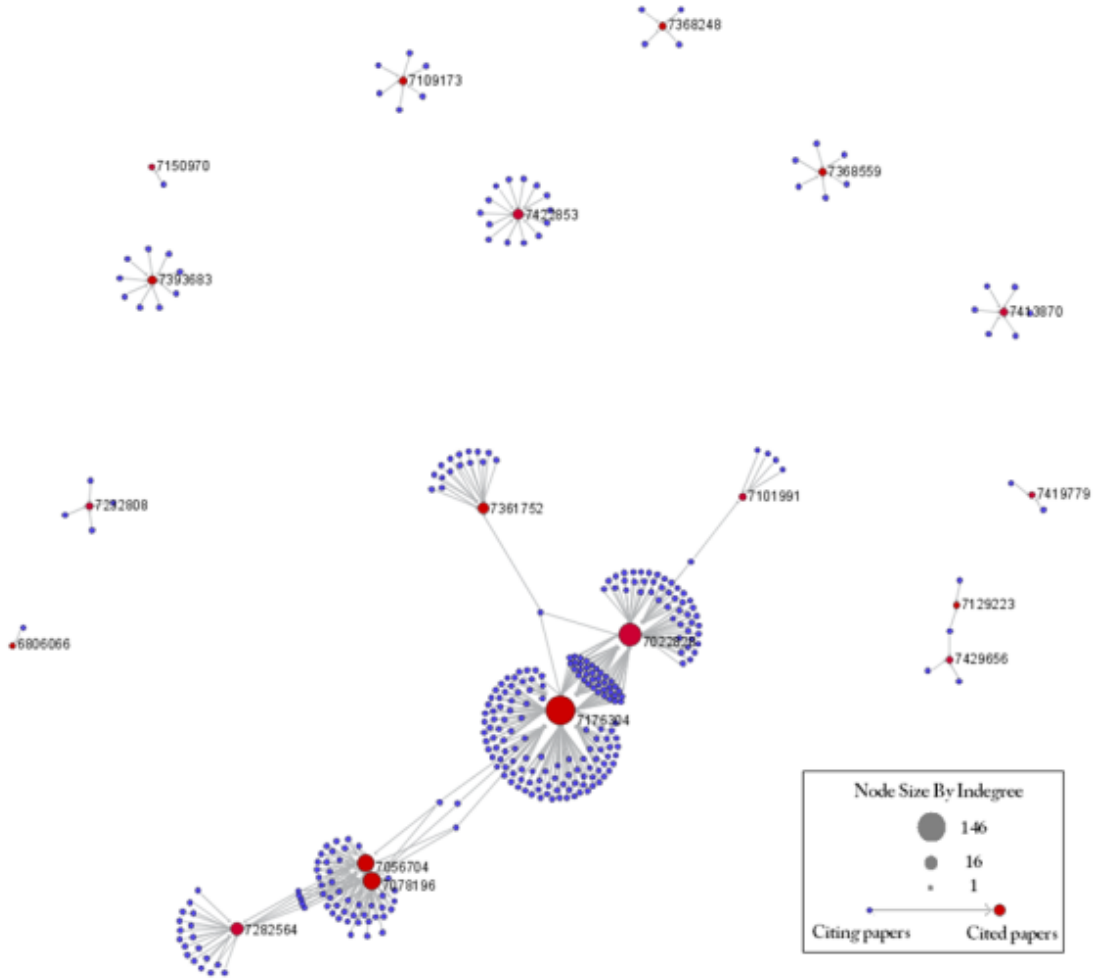
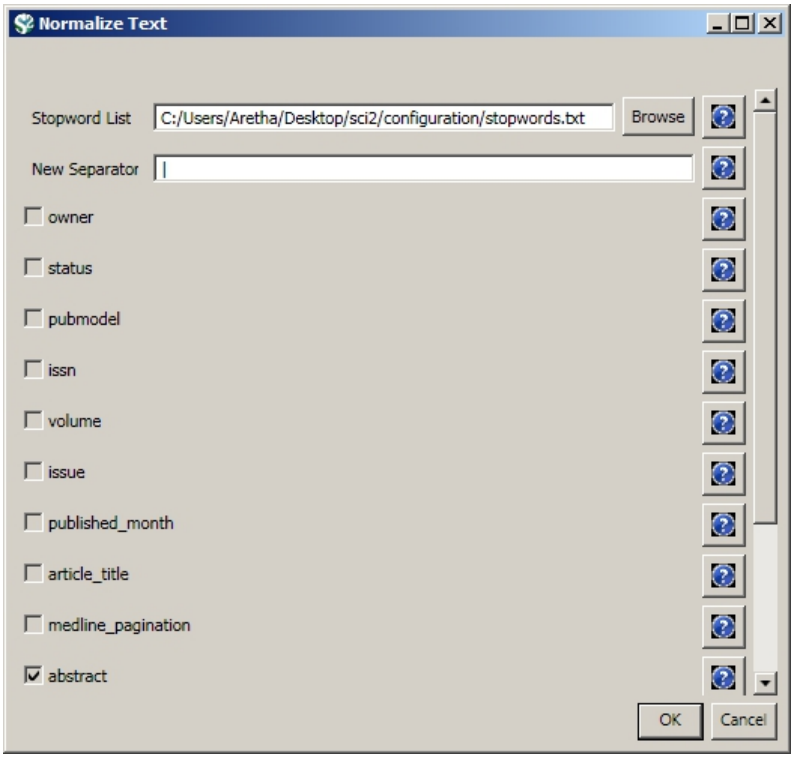


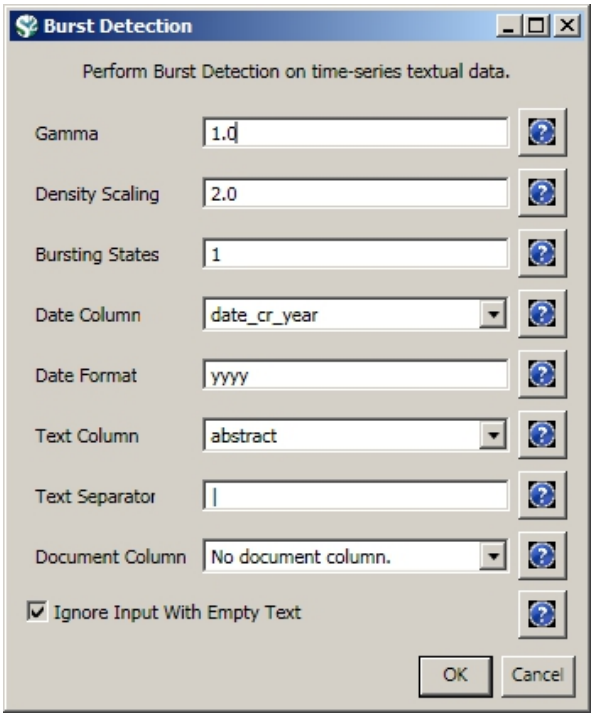
Figure 5.35: USPTO Patent citation network on RNAi

The SDB also outputs much more robust tables, for example '***yoursci2directory****/sampledata/scientometrics/sdb/RNAi/Medline_master_table.csv'.* This table includes full records of Medline papers, and will be used to find bursting terms from Medline abstracts dealing with RNAi. You can download the file by clicking here.

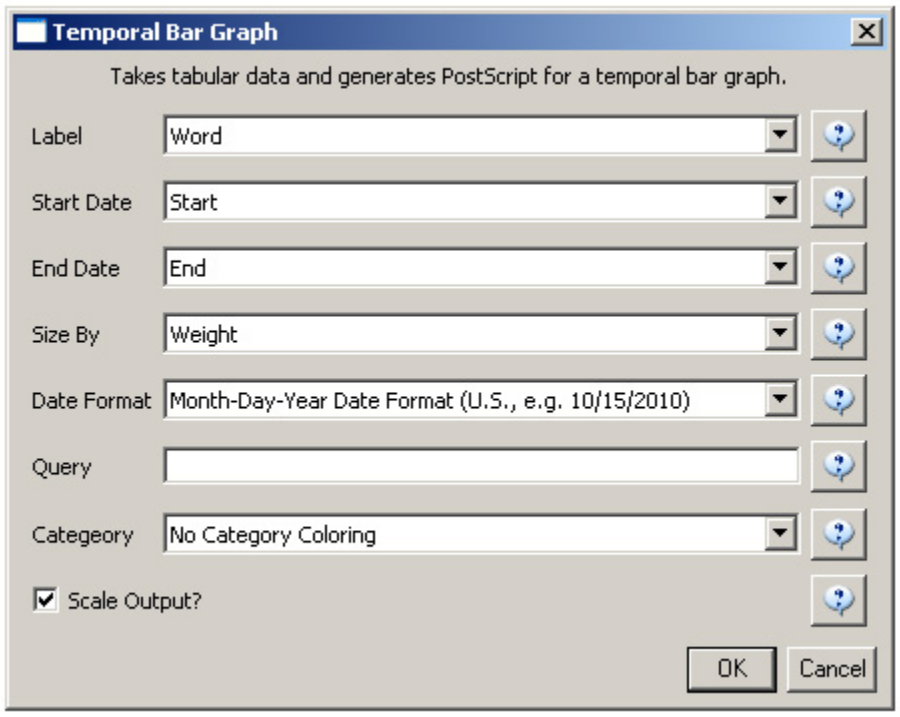
Load the file in Standard csv format and run '*Preprocessing > Topical > Lowercase, Tokenize, Stem, and Stopword Text*' with the following parameters:



Select the "with normalized abstract" table in the Data Manager and run '*Analysis* > *Topical* > *Burst Detection*' with the following parameters:



View the file "Burst detection analysis (date_cr_year, abstract): maximum burst level 1." There are more words than can easily be viewed with the horizontal bar graph, so sort the list by "Strength" and prune all but the strongest 10 words. In other words save the file from the data manager and sort the file by the weight column. Delete all but the top 10 rows. Save the file as a new .csv and load it into the Sci^2 Tool as a standard csv file.
Select the new table in the data manager and visualize it using '*Visualize* > *Temporal* > *Horizontal Bar Graph (not included version)Temporal Bar Graph*' with the following parameters:



Save and view the resulting PostScript file using the workflow described in section [2.4 Saving Visualizations for Publication](#).

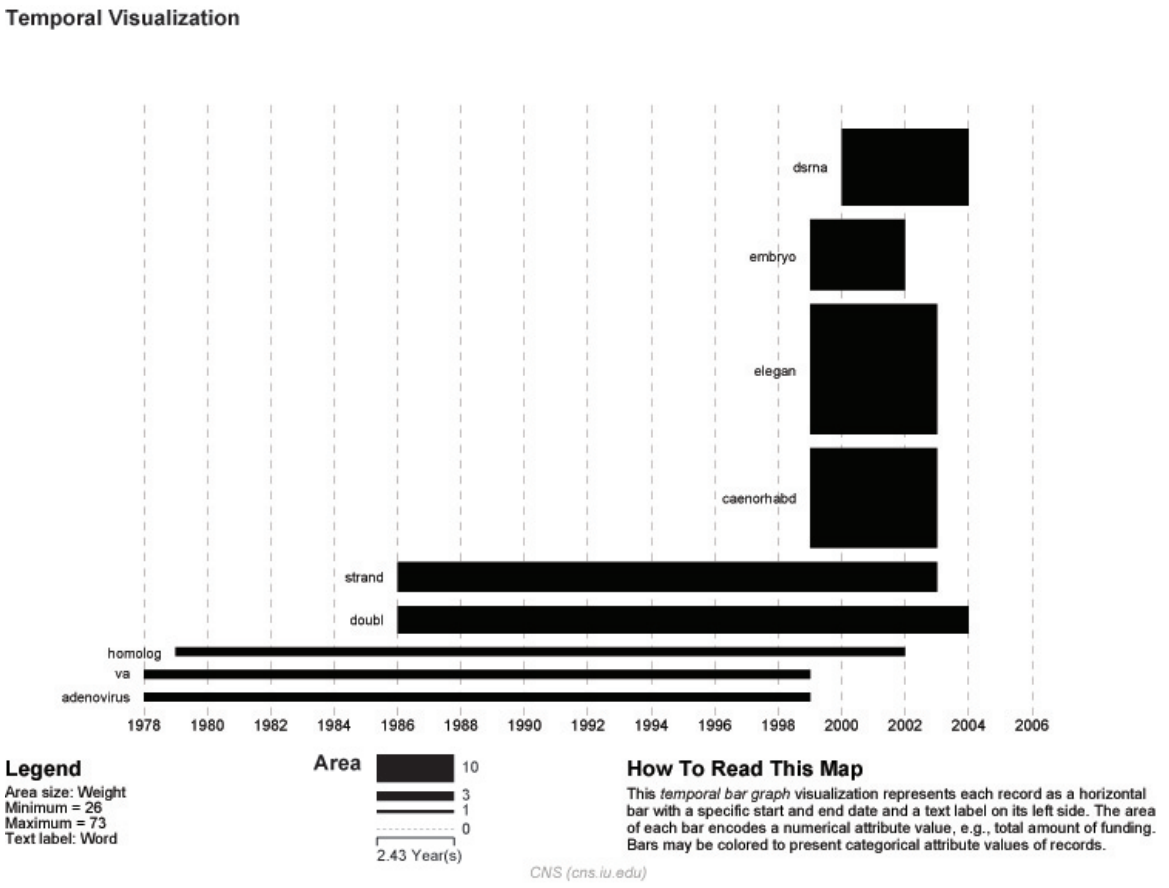


Figure 5.36: Top ten burst terms from Medline abstracts on RNAi Global Level Studies – Meso

To see the log file from this workflow save the 5.2.6 Mapping the Field of RNAi Research (SDB Data) log file.

5.2.7 Visualizing VIVO Data

There are three steps to visualizing VIVO Data in Sci2.
1. Query the desired institution using SPARQL to retrieve the VIVO Data
2. Using Sci2 run various preprocessing algorithms to clean up the data
3. Visualize the Data using GUESS

In this workflow the University of Florida VIVO data will be used to generate an organization Hierarchy Visualization.

Working with SPARQL to Retrieve VIVO Data

To retrieve all organizations and sub-organizations within the University of Florida (UFL) copy and paste the SPARQL query shown below into the text box at the University of Florida's SPARQLer - General purpose processor here.

```
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX core: <http://vivoweb.org/ontology/core#>

SELECT (str(?orgLabel) as ?organizationLabel) (str(?subLabel) as
?subOrganizationLabel)
WHERE
{

?org rdfs:label ?orgLabel;
     core:hasSubOrganization/rdfs:label ?subLabel .
}
```

Once you have copied the code into the text box you can select the desired output. For manipulation data in Sci2 CSV files will work best. Select CSV output and save the file to your desktop. Once the file has been saved rename the file as OrganizationToSubOraganization.csv. This data set will contain sub-organizations that do not belong to any part of the University of Florida system.

Using the following SPARQL query to retrieve the UFL's organizations list from the UFL's SPARQLer. Save the file as UFLOrganizations.csv.

```
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX core: <http://vivoweb.org/ontology/core#>

SELECT (str(?orgLabel) as ?UFLOrganizationLabel)
WHERE
{

<http://vivo.ufl.edu/individual/UniversityofFlorida>
core:hasSubOrganization*/rdfs:label ?orgLabel .
}
```

Now, we want to extract all the rows of the OrganizationToSubOraganization.csv that belong to UFL. To do this, extract the rows from the OrganizationToSubOraganization.csv where their organizationLabel exists in the

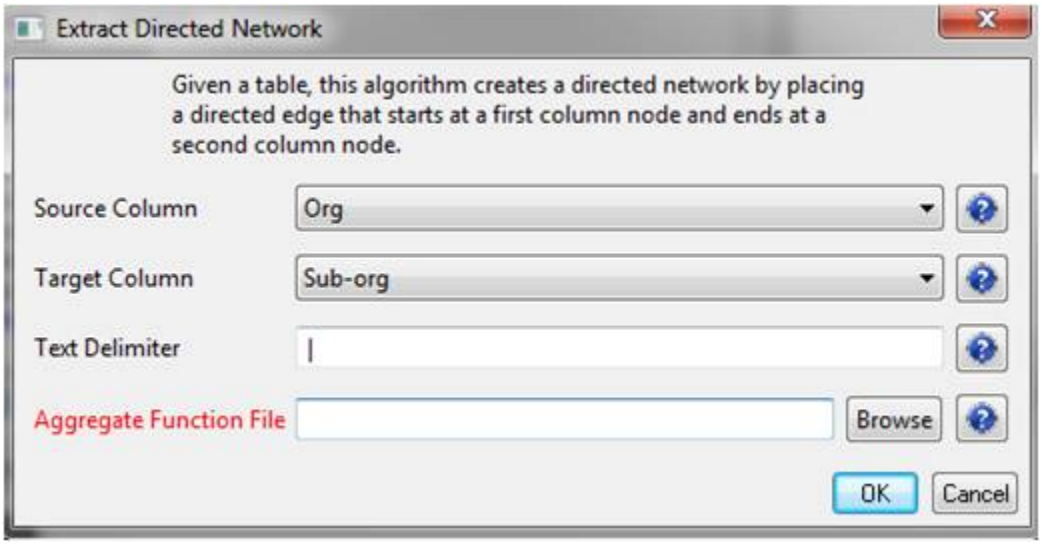
UFLOrganizations.csv. This can be done by using Excel, Python, etc. Save the result as UFLOrgHierarchy.csv

Loading VIVO Data in Sci2

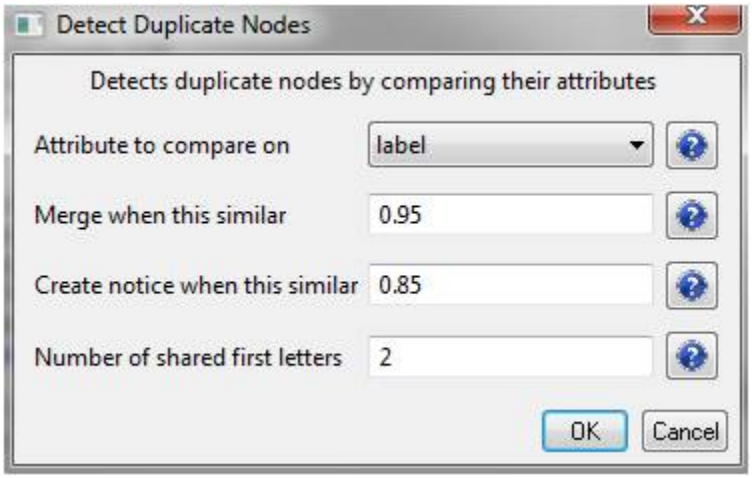
Now open the Sci2 Tool and load the VIVO data by selecting "*File > Load > UFLOrgHierarchy.org*"

Cleaning and Preprocessing VIVO Data

In the Sci2 Tool extract a directed network by running "*Data Preparation > Extract Directed Network*". Use the following parameters:



Next you will want to detect duplicate nodes. Run "*Data Preparation > Detect Duplicate Nodes"* with the default parameters:



A merge table and two log files will be generated. To remove the duplicate nodes, select the merge table and the directed network file, run "*Data Preparation > Update Network by Merging Nodes*".

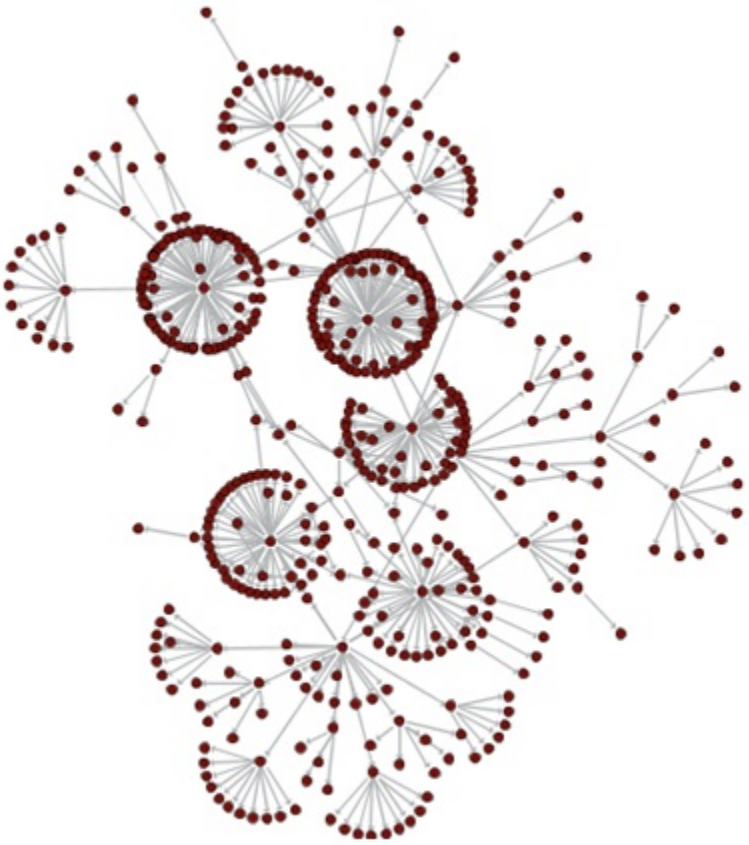
Next generate the in-degree attributes by using the 'Updated Network' file and running "*Analysis > Networks > Unweighted and Directed > Node Indegree"*. Then take the resulting network and generate the out-degree attributes by running "*Analysis > Networks > Unweighted and Directed > Node Outdegree".*

Visualizing in GUESS

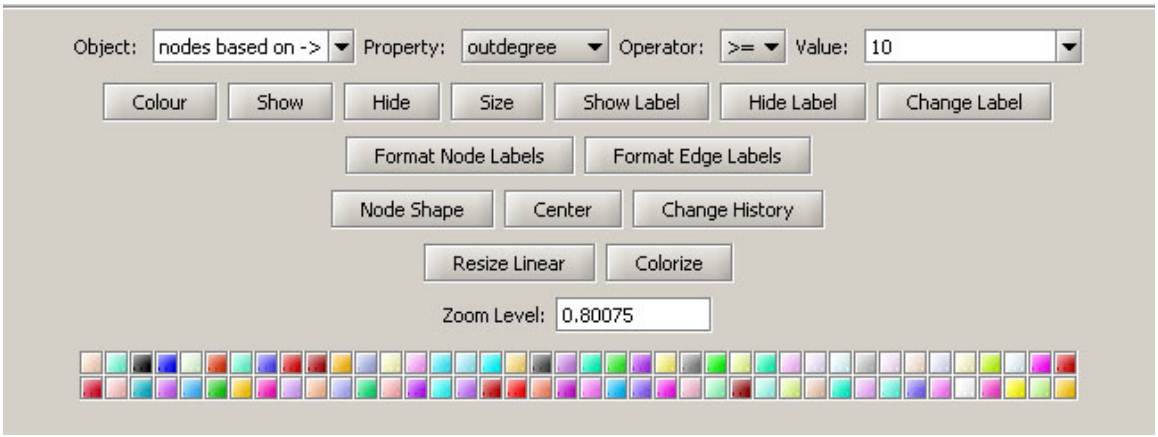
Take the network that results from the above cleaning and preprocessing and visualize it in GUESS by running "*Visu*

alization > Networks > GUESS". This will take some time due to the size of the data set.

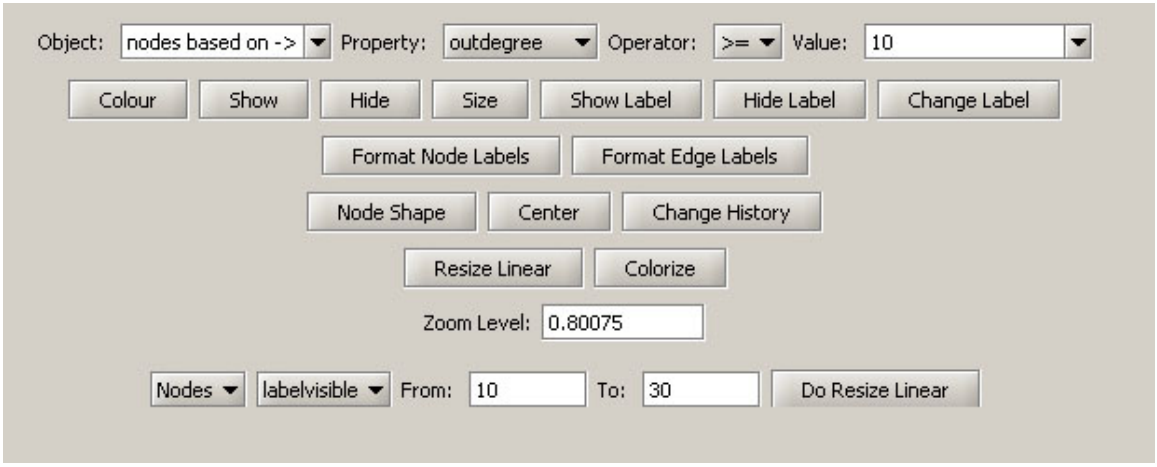
Once the network has been visualized change the layout by running "*Layout > GEM / SPRING / Bin Pack"* or any combination of these layouts to your satisfaction. The image below has used the SPRING layout and then Bin Pack.



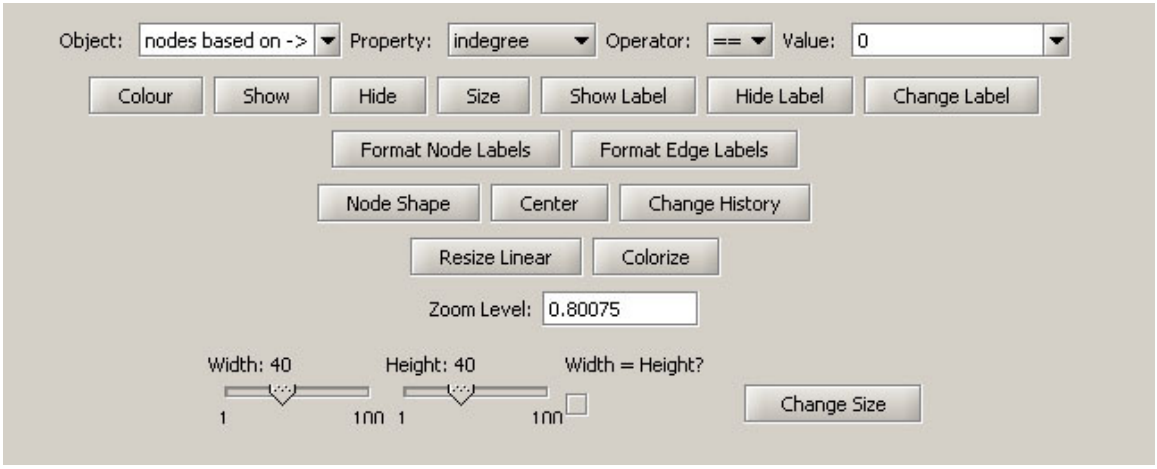
In order to further visualize the network you can use the Graph Modifier. To color the nodes that have more than 10 sub-organizations enter the parameters shown below into the Graph Modifier, select "Colour", and choose the desired color. You might also want to show the labels of these organizations.



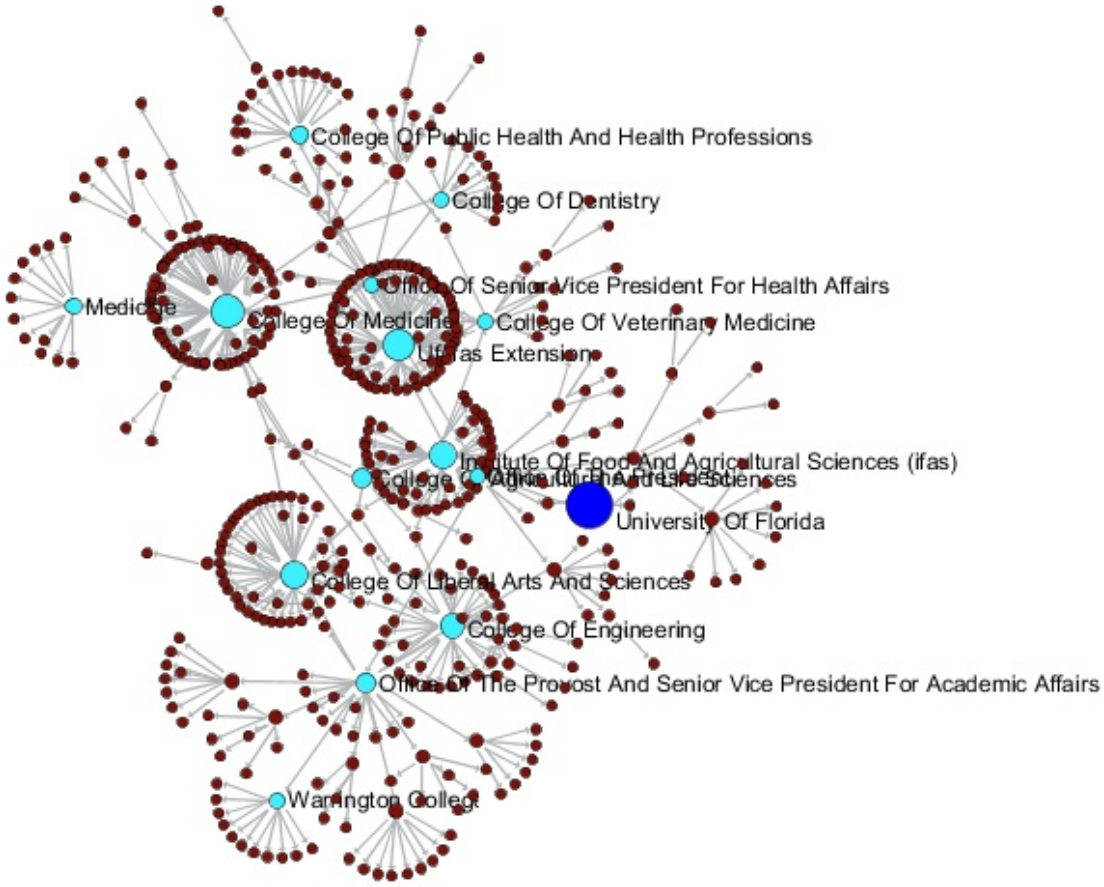
Next, you will want to re-size the nodes linearly based on the number of direct sub-organizations. To achieve this, click on the "Resize Linear" button and fill in the parameters as follows:



Finally, you can highlight the UFL node by coloring and resizing. Use the parameters as shown below to achive the highlight effect.



The result is a visualization of the UFL organization hierarchy.



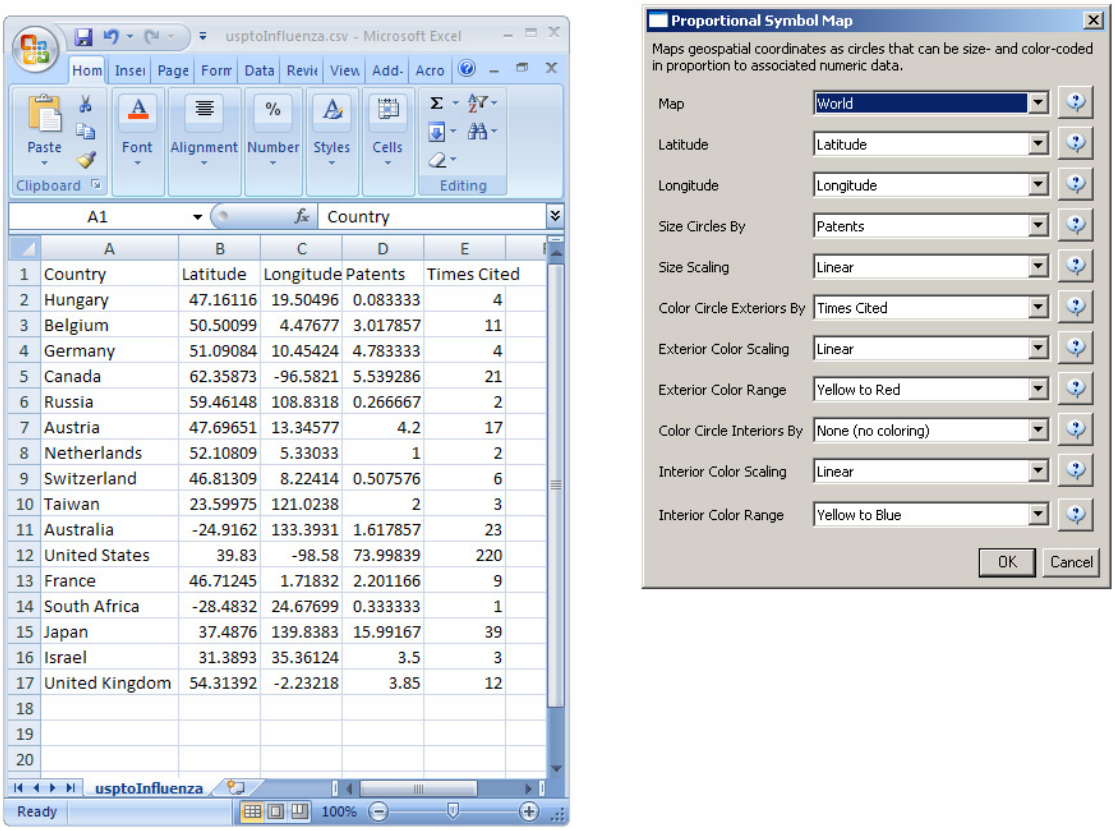
5.3 Global Level Studies - Macro

- 5.3.1 Geo USPTO (SDB Data)
- 5.3.2 Congressional District Geocoder

5.3.1 Geo USPTO (SDB Data)

usptoInfluenza.csv	
Time frame:	1865-2008
Region(s):	Miscellaneous
Topical Area(s):	Influenza
Analysis Type(s):	Geospatial Analysis

The file '*usptoInfluenza.csv*' was generated with an SDB search for patents containing the term "Influenza", and was heavily modified to produce a simple geographic table. Load it using '*File* > *Load*" and following this path: *'**your sci2directory**/sampledata/geo/usptoInfluenza.csv*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Then select "Standard csv format." See the data format in Figure 5.37 (left). Once loaded, select the dataset in the Data Manager and run *Visualization* > *Geospatial* > *Proportional Symbol Map* using the parameters in Figure 5.37 (right). The tool will output a PostScript visualization which can be viewed using GhostView (see section 2.4 Saving Visualizations for Publication and Figure 5.38).



	A	B	C	D	E
1	Country	Latitude	Longitude	Patents	Times Cited
2	Hungary	47.16116	19.50496	0.083333	4
3	Belgium	50.50099	4.47677	3.017857	11
4	Germany	51.09084	10.45424	4.783333	4
5	Canada	62.35873	-96.5821	5.539286	21
6	Russia	59.46148	108.8318	0.266667	2
7	Austria	47.69651	13.34577	4.2	17
8	Netherlands	52.10809	5.33033	1	2
9	Switzerland	46.81309	8.22414	0.507576	6
10	Taiwan	23.59975	121.0238	2	3
11	Australia	-24.9162	133.3931	1.617857	23
12	United States	39.83	-98.58	73.99839	220
13	France	46.71245	1.71832	2.201166	9
14	South Africa	-28.4832	24.67699	0.333333	1
15	Japan	37.4876	139.8383	15.99167	39
16	Israel	31.3893	35.36124	3.5	3
17	United Kingdom	54.31392	-2.23218	3.85	12
18					
19					
20					

Figure 5.37: Geospatial workflow with usptoinfluenza.csv data (left) and Geo Map parameters (right).

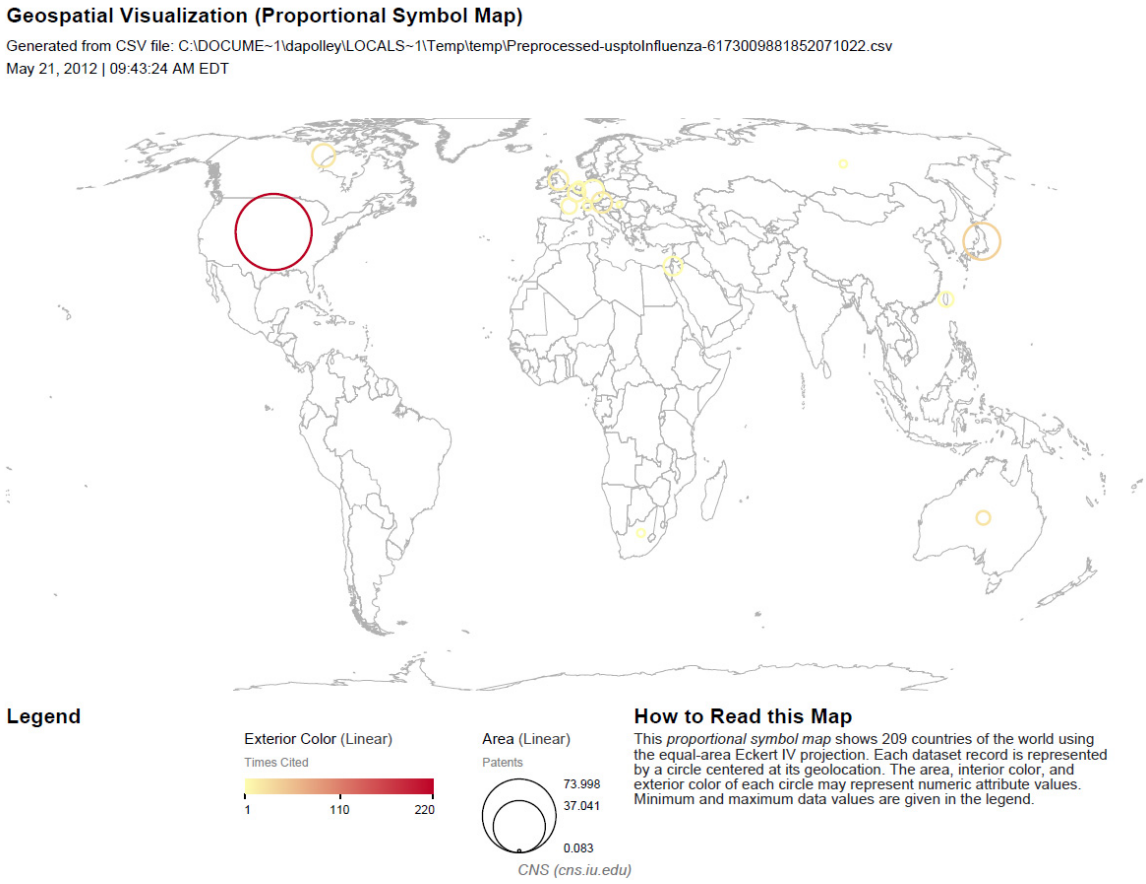
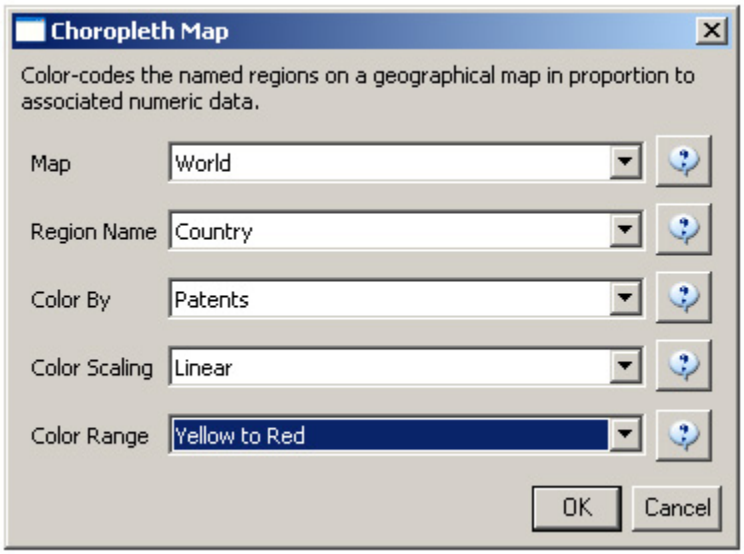


Figure 5.38: Geospatial map (circle annotations) of USPTO patent influemza data

To create a geospatial map with region coding, select *usptoinfluenza.csv* once again and then select *Visualization* > *Geospatial* > *Choropleth Map*. Use the following parameters:



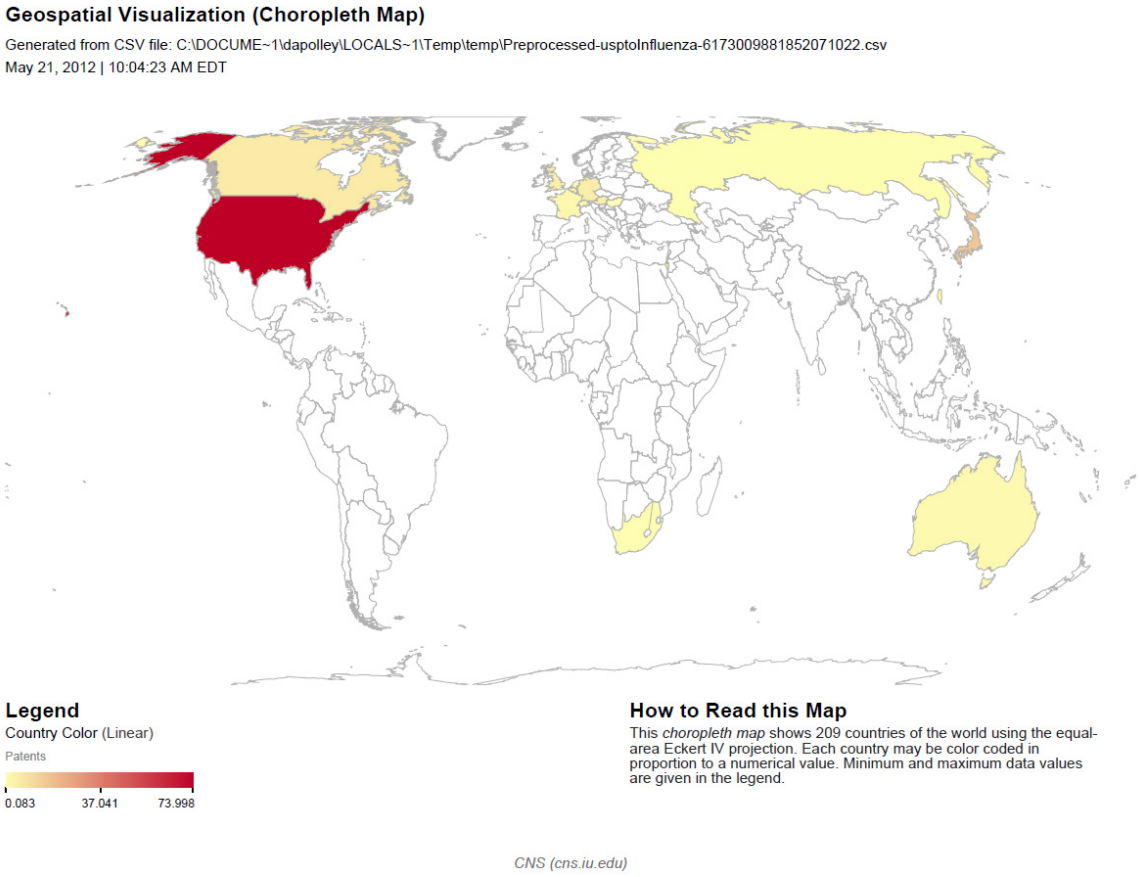


Figure 5.39: Geospatial map (Colored-Region) of USPTO Patent influenza data.

To see the log file from this workflow save the 5.3.1 Geo USPTO (SDB Data) log file.

> **Older Content**
>
> The following pertains to an older version of Sci2. The rest of this workflow may no longer be relevant with the recent changes in the geospatial visualization algorithms.

One can also create a US Geo Map with customized data by running the same workflow but selecting "US States" in "Map", see below.

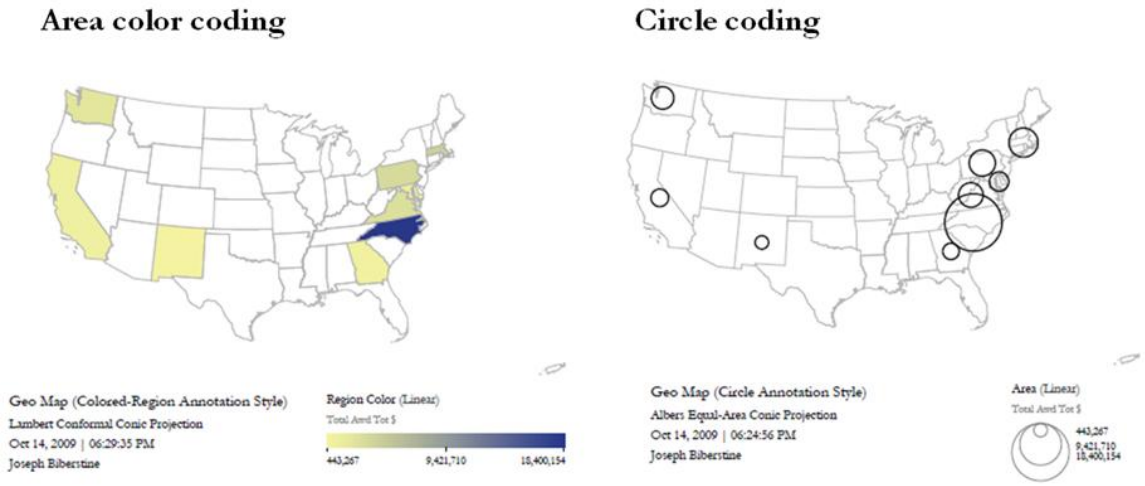


Figure 5.40: US map with area color coding and circle coding for aggregated data over states.

There are two available size scaling options, "Linear" and "Logarithmic". We recommend using logarithmic scaling for larger datasets.

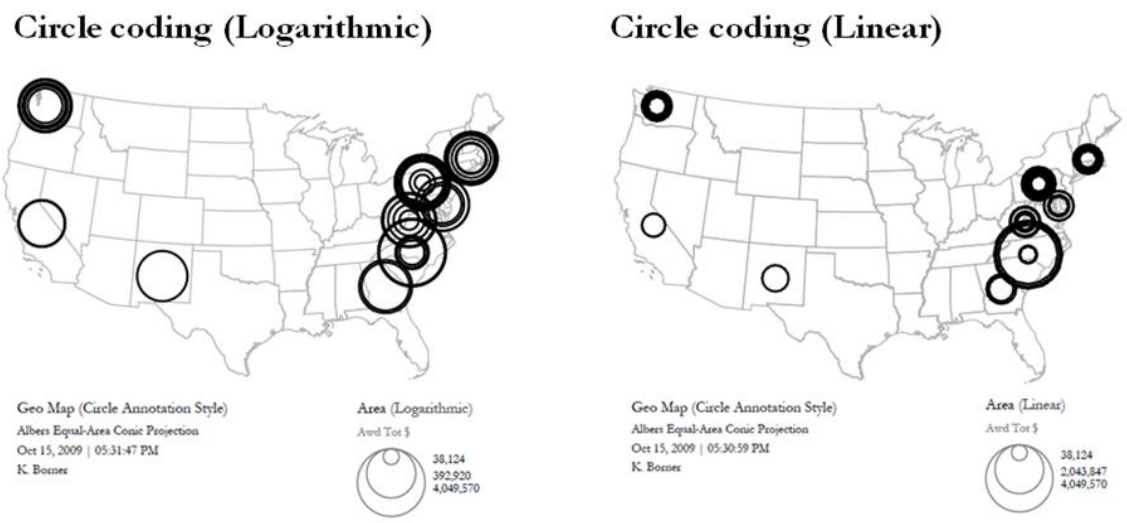
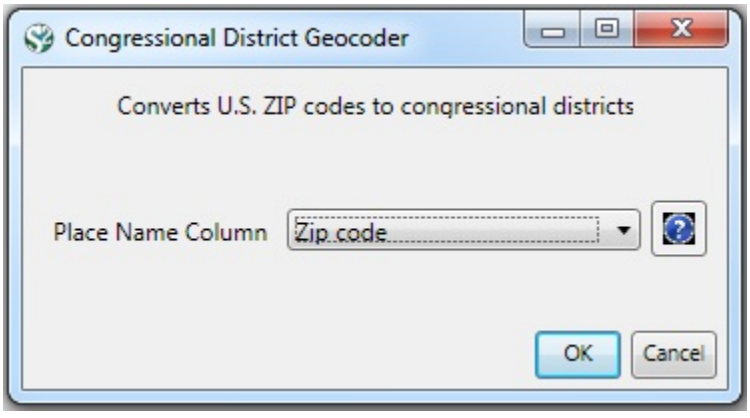


Figure 5.41: US geospatialmap of state-level data with logarithmic circle size scaling (left) and circle linear size scaling (right).

5.3.2 Congressional District Geocoder

zip code.csv	
Region(s):	United States
Analysis Type(s):	Geospatial Analaysis

To visualize Congressional Districts you must first extract that data from a dataset containing either ZIP codes or addresses. You can download the Congressional District Geocoder plugin here. You can load any file that contains 9-digit U.S. ZIP codes to be geocoded. A sample file can be loaded by using '*File* > *Load*' and following this path: '***yo ursic2directory****/sampledat/geo/zipcode.csv*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Load the file in Standard csv format.Then select the file in the data manager and use '*A nalysis* > *Geospatial* > *Congressional District Geocoder*' with the following parameters:



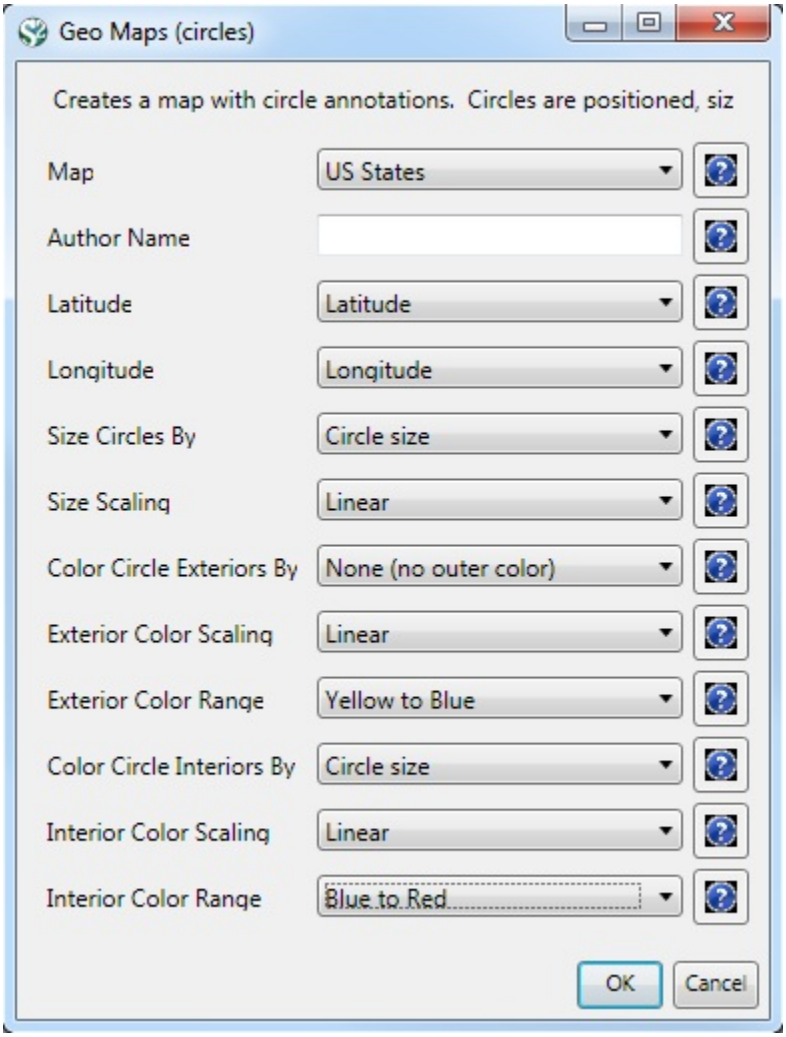
5-digits ZIP codes with multiple congressional districts, empty entries and invalid ZIP codes that failed to be geocoded will list in warning messages on the console.The output table contains all columns of the input table with three additional columns appended: Congressional district, latitude, and longitude.To view the output table save the file using '*File* > *Save...*' and selecting the desired save location (to view the file in Excel save it as a csv file). Once the file has been saved it can be viewed with your choice of program. Below the file has been opened as a csv file:

	A	B	C	D
1	Zip code	Congressional District	Latitude	Longitude
2	90095	CA-30	34.0735035	-118.6645815
3	4672	ME-02	45.818717	-69.0290345
4	232980568	VA-03	37.270472	-77.0699835

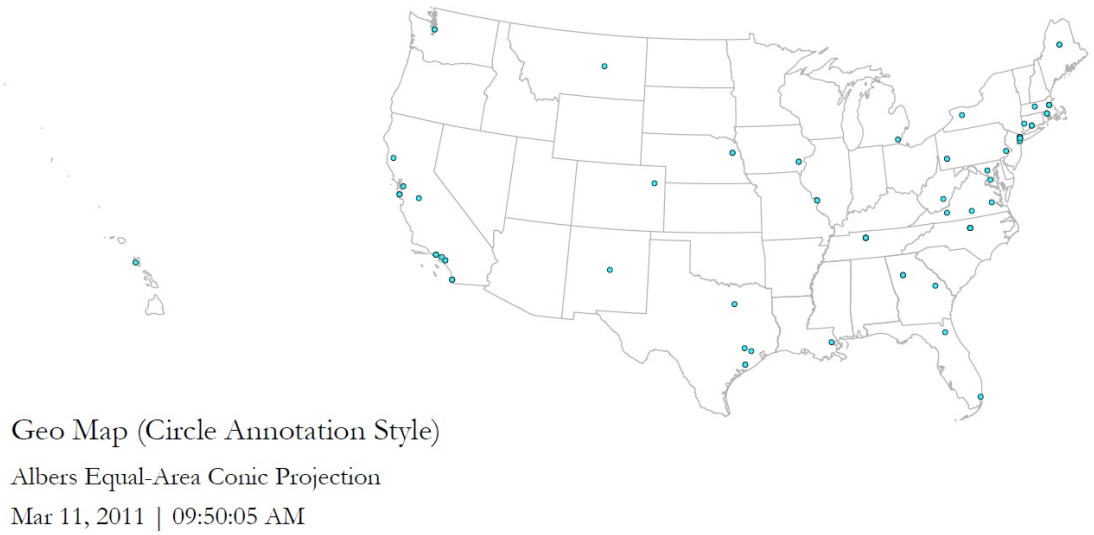
Before you you can visualize this data you will need to edit the csv file (shown above). It will be easiest to edit the file with Excel. First save the file from the data manager in Sci2 to a desired location. Open the file with Excel and add a column titles "Circle size". The value you chose does not matter, but it should be consistent across congressional districts. (Note: the smaller the value the more precise the visualization will be with regard to Congressional District.) Below 0.5 was chosen for circle size:

	A	B	C	D	E
1	Zip code	Congressional District	Latitude	Longitude	Circle size
2	90095	CA-30	34.0735035	-118.6645815	0.5
3	4672	ME-02	45.818717	-69.0290345	0.5
4	232980568	VA-03	37.270472	-77.0699835	0.5

Once the csv file has been edited, reload it into Sci2. To visualize the newly loaded dataset, select the file in the data manager. Then select *Visualization* > *Geospatial* > *Proportional Symbol Map* and use the following parameters:



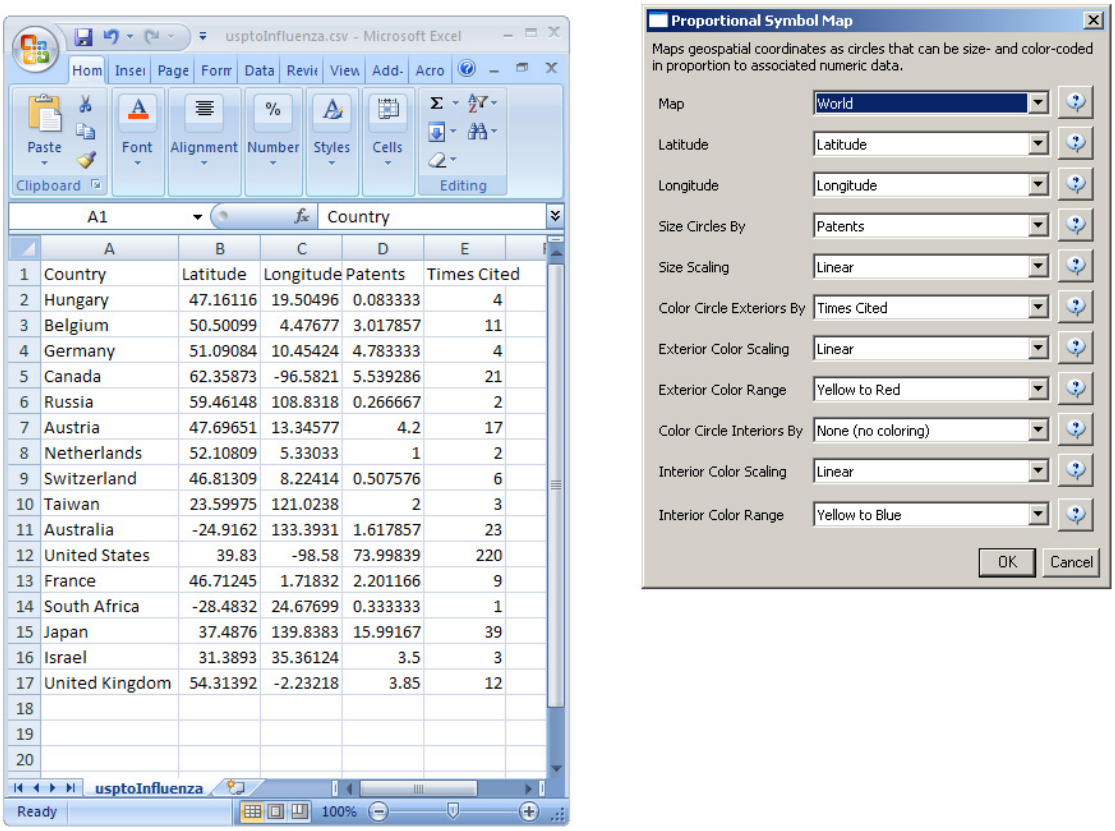
Note: to color the interior of the circles you must select a value for the "Color Circle Interior By" tab. Here Circle size was selected. This value is arbitrary, but it's consistency results in consistent coloring for the final visualization:



5.3.1 Geo USPTO (SDB Data)

usptoInfluenza.csv	
Time frame:	1865-2008
Region(s):	Miscellaneous
Topical Area(s):	Influenza
Analysis Type(s):	Geospatial Analysis

The file '*usptoInfluenza.csv*' was generated with an SDB search for patents containing the term "Influenza", and was heavily modified to produce a simple geographic table. Load it using '*File* > *Load*" and following this path: ***'your sci2directory****/sampledata/geo/usptoInfluenza.csv'* (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Then select "Standard csv format." See the data format in Figure 5.37 (left). Once loaded, select the dataset in the Data Manager and run *Visualization* > *Geospatial* > *Proportional Symbol Map* using the parameters in Figure 5.37 (right). The tool will output a PostScript visualization which can be viewed using GhostView (see section 2.4 Saving Visualizations for Publication and Figure 5.38).



	A	B	C	D	E
1	Country	Latitude	Longitude	Patents	Times Cited
2	Hungary	47.16116	19.50496	0.083333	4
3	Belgium	50.50099	4.47677	3.017857	11
4	Germany	51.09084	10.45424	4.783333	4
5	Canada	62.35873	-96.5821	5.539286	21
6	Russia	59.46148	108.8318	0.266667	2
7	Austria	47.69651	13.34577	4.2	17
8	Netherlands	52.10809	5.33033	1	2
9	Switzerland	46.81309	8.22414	0.507576	6
10	Taiwan	23.59975	121.0238	2	3
11	Australia	-24.9162	133.3931	1.617857	23
12	United States	39.83	-98.58	73.99839	220
13	France	46.71245	1.71832	2.201166	9
14	South Africa	-28.4832	24.67699	0.333333	1
15	Japan	37.4876	139.8383	15.99167	39
16	Israel	31.3893	35.36124	3.5	3
17	United Kingdom	54.31392	-2.23218	3.85	12
18					
19					
20					

Figure 5.37: Geospatial workflow with usptoinfluenza.csv data (left) and Geo Map parameters (right).

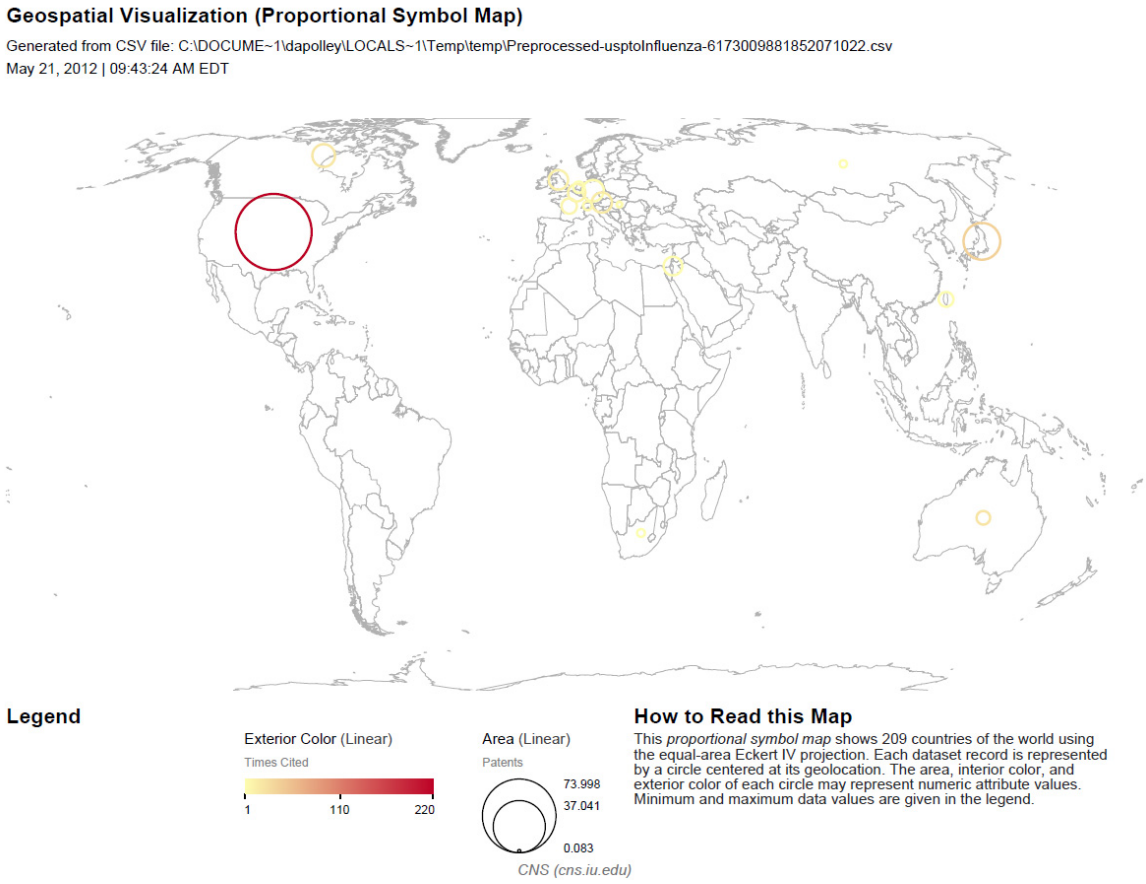
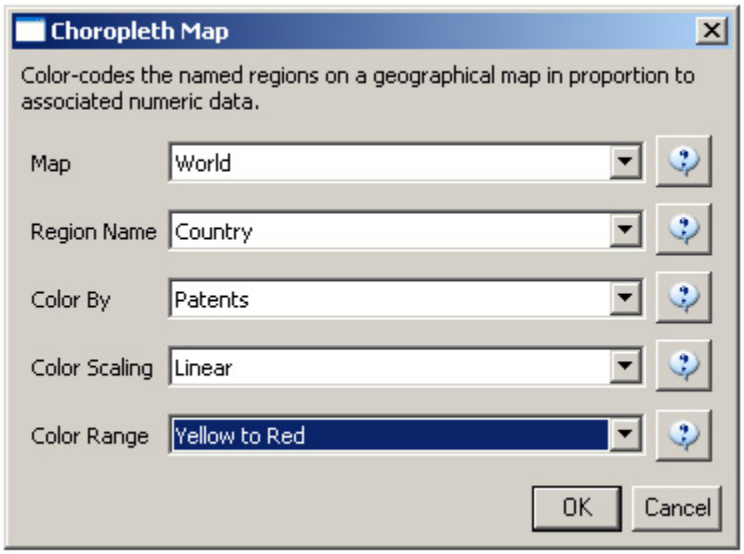


Figure 5.38: Geospatial map (circle annotations) of USPTO patent influemza data

To create a geospatial map with region coding, select *usptoinfluenza.csv* once again and then select *Visualization* > *Geospatial* > *Choropleth Map*. Use the following parameters:



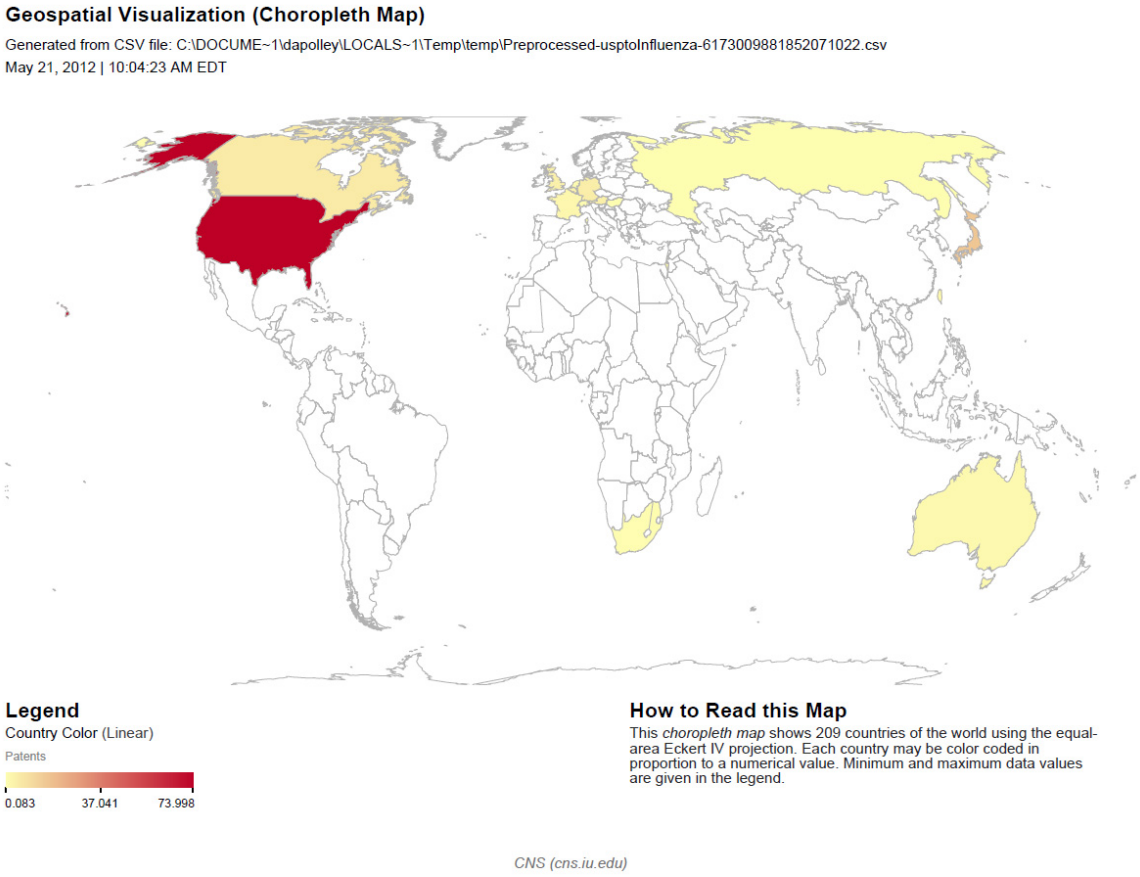


Figure 5.39: Geospatial map (Colored-Region) of USPTO Patent influenza data.

To see the log file from this workflow save the 5.3.1 Geo USPTO (SDB Data) log file.

> **Older Content**
>
> The following pertains to an older version of Sci2. The rest of this workflow may no longer be relevant with the recent changes in the geospatial visualization algorithms.

One can also create a US Geo Map with customized data by running the same workflow but selecting "US States" in "Map", see below.

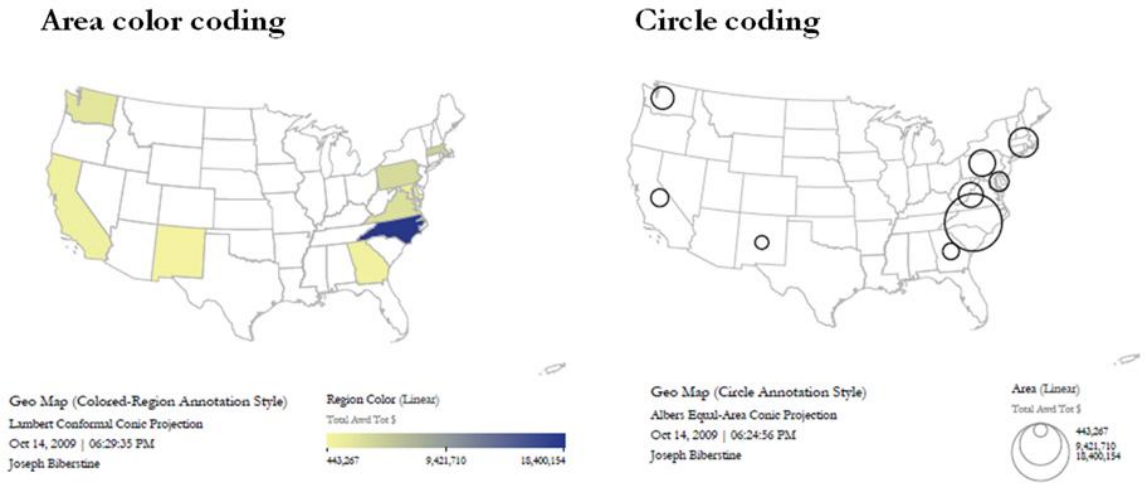


Figure 5.40: US map with area color coding and circle coding for aggregated data over states.

There are two available size scaling options, "Linear" and "Logarithmic". We recommend using logarithmic scaling for larger datasets.

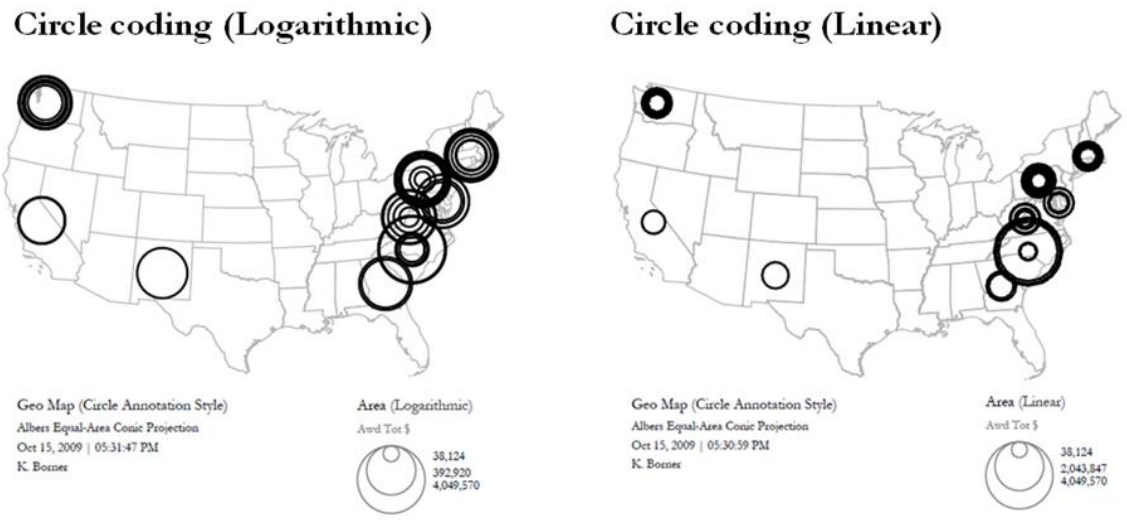
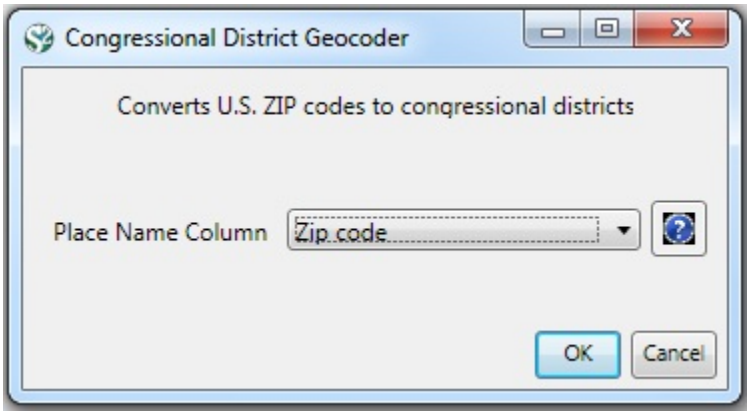


Figure 5.41: US geospatialmap of state-level data with logarithmic circle size scaling (left) and circle linear size scaling (right).

5.3.2 Congressional District Geocoder

zip code.csv	
Region(s):	United States
Analysis Type(s):	Geospatial Analaysis

To visualize Congressional Districts you must first extract that data from a dataset containing either ZIP codes or addresses. You can download the Congressional District Geocoder plugin here. You can load any file that contains 9-digit U.S. ZIP codes to be geocoded. A sample file can be loaded by using '*File > Load*' and following this path: '***yo ursic2directory****/sampledat/geo/zipcode.csv*' (if the file is not in the sample data directory it can be downloaded from 2.5 Sample Datasets). Load the file in Standard csv format.Then select the file in the data manager and use '*A nalysis > Geospatial > Congressional District Geocoder*' with the following parameters:



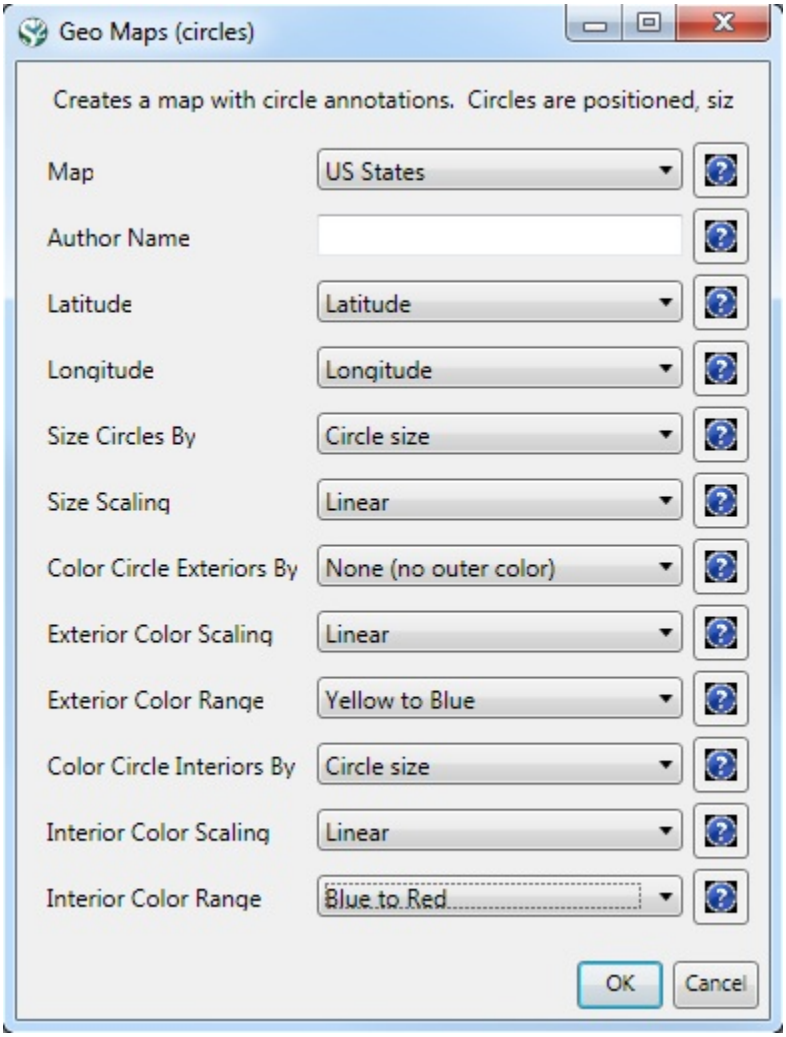
5-digits ZIP codes with multiple congressional districts, empty entries and invalid ZIP codes that failed to be geocoded will list in warning messages on the console.The output table contains all columns of the input table with three additional columns appended: Congressional district, latitude, and longitude.To view the output table save the file using '*File > Save...*' and selecting the desired save location (to view the file in Excel save it as a csv file). Once the file has been saved it can be viewed with your choice of program. Below the file has been opened as a csv file:

	A	B	C	D
1	Zip code	Congressional District	Latitude	Longitude
2	90095	CA-30	34.0735035	-118.6645815
3	4672	ME-02	45.818717	-69.0290345
4	232980568	VA-03	37.270472	-77.0699835

Before you you can visualize this data you will need to edit the csv file (shown above). It will be easiest to edit the file with Excel. First save the file from the data manager in Sci2 to a desired location. Open the file with Excel and add a column titles "Circle size". The value you chose does not matter, but it should be consistent across congressional districts. (Note: the smaller the value the more precise the visualization will be with regard to Congressional District.) Below 0.5 was chosen for circle size:

	A	B	C	D	E
1	Zip code	Congressional District	Latitude	Longitude	Circle size
2	90095	CA-30	34.0735035	-118.6645815	0.5
3	4672	ME-02	45.818717	-69.0290345	0.5
4	232980568	VA-03	37.270472	-77.0699835	0.5

Once the csv file has been edited, reload it into Sci2. To visualize the newly loaded dataset, select the file in the data manager. Then select *Visualization* > *Geospatial* > *Proportional Symbol Map* and use the following parameters:



Note: to color the interior of the circles you must select a value for the "Color Circle Interior By" tab. Here Circle size was selected. This value is arbitrary, but it's consistency results in consistent coloring for the final visualization:

Geo Map (Circle Annotation Style)
Albers Equal-Area Conic Projection
Mar 11, 2011 | 09:50:05 AM

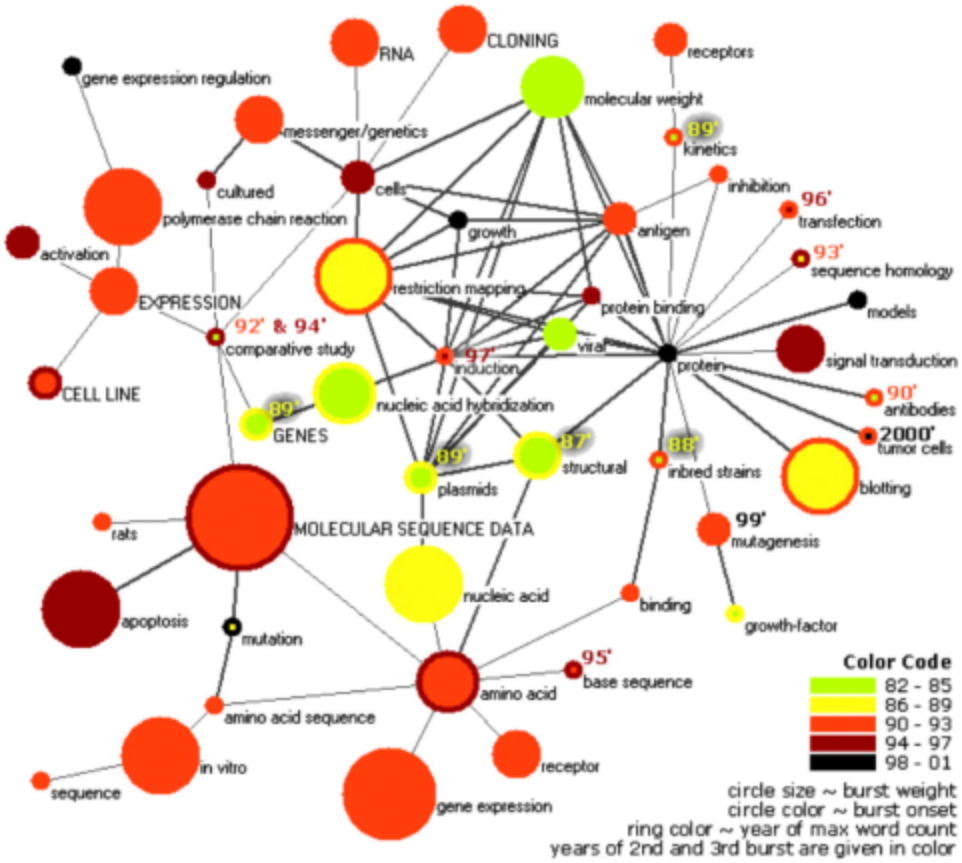
Created with Sci² Tool | Cyberinfrastructure for Network Science Center (http://cns.slis.indiana.edu)

6 Sample Science Studies & Online Services

6.1 Science Dynamics

6.1.1 Mapping Topics and Topic Bursts in PNAS (2004)

By *Ketan K. Mane & Katy Börner (2004)*
Scientific research is highly dynamic. New areas of science continually evolve; others gain or lose importance, merge, or split. Due to the steady increase in the number of scientific publications, it is hard to keep an overview of the structure and dynamic development of one's own field of science, much less all scientific domains.
However, knowledge of "hot" topics, emergent research frontiers, or change of focus in certain areas is a critical component of resource allocation decisions in research laboratories, governmental institutions, and corporations. This paper demonstrates the utilization of Kleinberg's burst detection algorithm, co-word occurrence analysis, and graph layout techniques to generate maps that support the identification of major research topics and trends. The approach was applied to analyze and map the complete set of papers published in PNAS in the years 1982--2001. Six domain experts examined and commented on the resulting maps in an attempt to reconstruct the evolution of major research areas covered by PNAS.

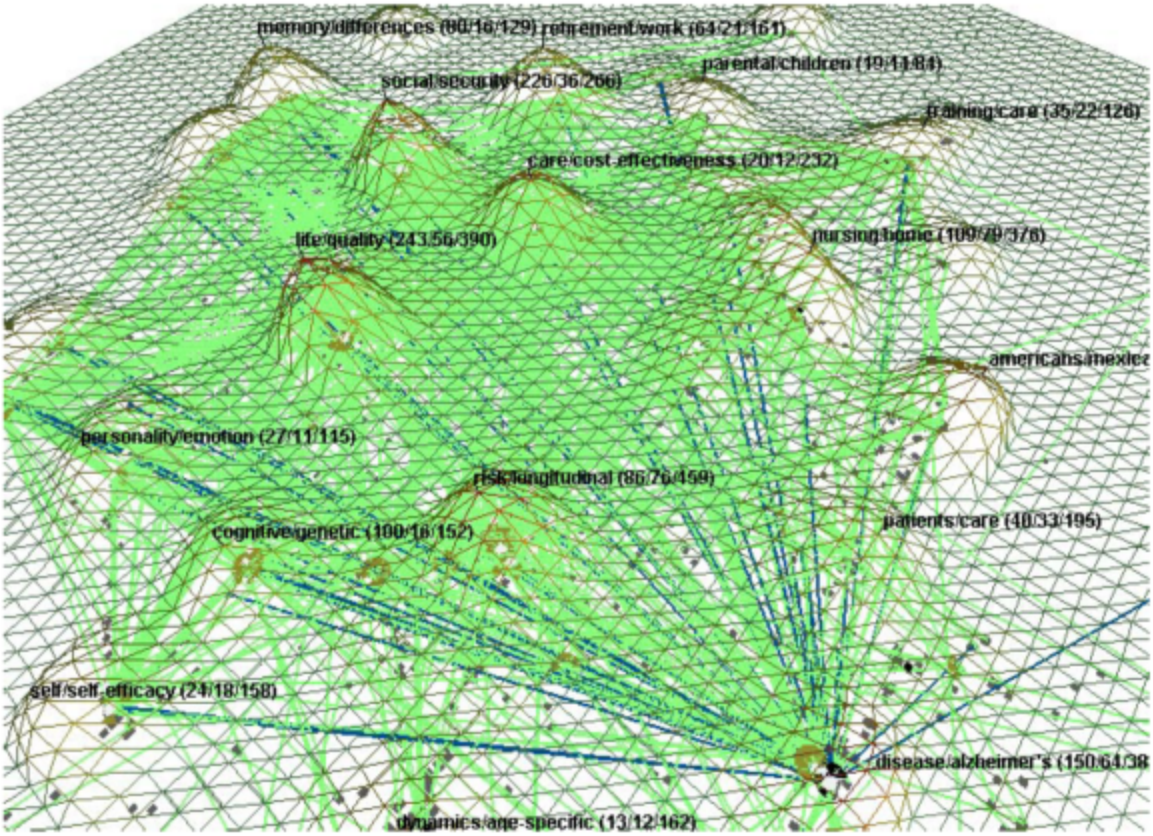


Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982--2001.

6.2 Local Input-Output and ROI Studies

6.2.1 Indicator-Assisted Evaluation and Funding of Research: Visualizing the Influence of Grants on the Number and Citation Counts of Research Papers (2003)

By *Kevin W. Boyack & Katy Börner (2003)*
This article reports research on analyzing and visualizing the impact of governmental funding on the amount and citation counts of research publications. For the first time, grant and publication data appear interlinked in one map. We start with an overview of related work and a discussion of available techniques. A concrete example - grant and publication data from Behavioral and Social Science Research, one of four extramural research programs at the National Institute on Aging (NIA) - is analyzed and visualized using the VxInsight® visualization tool. The analysis also illustrates current existing problems related to the quality and existence of data, data analysis, and processing. The article concludes with a list of recommendations on how to improve the quality of grant-publication maps and a discussion of research challenges for indicator-assisted evaluation and funding of research.

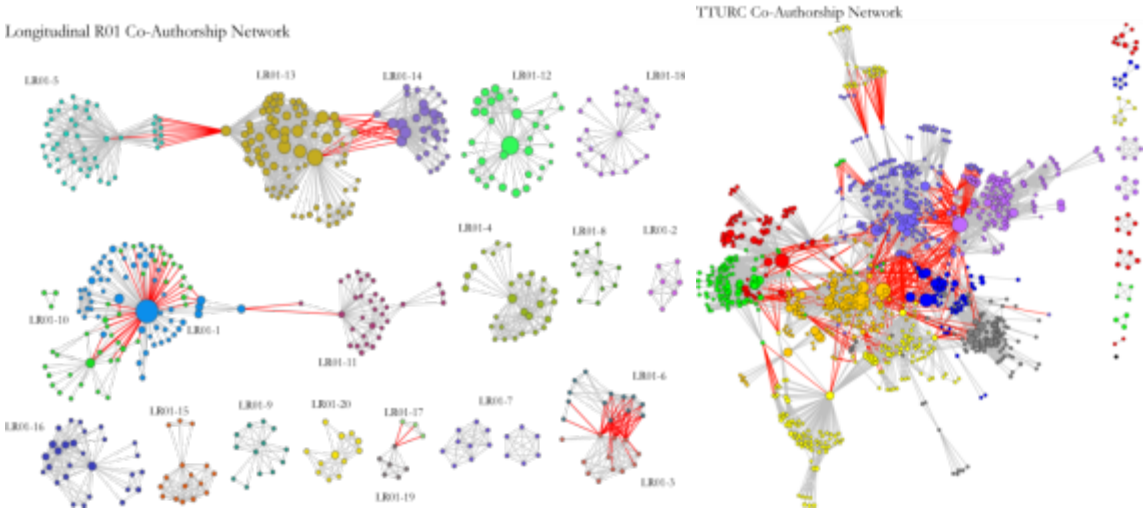


Author supplied linkage patterns (light gray lines) from grants to publications with links highlighted as dark lines for grant 01 P50 AG11715-01.

6.2.2 Mapping Transdisciplinary Tobacco Use Research Centers Publications (forthcoming)

By Angela Zoss & Katy Börner (forthcoming)

This paper reports the results of a scientometric study aimed at evaluating and comparing investigator initiated R01-research and Transdisciplinary Tobacco Use Research Centers (TTURC) funded by the National Institutes of Health (NIH) between 1999 to 2009. Specifically, the study shows the results of a geospatial and topical analysis, a network analysis of collaboration networks, a funding-input vs. publication-output analysis, and a temporal analysis of data trends and coverage. The results were interpreted by tobacco domain experts providing insight into the overall structure and evolution of tobacco research collaborations, interdisciplinary integration, and impact on science as a whole. The study complements efforts that use the very same controlled R01-TTURC dataset in two ways: First, it shows major differences in collaboration patterns and topical coverage of TTURC funded projects. TTURC co-author networks have small world characteristics making them robust to the frequent movement of people in academia while supporting efficient diffusion of information and expertise. Second, TTURC projects by design have a larger topic coverage and wider spectrum of basic to applied research and practice. This is reflected in their topic coverage but also in the ratio of funding-input versus paper-output – TTURC output goes beyond simply producing papers. We conclude with recommendations on how to improve future evaluations of transdisciplinary research centers.



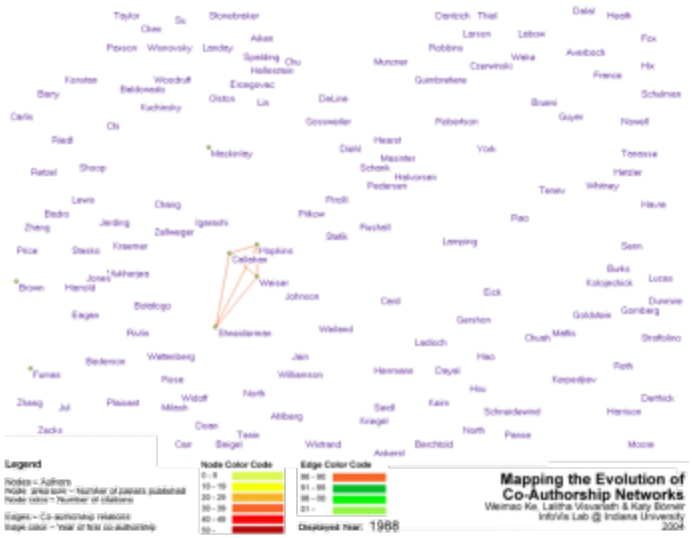
Compare R01 investigator based funding with TTURC Center awards in terms of number of publications and evolving co-author networks.

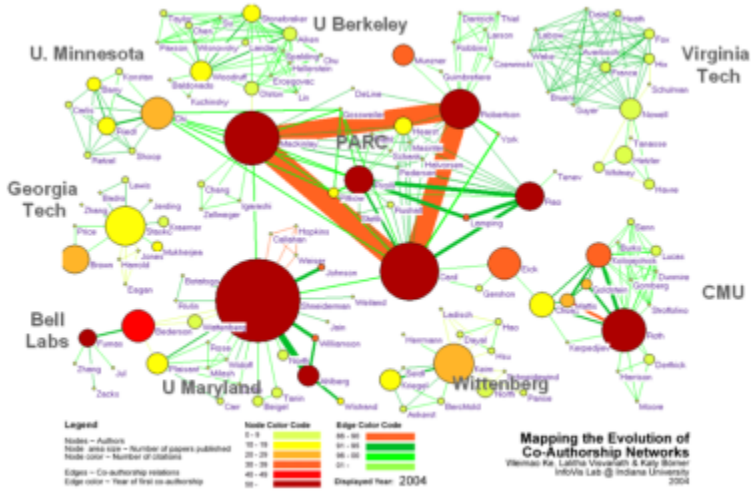
6.3 Local and Global Science Studies

6.3.1 Mapping the Evolution of Co-Authorship Networks (2004)

By Weimo Ke, Katy Börner, & Lalitha Viswanath (2004)

The presented work aims to identify major papers and their interrelations, topic trends over time, as well as major authors and their evolving co-authorship networks in the IV Contest 2004 data set. Paper-citation, co-citation, word co-occurrence, burst analysis and co-author analysis were used to analyze the data set. The results are visually presented as graphs, static Pajek visualizations, and animated network layouts.





Mapping the Evolution of Co-Authorship Networks

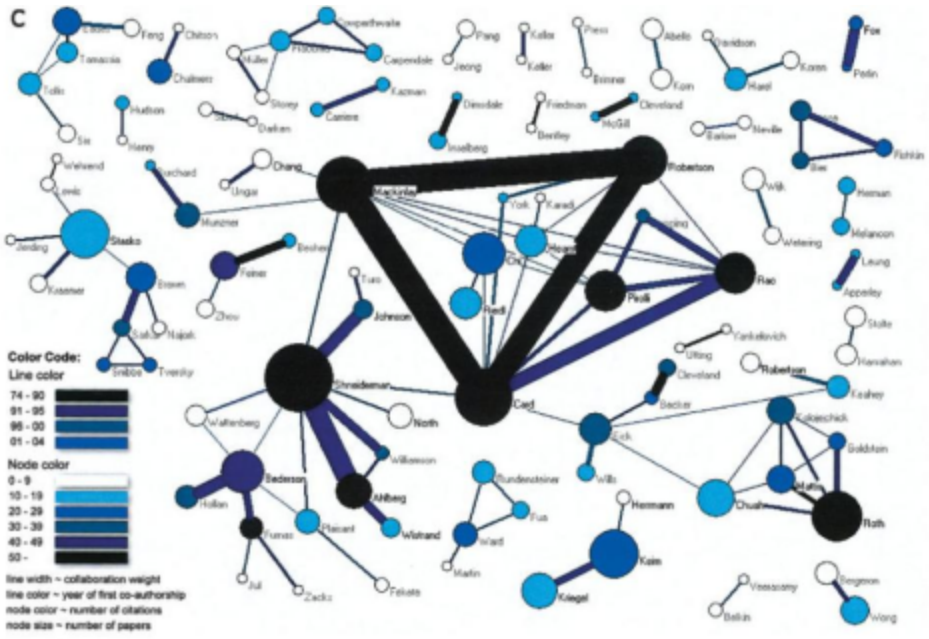
Animated gif is at http://iv.cns.iu.edu/ref/iv04contest/Ke-Borner-Viswanath.gif.

6.3.2 Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams (2005)

By Katy Börner, Luca Dall'Asta, Weimo Ke, & Alessandro Vespignani (2005)

This article introduces a suite of approaches and measures to study the impact of co-authorship teams based on the number of publications and their citations on a local and global scale. In particular, we present a novel weighted graph representation that encodes coupled author-paper networks as a weighted co-authorship graph. This weighted graph representation is applied to a dataset that captures the emergence of a new field of science and comprises 614 articles published by 1036 unique authors between 1974 and 2004. To characterize the properties and evolution of this field, we first use four different measures of centrality to identify the impact of authors. A global statistical analysis is performed to characterize the distribution of paper production and paper citations and its correlation with the co-authorship team size. The size of co-authorship clusters over time is examined. Finally, a novel local, author-centered measure based on entropy is applied to determine the global evolution of the field and the identification of the contribution of a single author's impact across all of its co-authorship relations. A

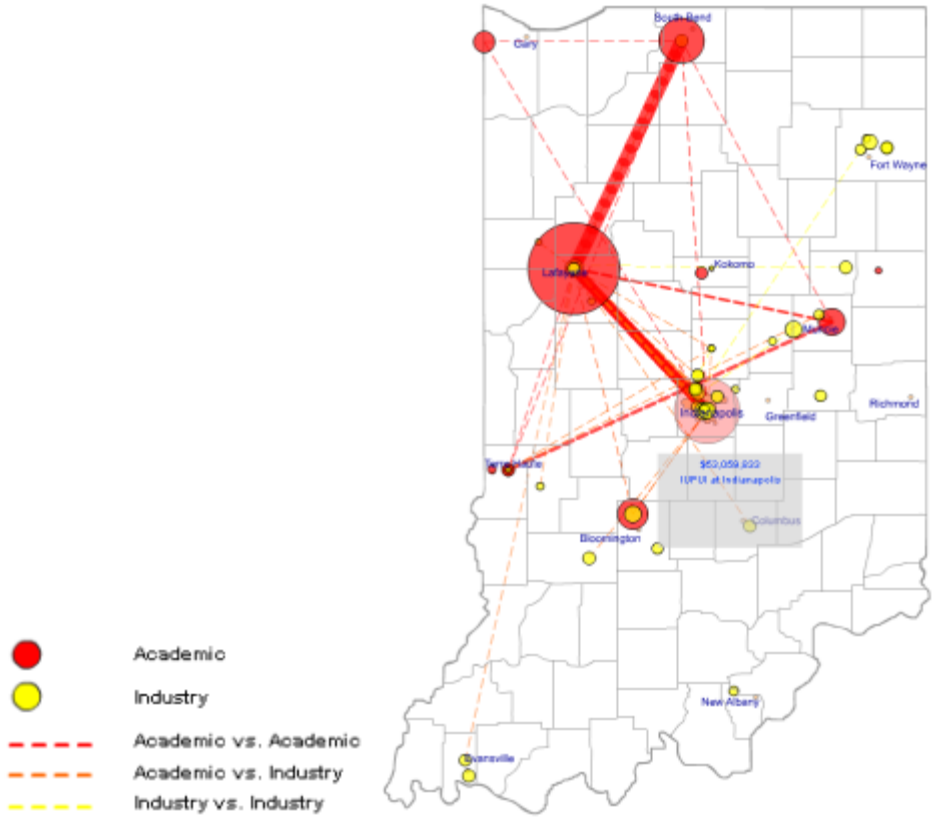
visualization of the growth of the weighted co-author network, and the results obtained from the statistical analysis indicate a drift toward a more cooperative, global collaboration process as the main drive in the production of scientific knowledge.



Weighted co-author network for papers published in 74--04

6.3.3 Mapping Indiana's Intellectual Space

This project aimed to identify pockets of innovation, pathways that ideas take to make it into products, and existing academia-industry collaborations. Submitted and awarded proposals for 2001-2006 were overlaid on a map of Indiana. Geolocations of academic investigators are given in red, industry collaborators are in yellow. Circle size denotes total award amount per geolocation. Linkages are color and line coded to distinguish within academia, within industry, and academia-industry collaborations. The interactive interface supports the selection of different years resulting in year-specific data overlays; clicking on any circle which brings up a table with all proposals and awards for this geolocation together with their titles, investigators and dollar amounts.



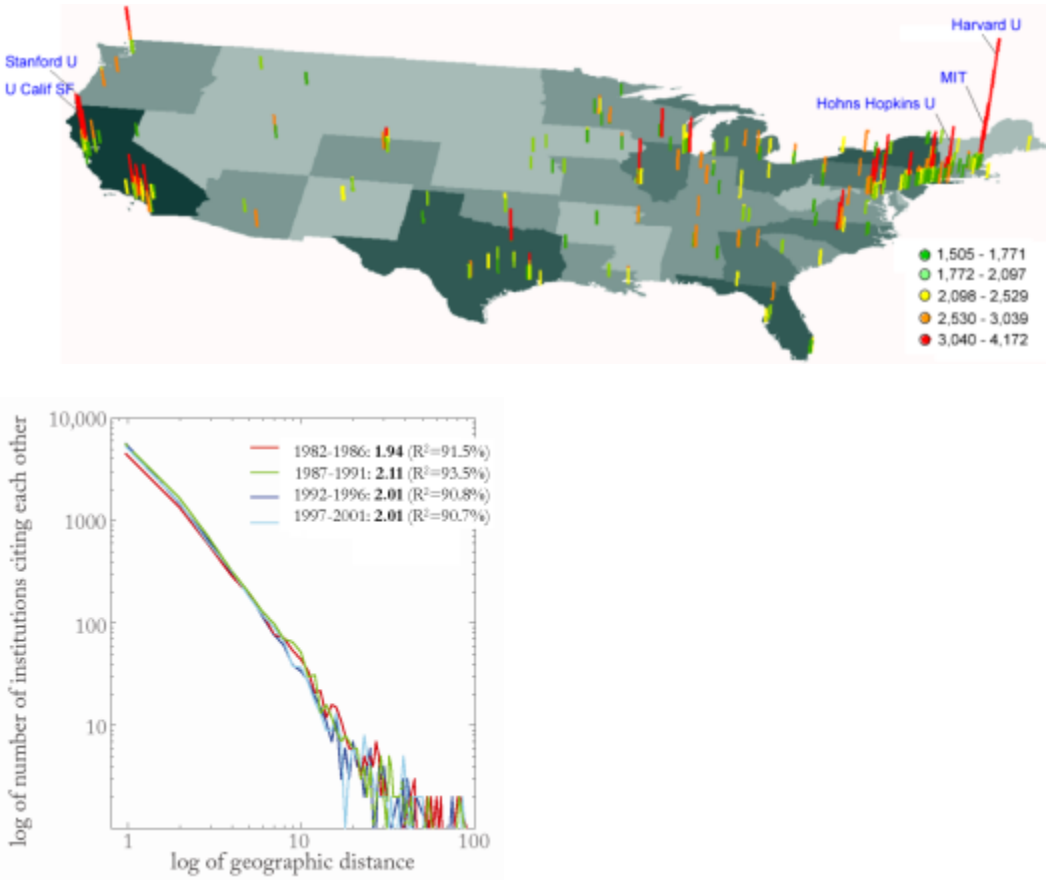
Submitted and Awarded Proposals in Indiana (2001-2006)

Reference: Unpublished.

6.3.4 Mapping the Diffusion of Information Among Major U.S. Research Institutions (2006)

By Katy Börner, Shashikant Penumarthy, Mark Meiss, & Weimao Ke (2006)

This paper reports the results of a large scale data analysis that aims to identify the information production and consumption among top research institutions in the United States. A 20-year publication data set was analyzed to identify the 500 most cited research institutions and spatio-temporal changes in their inter-citation patterns. A novel approach to analyzing the dual role of institutions as information producers and consumers and to study the diffusion of information among them is introduced. A geographic visualization metaphor is used to visually depict the production and consumption of knowledge. The highest producers and their consumers as well as the highest consumers and their producers are identified and mapped. Surprisingly, the introduction of the Internet does not seem to affect the distance over which information diffuses as manifested by citation links. The citation linkages between institutions fall off with the distance between them, and there is a strong linear relationship between the log of the citation counts and the log of the distance. The paper concludes with a discussion of these results and an outlook for future work.



Geographic location and number of received citations for the top 500 institutions (top) and log-log plot showing the variation of the number of institutions that cite each other over geographic distance among them for each of the four time slices. The distance was calculated by applying the Euclidean form formulae to xy coordinates obtained using the Albers projection. 1.5 units of geographic distance equal approximately 100 miles (bottom).

6.3.5 Research Collaborations by the Chinese Academy of Sciences (2009)

By Weixia (Bonnie) Huang, Russell J. Duhon, Elisha F. Hardy, & Katy Börner (2009)

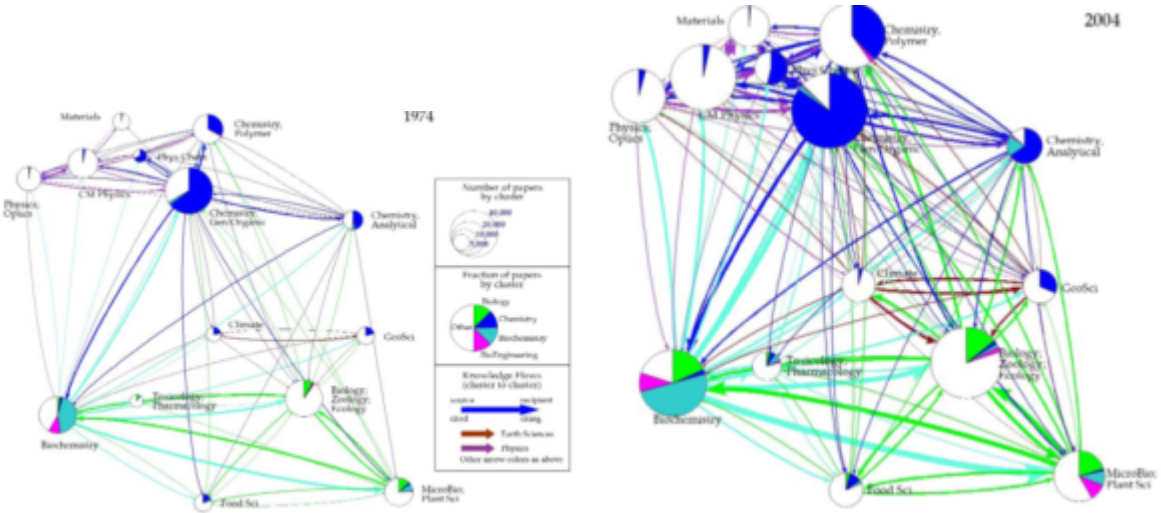
This map highlights the research co-authorship collaborations of the Chinese Academy of Sciences with locations in China and countries around the world. The large geographic map shows the research collaborations of all CAS institutes. Each smaller geographic map shows the research collaborations by the CAS researchers in one province-level administrative division. Collaborations between CAS researchers are not included in the data. On each map, locations are colored on a logarithmic scale by the number of collaborations from red to yellow. The darkest red is 3,395 collaborations by all of CAS with researchers in Beijing. Also, flow lines are drawn from the location of focus to all locations collaborated with. The width of the flow line is linearly proportional to the number of collaborations with the locations it goes to, with the smallest flow lines representing one collaboration and the largest representing differing amounts on each geographic map.

Collaboration and knowledge diffusion via co-author networks

6.3.6 Mapping the Structure and Evolution of Chemistry Research (2009)

By Kevin W. Boyack, Katy Börner, & Richard Klavans (2009)

How does our collective scholarly knowledge grow over time? What major areas of science exist and how are they interlinked? Which areas are major knowledge producers; which ones are consumers? Computational scientometrics – the application of bibliometric/scientometric methods to large-scale scholarly datasets – and the communication of results via maps of science might help us answer these questions. This paper represents the results of a prototype study that aims to map the structure and evolution of chemistry research over a 30 year time frame. Information from the combined Science (SCIE) and Social Science (SSCI) Citations Indexes from 2002 was used to generate a disciplinary map of 7,227 journals and 671 journal clusters. Clusters relevant to study the structure and evolution of chemistry were identified using JCR categories and were further clustered into 14 disciplines. The changing scientific composition of these 14 disciplines and their knowledge exchange via citation linkages was computed. Major changes on the dominance, influence, and role of Chemistry, Biology, Biochemistry, and Bioengineering over these 30 years are discussed. The paper concludes with suggestions for future work.



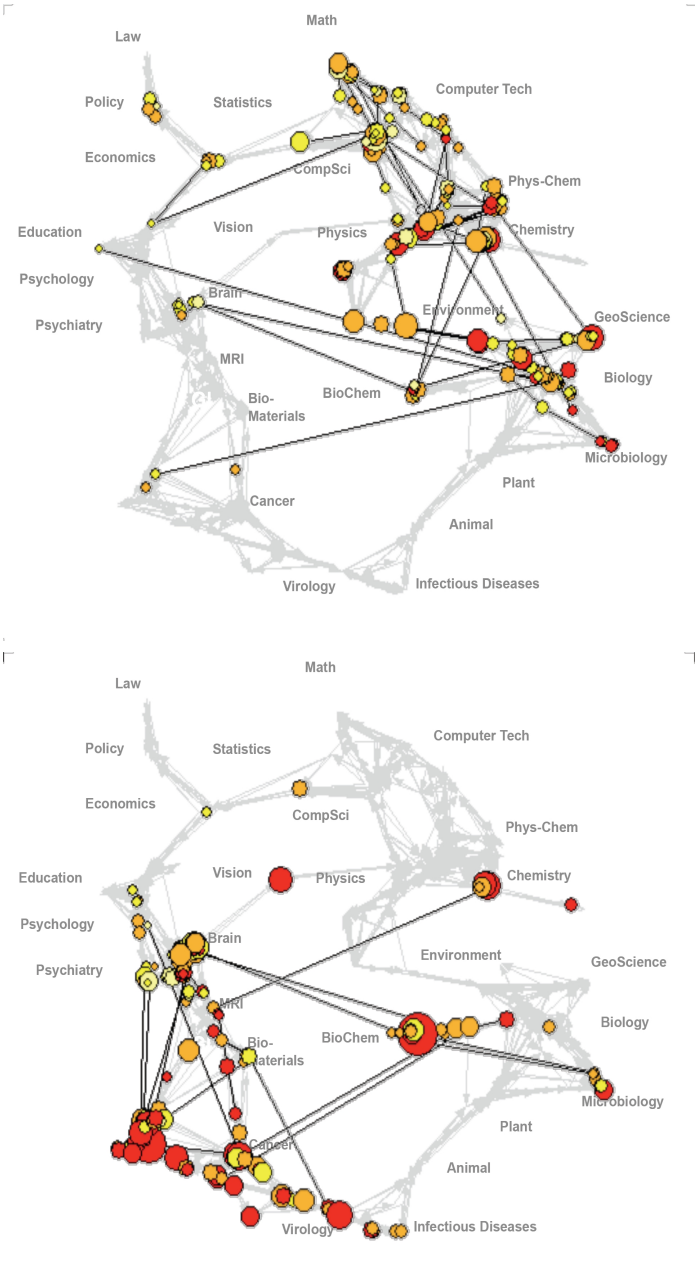
Map of the 14 disciplines, fractions of papers by field for each discipline, and knowledge flows between disciplines for 1974 (left) and 2004 (right). These 14 disciplines are further aggregated into six groups, represented by the 6 colors shown in the legend.

6.3.7 Science Map Applications: Identifying Core Competency (2007)

By Kevin W. Boyack, Katy Börner, & Richard Klavans (2009)

The 2002 base map represents journal cluster interrelations but is invariant to rotation and mirroring. The map was oriented to place mathematics at the top and the physical sciences on the right. The ordering of disciplines is similar

to what has been shown in other maps of science: as one progresses clockwise around the map, one progresses from mathematics through the physical sciences (Engineering, Physics, Chemistry), to the earth sciences, life sciences, medical sciences, and social sciences. The social sciences link back to computer science (near the top of the map), which has strong linkages to mathematics and engineering.
Just like a map of the world can be used to communicate the location of minerals, soil types, political boundaries, population densities, etc., a map of science can be used to locate the position of scholarly activity. The profiles for the U.S. NIH (National Institutes of Health) and NSF (National Science Foundation) are shown below, and were calculated by matching the principal investigators and their institutions from grants funded in 1999 to first authors and institutions of papers indexed in 2002. This type of paper-to-grant matching will produce some false positives. On the whole, however, it is a conservative approach in that it only considers a single time-lag between funding and publication (3 years in this case), and it does not match on secondary authors. The 14,367 NIH matches and 10,054 NSF matches are large samples, ensuring that the aggregated profiles are representative of the actual funding profiles of the agencies. It serves as a good example of how journal level, or disciplinary, maps can be used to display aggregated information obtained from paper-level analysis.

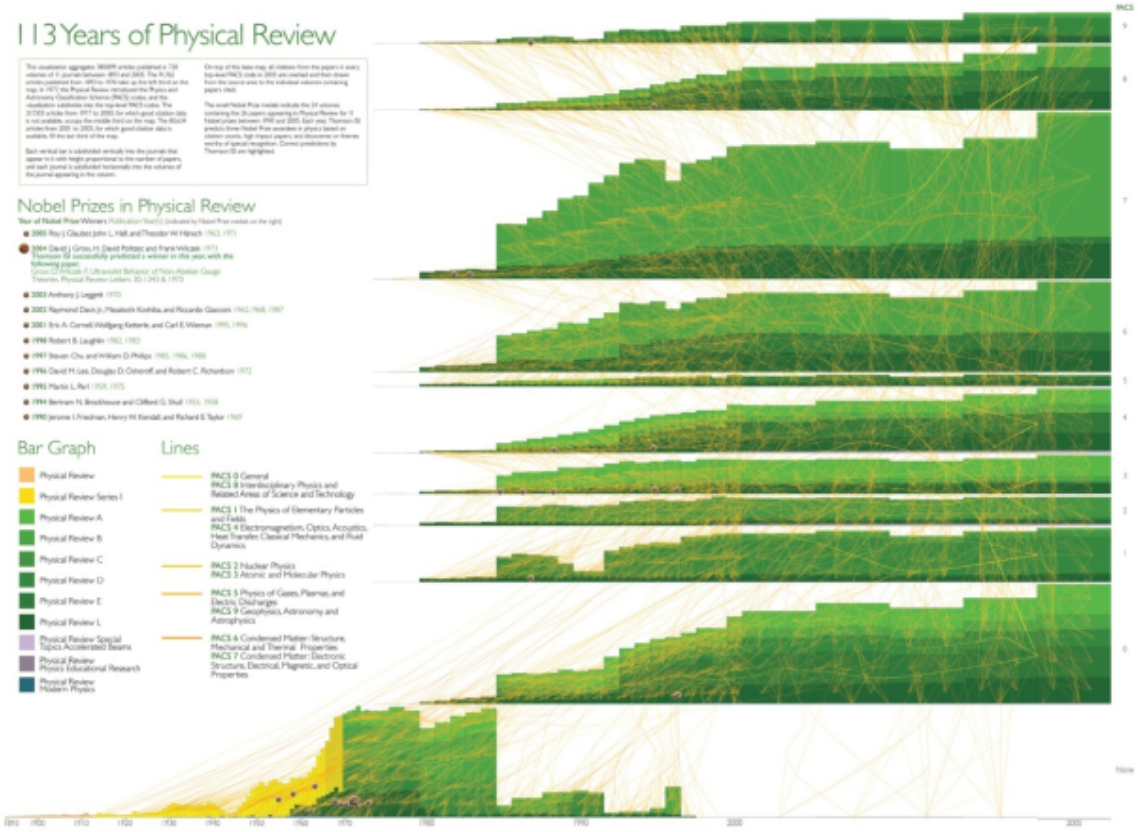


Funding Patterns of the National Science Foundation (left) and the National Institute of Health (right).

6.4 Modeling Science

6.4.1 113 Years of Physical Review: Using Flow Maps to Show Temporal and Topical Citation (2008)

By Bruce W. Herr II, Russell Jackson Duhon, Katy Börner, Elisha F. Hardy, & Shashikant Penumarthy (2008)
We visualize 113 years of bibliographic data from the American Physical Society. The 389,899 documents are laid out in a two dimensional time-topic reference system. The citations from 2005 papers are overlaid as flow maps from each topic to the papers referenced by papers in the topic making intercitation patterns between topic areas visible. Paper locations of Nobel Prize predictions and winners are marked. Finally, though not possible to reproduce here, the visualization was rendered to, and is best viewed on, a 24" x 30" canvas at 300 dots per inch.

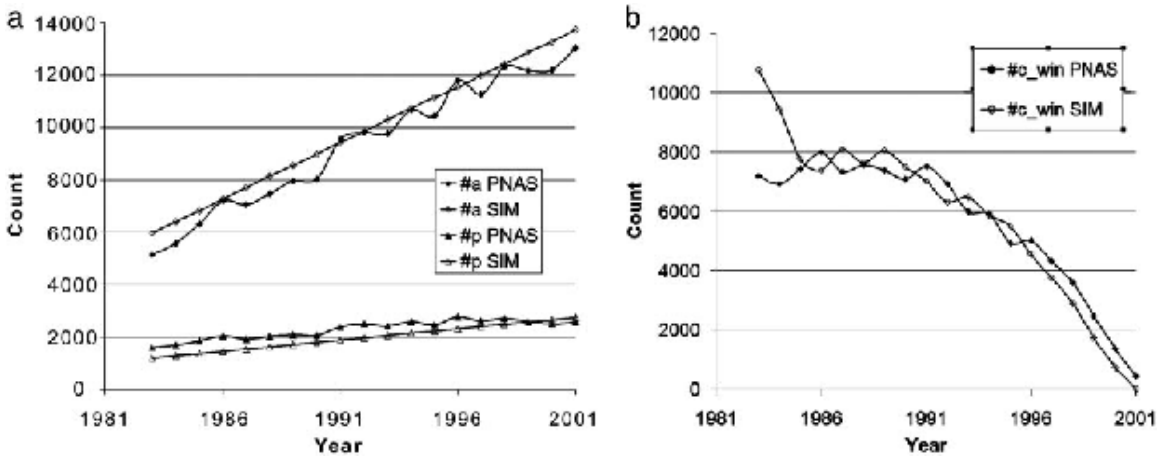


113 Years of bibliogaphic data from the American Physical Society

6.4.2 The Simultaneous Evolution of Author and Paper Networks (2004)

By Katy Börner, Jeegar Maru, & Robert Goldstone (2004)
There has been a long history of research into the structure and evolution of mankind's scientific endeavor. However, recent progress in applying the tools of science to understand science itself has been unprecedented because only recently has there been access to high-volume and high-quality data sets of scientific output (e.g., publications, patents, grants) and computers and algorithms capable of handling this enormous stream of data. This article reviews major work on models that aim to capture and recreate the structure and dynamics of scientific evolution. We then introduce a general process model that simultaneously grows coauthor and paper citation networks. The statistical and dynamic properties of the networks generated by this model are validated against a 20-year data set of articles published in PNAS. Systematic deviations from a power law distribution of citations to papers are well fit by a model that incorporates a partitioning of authors and papers into topics, a bias for authors to

cite recent papers, and a tendency for authors to cite papers cited by papers that they have read. In this TARL model (for topics, aging, and recursive linking), the number of topics is linearly related to the clustering coefficient of the simulated paper citation network.



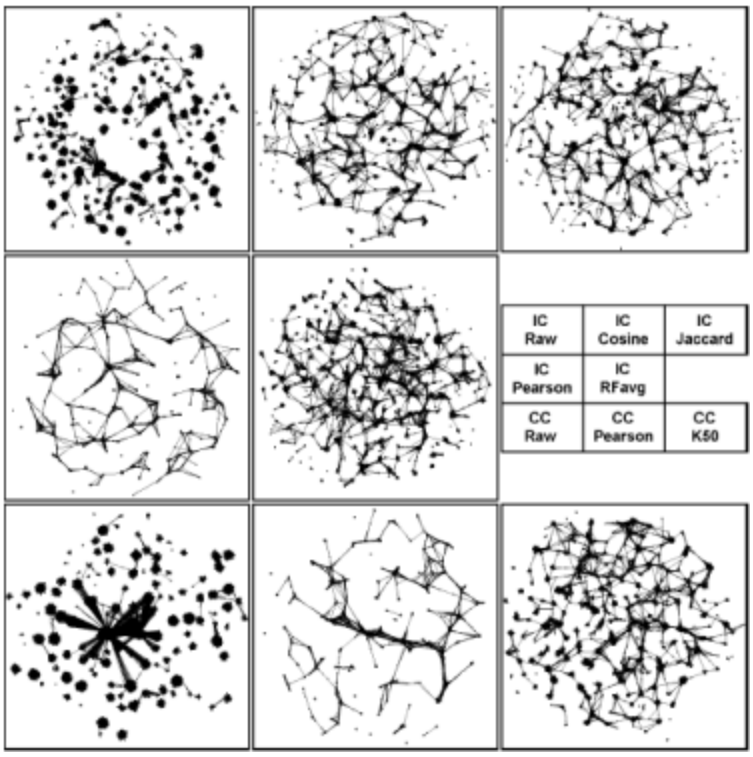
Total number of actual and simulated papers (#p) and authors (#a) (a) and received citations (#c_win) (b).

6.5 Accuracy Studies

6.5.1 Mapping the Backbone of Science (2005)

By Kevin W. Boyack, Richard Klavans, & Katy Börner (2005)

This paper presents a new map representing the structure of all of science, based on journal articles, including both the natural and social sciences. Similar to cartographic maps of our world, the map of science provides a bird's eye view of today's scientific landscape. It can be used to visually identify major areas of science, their size, similarity, and interconnectedness. In order to be useful, the map needs to be accurate on a local and on a global scale. While our recent work has focused on the former aspect,1 this paper summarizes results on how to achieve structural accuracy.
Eight alternative measures of journal similarity were applied to a data set of 7,121 journals covering over 1 million documents in the combined Science Citation and Social Science Citation Indexes. For each journal similarity measure we generated two-dimensional spatial layouts using the force-directed graph layout tool, VxOrd. Next, mutual information values were calculated for each graph at different clustering levels to give a measure of structural accuracy for each map. The best co-citation and inter-citation maps according to local and structural accuracy were selected and are presented and characterized. These two maps are compared to establish robustness. The inter-citation map is then used to examine linkages between disciplines. Biochemistry appears as the most interdisciplinary discipline in science.

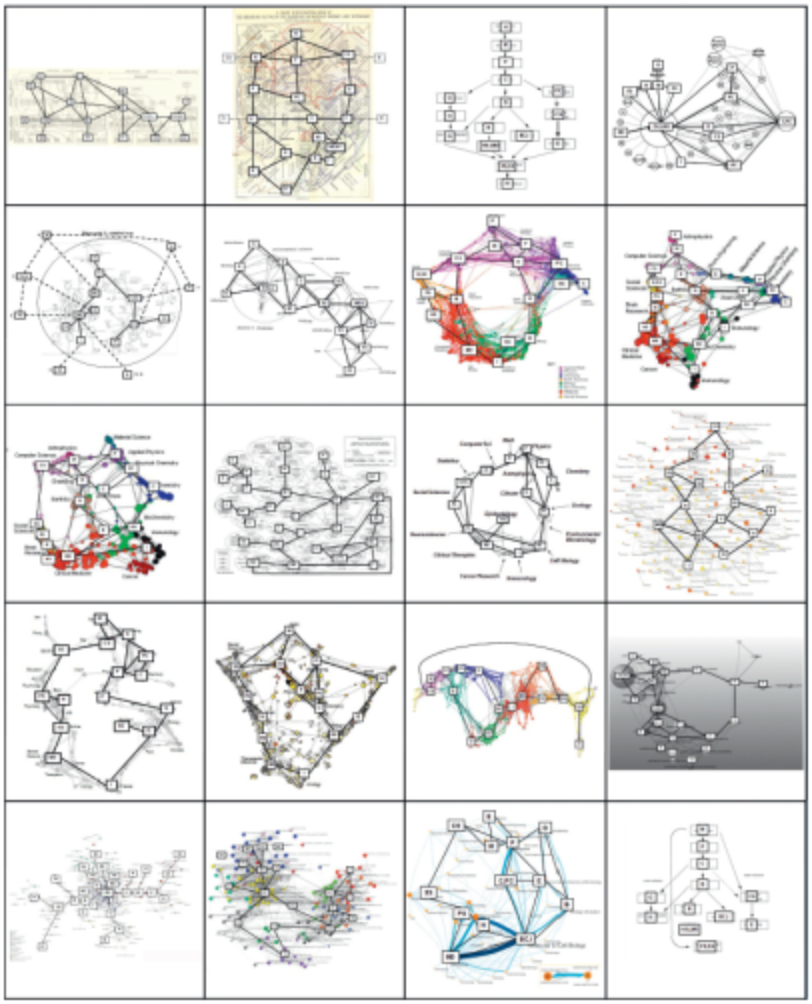


Maps of science generated from eight different journal-journal similarity measures. Dots represent journals. Lines represent the edges remaining at the end of the VxOrd runs. Similarity measures corresponding to the various map panels are listed in the middle right panel.

6.5.2 Toward a Consensus Map of Science (2009)

By Richard Klavans & Kevin Boyack (2009)

A consensus map of science is generated from an analysis of 20 existing maps of science. These 20 maps occur in three basic forms: hierarchical, centric, and noncentric (or circular). The consensus map, generated from consensus edges that occur in at least half of the input maps, emerges in a circular form. The ordering of areas is as follows: mathematics is (arbitrarily) placed at the top of the circle, and is followed clockwise by physics, physical chemistry, engineering, chemistry, earth sciences, biology, biochemistry, infectious diseases, medicine, health services, brain research, psychology, humanities, social sciences, and computer science. The link between computer science and mathematics completes the circle. If the lowest weighted edges are pruned from this consensus circular map, a hierarchical map stretching from mathematics to social sciences results. The circular map of science is found to have a high level of correspondence with the 20 existing maps, and has a variety of advantages over hierarchical and centric forms. A onedimensional Riemannian version of the consensus map is also proposed.



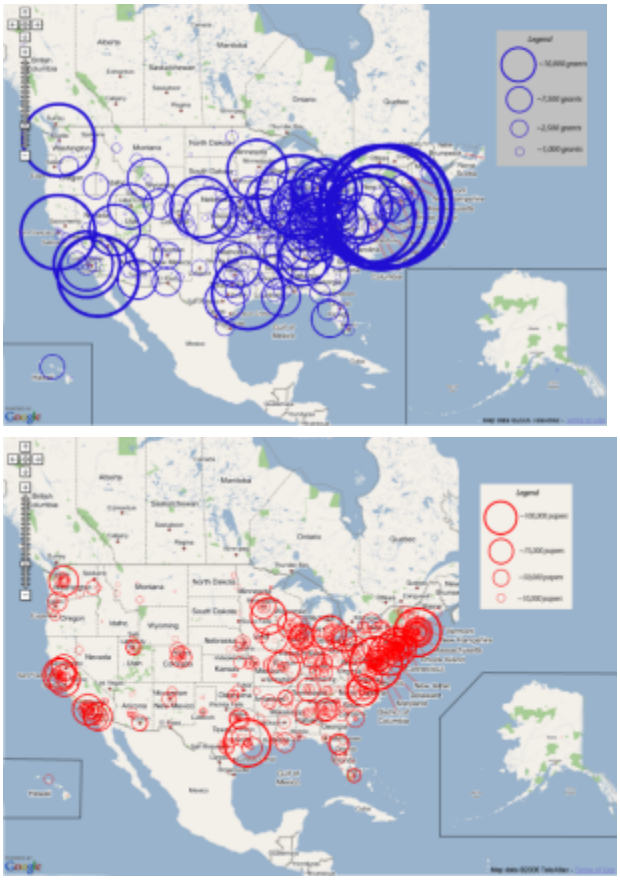
Images of the 20 maps of science that were used in this study along with their codings. The 20 maps are shown in the same order in which they are listed in Table 1, from upper left to lower right.

6.6 Databases and Tools

6.6.1 The Scholarly Database and Its Utility for Scientometrics Research (2009)

By Gavin LaRowe, Sumeet Ambre, John Burgoon, Weimao Ke, & Katy Börner (2009)

The Scholarly Database aims to serve researchers and practitioners interested in the analysis, modeling, and visualization of large-scale data sets. A specific focus of this database is to support macro-evolutionary studies of science and to communicate findings via knowledge-domain visualizations. Currently, the database provides access to about 18 million publications, patents, and grants. About 90% of the publications are available in full text. Except for some datasets with restricted access conditions, the data can be retrieved in raw or pre-processed formats using either a web-based or a relational database client. This paper motivates the need for the database from the perspective of bibliometric/scientometric research. It explains the database design, setup, etc., and reports the temporal, geographical, and topic coverage of data sets currently served via the database. Planned work and the potential for this database to become a global testbed for information science research are discussed at the end of the paper.

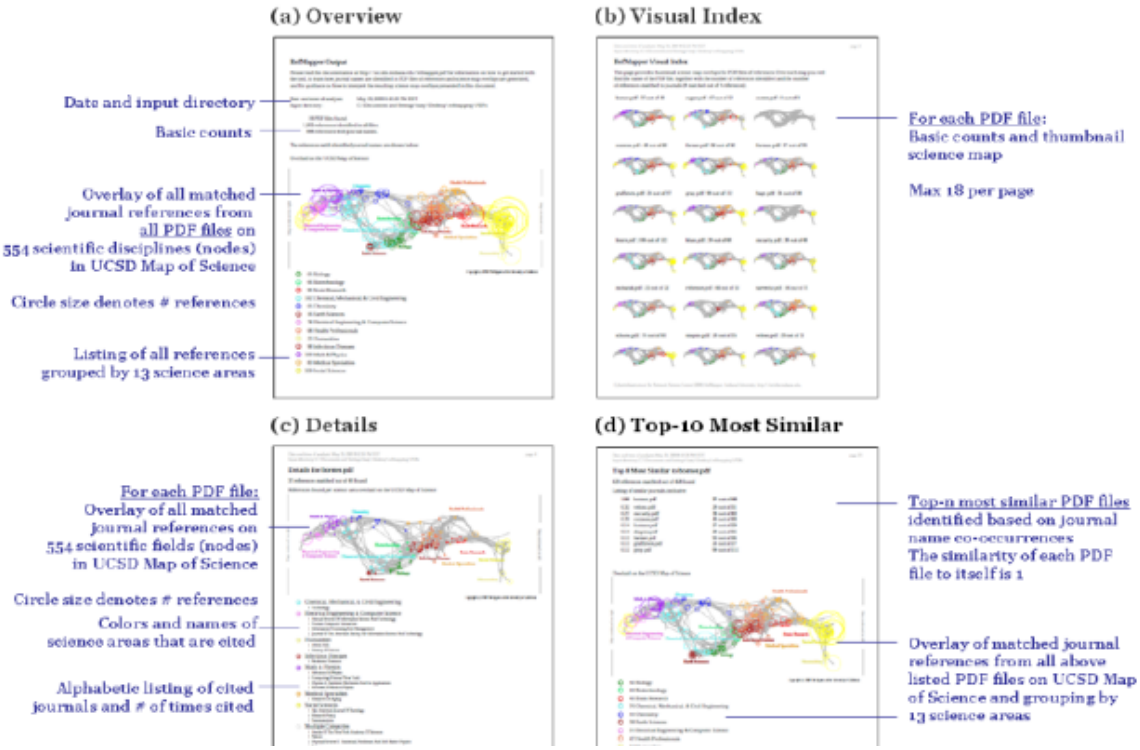


Map of NIH Grants (top) and MEDLINE Publications (bottom)

6.6.2 Reference Mapper

By Russel J. Duhon, Katy Börner (2007)

The RefMapper tool supports the automatic detection, mapping, and clustering of grant awards and proposals based on citation references. It might be used to group proposals for review or to communicate the topic coverage of a proposal/funding portfolio. The tool uses a master list of 18,351 journal names that are indexed by Scopus and Reuters/Thomson Scientific (ISI SCI, SSCI, and A&H Indexes) and a lookup table of 57,860 different abbreviations for those journal names. It science-locates identified journals on the 554 scientific areas of the UCSD Map of Science (Klavans, Boyack, 2007). Each of the 13 main scientific disciplines is labeled and color coded in a metaphorical way, e.g., Medicine is blood red and Earth Sciences are brown as soil. The RefMapper also identifies clusters based on reference co-occurrence similarity. The RefMapper tool was made available as a plugin to the Network Workbench (NWB Team, 2006; Cyberinfrastructure for Network Science Center, 2009). It can be downloaded for Windows and for Mac.

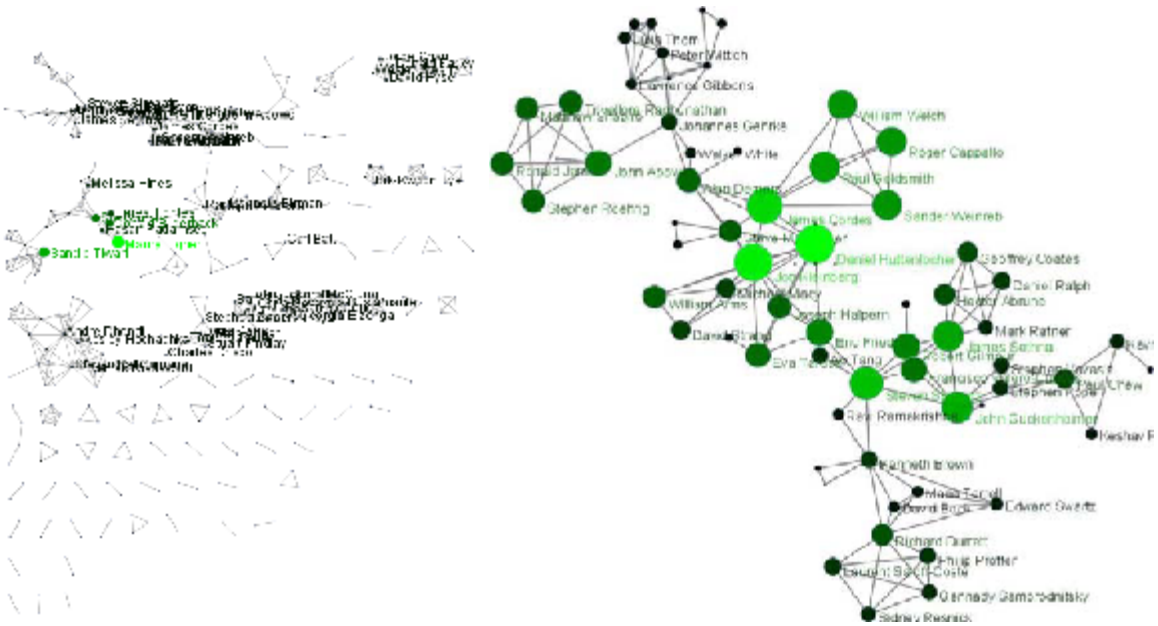


Presentation of RefMapper analysis results

6.6.3 Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Scholarly Database and the Network Workbench Tool (2009)

By Katy Börner, Bonnie (Weixia) Huang, Micah Linnemeier, Russell J. Duhon, Patrick Phillips, Ninali Ma, Angela Zoss, Hanning Guo, & Mark A. Price (2009)

The enormous increase in digital scholarly data and computing power combined with recent advances in text mining, linguistics, network science, and scientometrics make it possible to scientifically study the structure and evolution of science on a large scale. This paper discusses the challenges of this 'BIG science of science' – also called 'computational scientometrics' research – in terms of data access, algorithm scalability, repeatability, as well as result communication and interpretation. It then introduces two infrastructures: (1) the Scholarly Database (SDB) (http://sdb.cns.iu.edu), which provides free online access to 20 million scholarly records – papers, patents, and funding awards which can be cross-searched and downloaded as dumps, and (2) Scientometrics-relevant plug-ins of the open-source Network Workbench (NWB) Tool (http://nwb.cns.iu.edu). The utility of these infrastructures is then exemplarily demonstrated in three studies: a comparison of the funding portfolios and co-investigator networks of different universities, an examination of paper-citation and co-author networks of major network science researchers, and an analysis of topic bursts in streams of text. The paper concludes with a discussion of related work that aims to provide practically useful and theoretically grounded cyberinfrastructure in support of computational scientometrics research, practice, and education.



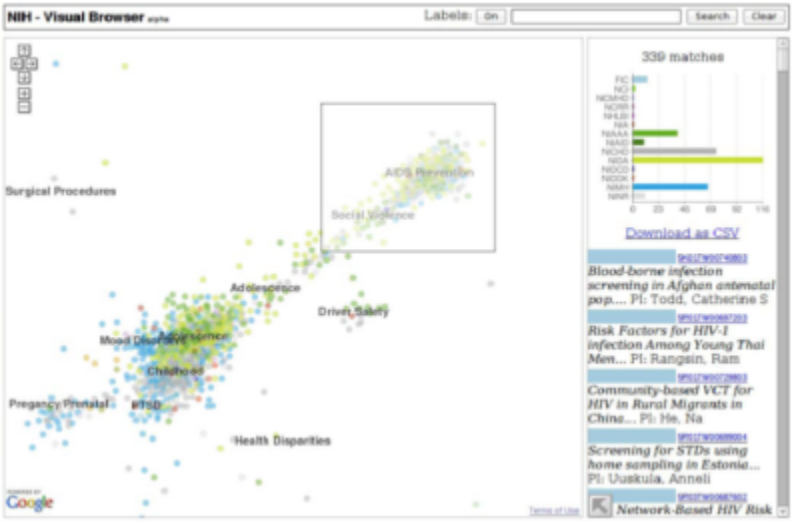
Complete network (left) and largest component (right) of Cornell University's co-investigator network (67 nodes).

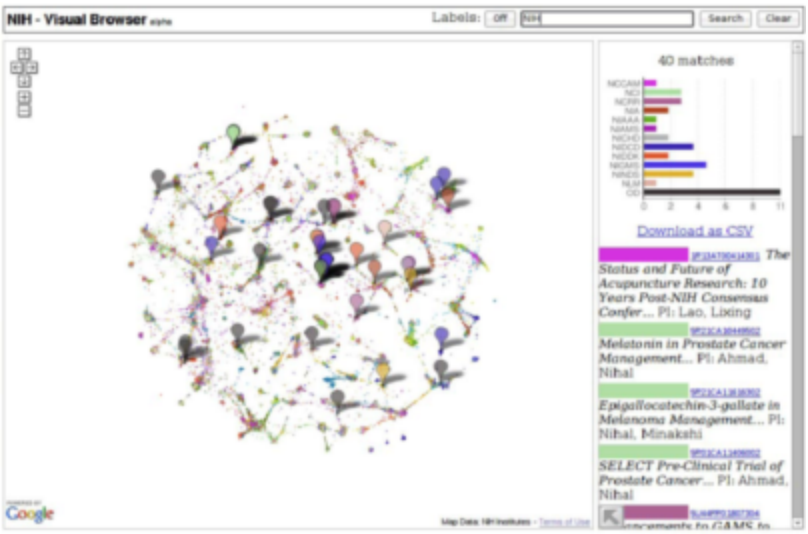
6.7 Interactive Online Services

6.7.1 The NIH Visual Browser: An Interactive Visualization of Biomedical Research (2009)

By Bruce W. Herr II, Edmund M. Talley, Gully A.P.C. Burns, David Newman, & Gavin LaRowe (2009)

This paper presents a technical description of the methods used to generate an interactive, two-dimensional visualization of 60,568 grants funded by the National Institutes of Health in 2007. The visualization is made intelligible by providing interactive features for assessing the data in a web-based visual browser, see http://www.nihmaps.org. The key features include deep zooming, selection, full-text querying, overlays, color-coding schemes, and multi-level labeling. Major insights, broader applicability, and future directions are discussed.



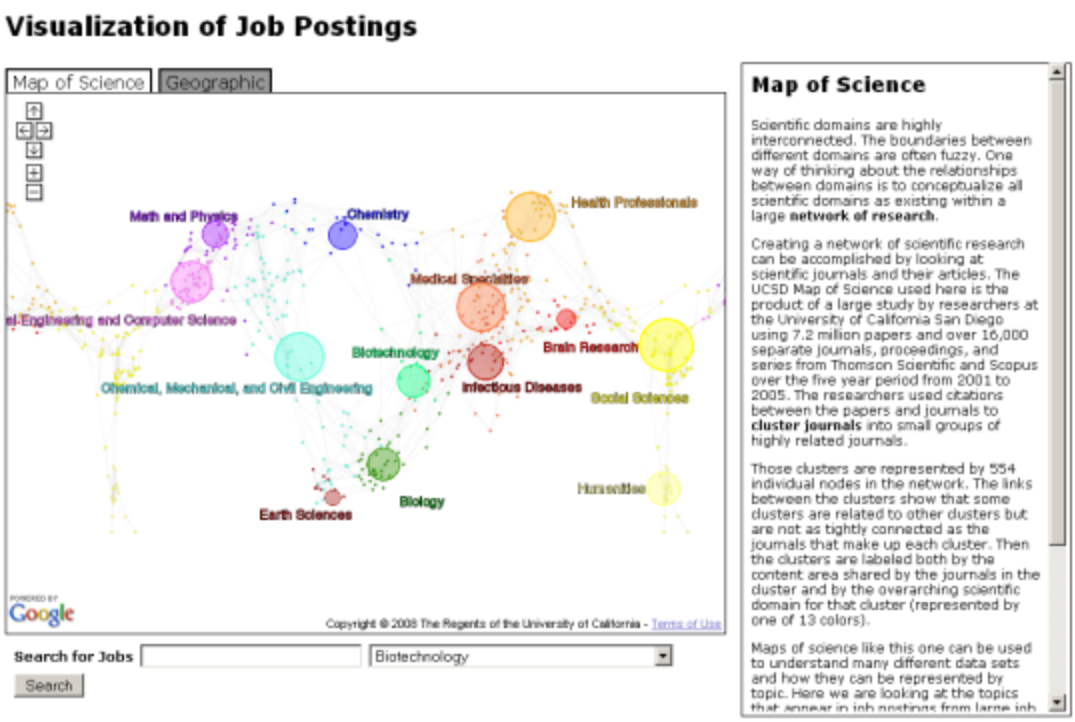


Cluster selection with results shown in the right-hand column (top) and Query for 'NIH' in the title with results shown on the map and in the right-hand column (bottom).

6.7.2 Interactive World and Science Map of S&T Jobs (2010)

By Angela Zoss, Michael Connover, Katy Börner (2010)

This paper details a methodology for capturing, analyzing, and communicating one specific type of real time data: advertisements of currently available academic jobs. The work was inspired by the American Recovery and Reinvestment Act of 2009 (ARRA) that provides approximately $100 billion for education, creating a historic opportunity to create and save hundreds of thousands of jobs. Here, we discuss methodological challenges and practical problems when developing interactive visual interfaces to real time data streams such as job advertisements. Related work is discussed, preliminary solutions are presented, and future work is outlined. The presented approach should be valuable to deal with the enormous volume and complexity of social and behavioral data that evolve continuously in real time, and analyses of them need to be communicated to a broad audience of researchers, practitioners, clients, educators, and interested policymakers, as originally suggested by Hemmings and Wilkinson.



High level view of the Map of Science visualization. The map is circular, so areas of the map are repeated side to side as users scroll back and forth. Postings are clustered by 13 main scientific domains at the high zoom level and the 554 subdisciplines at the lower zoom level.

7 Extending the Sci2 Tool

7.1 CIShell Basics

The CIShell Platform has been specifically design around the idea of the algorithm. It is the central and most important concept. Algorithms are fully defined and self-contained bits of execution. They can do many things: data conversion, data analysis, and even spawn whole external programs if needed. Algorithms are well defined black boxes, which can contain either Java code or any program which can be compiled. Creating new algorithms is the primary method to extend CIShell tool's functionality, or creating new CIShell-based tools.
CIShell is based on OSGi, which is a plugin and service based framework. Practically this means that OSGi functionality is divided into plugins or bundles (Java jar files with some additional special files), each of which contains code to create some number of services at runtime. These services are the main actors in the OSGi environment. In CIShell almost all services are algorithms, which means they conform to a certain interface, allowing algorithms to interoperate with the CIShell environment and each other in a well-defined way. Specifically, algorithms accept Data[], user-input parameters, and a CIShellContext. They output Data[].

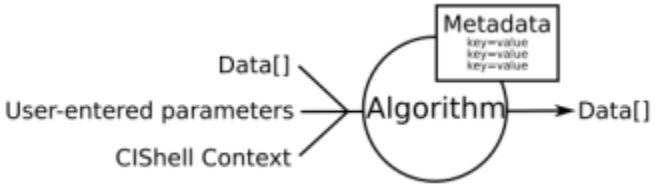


Figure 8.1: Input data, paramters,metadataof dataset and algorithm services

CIShell provides an environment which makes it easy for users to interact with a set of algorithms in the form of an executable tool. This environment includes a Menu Manager which allows users to invoke algorithms, a Data

Manager which serves as workspace to hold data while users run a series of algorithms on it, a scheduler which monitors algorithms as they run, a conversion service which converts data between various types so algorithms can operate on data in the format of their choice, and several other services. Since all of this is provided by default, developers can maintain focus on the algorithms they wish to implement without having to reinvent the entire supporting infrastructure.
Currently the CIShell environment is implemented as an Eclipse-based desktop tool, but the CIShell interfaces are defined in such as way that the environment could be implemented in a variety of ways, e.g., a web-based service. Since CIShell is so closely tied to OSGi, many references will be made to OSGi in the developer documentation. To fully understand the details of how CIShell works, it is often necessary to understand certain aspects of OSGi, however most developers should be able to begin working with CIShell without understanding OSGi.

7.2 Creating and Sharing New Algorithm Plugins

The Sci^2 Tool is an 'empty shell' filled with plugins. Some plugins run on the core architecture, OSGi and CIShell. Others convert loaded data into in-memory objects, formatted for different algorithms to read it. The algorithm plugins themselves can be divided into different menus, in this case data processing, preprocessing, analysis, modeling, visualization, and scientometrics. Users are not limited to using pre-packaged plugins; instead, they can create, download, share, and import their own.
To use an alternative plugin, simply copy the *.jar* file you created or downloaded into '**yoursci2directory/plugins/*' and then look for the plugin's name in the Sci^2 Tool menu structure.

A step-by-step guide to developing new plugins can be found at the CIShell Manual.

7.3 Tools That Use OSGi and_or CIShell

Recently, a number of other efforts adopted OSGi and/or CIShell. Among them are:

Cytoscape (http://www.cytoscape.org) lead by Trey Ideker, UCSD is an open source bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data (Shannon, Markiel et al. 2002).

Taverna Workbench (http://taverna.sourceforge.net) lead by Carol Goble, University of Manchester, UK is a free software tool for designing and executing workflows (Hull, Wolstencroft et al. 2006). Taverna allows users to integrate many different software tools, including over 30,000 web services from many different domains, such as chemistry, music and social sciences. The myExperiment (http://www.myexperiment.org) social web site supports finding and sharing of workflows and has special support for Taverna workflows (De Roure, Goble et al. 2009). Currently, Taverna uses Raven at its core but a reimplementation using OSGi is underway.

MAEviz (https://wiki.ncsa.uiuc.edu/display/MAE/Home) managed by Shawn Hampton, NCSA is an open-source, extensible software platform which supports seismic risk assessment based on the Mid-America Earthquake (MAE) Center research in the Consequence-Based Risk Management (CRM) framework (Elnashai, Spencer et al. 2008). It uses the Eclipse Rich Client Platform (RCP) that includes Equinox, a component framework based on the OSGi standard. The 125 MAEvis plugins consist of 6 core plugins, 7 plugins related to the display of hazard, building, and bridges, and lifeline data, 11 network and social science plugins, and 2 report visualization plugins. Bard (previously NCSA-GIS) has 11 in core plugins, 2 relevant for networks and 10 for visualization. The analysis framework has 6 core plugins. Ogrescript has 14 core plugins. A total of 54 core Eclipse OSGI plugins are used such as org.eclipse.core*, org.eclipse.equinox*, org.eclipse.help*, org.eclipse.osgi*, org.eclipse.ui*, and org.eclipse.update* (https://wiki.ncsa.uiuc.edu/display/MAE/OSGI+Plugins).

TEXTrend (http://www.textrend.org) lead by George Kampis, Eötvös University, Hungary develops a framework for the easy and flexible integration, configuration, and extension of plugin-based components in support of natural language processing (NLP), classification/mining, and graph algorithms for the analysis of business and governmental text corpuses with an inherently temporal component (Kampis, Gulyas et al. 2009). TEXTrends recently adopted OSGi/CIShell for the core architecture and the first seven plugins are IBMs Unstructured

Information Management Architecture (UIMA) (http://incubator.apache.org/uima), the data mining, machine learning, classification and visualization toolset WEKA (http://www.cs.waikato.ac.nz/ml/weka), Cytoscape, Arff2xgmml converter, R (http://www.r-project.org) via iGgraph and scripts (http://igraph.sourceforge.net), and yEd. Upcoming work will focus on integrating the Cfinder clique percolation analysis and visualization tool (http://www.cfinder.org), workflow support, and web services.

Note that the Sci^2 Tool uses plugins from several other efforts/tools such as the Information Visualization cyberinfrastructure (http://iv.cns.iu.edu), the Network Workbench (http://nwb.cns.iu.edu), and TEXTrend. As the functionality of OSGi/CIShell-based software frameworks improves and the number and diversity of dataset and algorithm plugins increases, the capabilities of custom tools will expand.

8 Relevant Datasets and Tools

8.1 Datasets

Aggregate

- NanoBank – http://www.nanobank.org/
- NWB Datasets - https://nwb.cns.iu.edu/community/?n=Datasets.HomePage
- Scholarly Database – http://sdb.cns.iu.edu/

Publications

- Google Scholar - http://scholar.google.com/
- ISI Web of Knowledge, Web of Science - http://apps.isiknowledge.com/
- JSTOR - http://www.jstor.org/
- Ley, Michael. The DBLP Computer Science Bibliography. Universität Trier.
- National Federation of Abstracting and Information Services. (1990). NFAIS abstract dataset, 1957-1990. Available at http://www.nfais.org/
- Office of Inspector General - http://www.oig.hhs.gov/fraud/exclusions.asp
- PsychInfo - http://www.apa.org/psycinfo/
- PubMed - http://www.ncbi.nlm.nih.gov/pubmed/
- PubMed Central - http://www.pubmedcentral.nih.gov/
- Research Papers in Economics - http://repec.org/
- Scopus - http://www.scopus.com/
- Stanford Linear Accelerator Center. (2009). SPIRES-HEP Database SPIRES-HEP - http://www.slac.stanford.edu/spires/

Patents

- École Polytechnique Fédérale de Lausanne. (2009). CEMI's PATSTAT Knowledge Base. http://wiki.epfl.ch/patstat
- EPO - http://www.epo.org/patents/patent-information/subscription/gpi.html
- Patent Lens: Initiative for Open Innovation. http://www.patentlens.net/daisy/patentlens/patentlens.html
- PATSTAT - http://wiki.epfl.ch/patstat/whatis
- Public Intellectual Property Resource - http://search.pipra.org/

Intellectual Property

- SparkIP. (2007). Spark-IP: The World's Leading IP Research and Marketplace Platform. http://www.sparkip.com
- Funding

- eJacket - https://www.ejacket.nsf.gov/
- Environmental Impact Database - http://www.epa.gov/oecaerth/nepa/eisdata.html
- National Science Foundation. Find Funding - http://nsf.gov/funding/
- NEH Funded Projects Query - https://securegrants.neh.gov/publicquery/main.aspx
- NIH RePORTER - http://projectreporter.nih.gov/reporter.cfm
- Research.gov - http://www.research.gov/
- USAspending.gov - http://www.usaspending.gov/

Federal Reports

- EuroStats - http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/
- National Academies Reports - http://www.nationalacademies.org/publications/
- OECD Statistics - www.oecd.org/statistics
- Science and Engineering Indicators 2006. Arlington, VA: National Science Foundation.
- SRS S&E - http://www.nsf.gov/statistics/

Surveys

- Taulbee Survey of CS Salaries - http://www.cra.org/statistics/

Science Databases

- FAO - http://www.fao.org/agris/search/search.do
- NCBI Plant Genomes - http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html
- NCBI: GenBank. http://www.ncbi.nlm.nih.gov/Genbank
- TAIR - http://www.arabidopsis.org/

Other

- Carter, Susan B., Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, Gavin Wright (Eds.). (2006). The Historical Statistics of the United States (Millennial Edition ed.): Cambridge University Press. http://hsus.cambridge.org/HSUSWeb/HSUSEntryServlet (accessed on 4/29/2009)
- State of Utah. (2009). Foreign Labor Certification Data Center: Online Wage Library. http://www.flcdatacenter.com/ (accessed on 5/1/2009).
- The University of California, Davis. (2009). The Public Intellectual Property Resource for Agriculture. http://www.pipra.org/ (accessed on 4/30/2009).
- United States Central Intelligence Agency. (2008). The World Factbook: United States. https://www.cia.gov/library/publications/the-world-factbook/print/us.html (accessed on 9/29/2008).
- World Intellectual Property Organization (WIPO). (2007). WIPO Patent Report: Statistics on Worldwide Patent Activities.
- Census Data - http://factfinder.census.gov/

8.2 Network Analysis and Other Tools

Table 8.1 provides an overview of existing tools used in scientometrics research, see also (Fekete and Börner-chairs 2004). The tools are sorted by the date of their creation. Domain refers to the field in which they were originally developed such as social science (SocSci), scientometrics (Scientom), biology (Bio), geography (Geo), and computer science (CS). Coverage aims to capture the general functionality and types of algorithms available, e.g., Analysis and Visualization (A+V), see also description column.

Table 8.1 Network analysis and visualization tools commonly used in scientometrics research.

Tool	Year	Domain	Coverage	Description	UI	Open Source	Operating System	References

S&T Dynamics Toolbox	1985	Scientom	A + V	Tools from Loet Leydesdorff for organization analysis, and visualization of scholarly data.	Command-line	No	Windows	(Leydesdorff 2008)
In Flow	1987	SocSci	A + V	Social network analysis software for organizations with support for what-if analysis.	Graphical	No	Windows	(Krebs 2008)
Pajek	1996	SocSci*	A + V	A network analysis and visualization program with many analysis algorithms, particularly for social network analysis.	Graphical	No	Windows	(Batagelj and Mrvar 1998)

MANET	1996	Statistics	A + V	MANET is a tool for exploring data, providing a range of graphical tools for studying multivariate features.	Graphical	Yes	Mac	(Unwin 1996)
ExplorN	1997	Statistics	A +V	ExplorN is a data analysis and visualization tool that supports scatterplot matrices, parallel coordinate plots, incon-enhanced three-dimensional seteoscopic plots, and more.	Graphical	Yes	All Major	(Carr et al. 1997)
XGobi	1998	Statistics	A + V	XGobi is a data visualization system for viewing high-dimensional data.	Graphical	Yes	Windows, Linux	(Swayne et al. 1998)

UCINet	2000	SocSci*	A + V	Social network analysis software particularly useful for exploratory analysis.	Graphical	No	Windows	(Borgatti, Everett et al. 2002)
XploRe	2000	Statistics	A + V	XploRe allows for multidimensional analysis, non- and semiparametric modelling, and analysis of financial markets.	Graphical	Yes	Windows, Linux	(Härdle et al. 2000)
Boost Graph Library	2000	CS	Analysis and Manipulation	Extremely efficient and flexible C++ library for extremely large networks.	Library	Yes	All Major	(Siek, Lee et al. 2002)

nViZn	2000	Statistics	A + V	nViZn is a Java foundation for analytical graphics, best understood as a geometric analytical engine that allows for the visualization of statistical data.	Command-line	No	All Major	(Wilkinson et al. 2000)
Common GIS	2001	GeoVis	A + V	Common GIS is a tool for visualizing spatial data and allows for exploratory data analysis.	Graphical	Yes	Web based	(Andrienko et al. 2003)
Visone	2001	SocSci	A + V	Social network analysis tool for research and teaching, with a focus on innovative and advanced visual methods.	Graphical	No	All Major	(Brandes and Wagner 2008)

GeoVISTA	2002	Geo	GeoVis	GIS software that can be used to lay out networks on geospatial substrates.	Graphical	Yes	All Major	(Takatsuka and Gahegan 2002)
Cytoscape	2002	Bio*	Visualization	Network visualization and analysis tool focusing on biological networks, with particularly nice visualizations.	Graphical	Yes	All Major	(Cytoscape-Consortium 2008)
Mondrian	2002	Statistics	A + V	Mondrian is a general purpose data visualization program particularly useful when working with categorical data, geographical data and large data sets.	Graphical	Yes	All Major	(Theus 2002)

NetworkX	2002	Networks	A + V	NetworkX is a Python language software package that allows for the analysis and visualization of complex networks.	Command-line	Yes	All Major	(Hagberg, Swart, & S Chult 2008)
Tulip	2003	CS	Visualization	Graph visualization software for networks over 1,000,000 elements.	Graphical	Yes	All Major	(Auber 2003)
iGraph	2003	CS	Analysis and Manipulation	A library for classic and cutting edge network analysis usable with many programming languages.	Library	Yes	All Major	(Csárdi and Nepusz 2006)

CrystalVision	2003	Statistics	A + V	ExlporN is a data visualization program that focuses on parallel coordinate plots, scatterplots, and grand tour animations.	Graphical	Yes	All Major	(Wegman and Dorfman 2003)
CiteSpace	2004	Scientom	A + V	A tool to analyze and visualize scientific literature, particularly co-citation structures.	Graphical	Yes	All Major	(Chen 2006)
HistCite	2004	Scientom	A + V	Analysis and visualization tool for data from the Web of Science.	Graphical	No	Windows	(Garfield 2008)
R	2004	Statistics	A + V	A statistical computing language with many libraries for sophisticated network analyses.	Command-line	Yes	All Major	(Ihaka and Gentleman 1996)

GraphViz	2004	Networks	Visualization	Flexible graph visualization software.	Graphical	Yes	All Major	(AT&T-Research-Group 2008)
OpenGeoDa	2005	Geo	A + V	OpenGeoDa is a tool for exploratory spatial data analysis (ESDA) on lattice data (points and polygons).	Graphical	Yes	All major	(Anselin 2005)
Prefuse	2005	Visualization	Visualization	A general visualization framework with many capabilities to support network visualization and analysis.	Library	Yes	All Major	(Heer, Card et al. 2005)
VisuaLyzer	2005	Networks	Visualization	VisuaLyzer is a tool for importing data from Edgelists/Edgearray, Excel or GraphML formats and generating newtork visualizations.	Graphical	No	All Major	(Reshef 2009)

NWB Tool	2006	Bio, SocSci, Scientom	A + V	Network analysis & visualization tool conducive to new algorithms supportive of many data formats.	Graphical	Yes	All Major	(Huang 2007)
BibExcel	2006	Scientom	A + V	Transforms bibliographic data into forms usable in Excel, Pajek, NetDraw, and other programs.	Graphical	No	Windows	(Persson 2008)
tnet	2006	Netowrks	A + V	Tnet is a tool for social network analaysis of weighted, two-mode, and lognitudinal netowrks.	Graphcial	Yes	All Major	(Barrat et al. 2004)
KrackPlot	2006	Networks	A + V	KrackPlot is a netowrk analysis program designed specifically for the analysis of social networks.	Graphical	Yes	All Major	(McGrath and Blythe 2004)

libSNA	2006	Networks	A + V	This open-source social network analysis tool is under development by Abe Usher.	Graphical	Yes	All Major	(Terna 2008)
GUESS	2007	Networks	Visualization	A tool for visual graph exploration that integrates a scripting environment.	Graphical	Yes	All Major	(Adar 2007)
Publish or Perish	2007	Scientom	Data Collection and Analysis	Harvests and analyzes data from Google Scholar, focusing on measures of research impact.	Web-based	No	Windows, Linux	(Harzing 2008)
Commetrix	2007	Networks	A + V	Commetrix software allows users to dynamically analyze and create rich network maps.	Graphical	Yes	Windows (any system that supports Java)	(Trier and Bobrik 2007)

PEGASUS	2008	Networks	A + V	Overview PEGASUS is a Peta-scale graph mining system, fully written in Java that provides large scale algorithms for important graph mining tasks.	Command-line	Yes	All Major	(Kang, Tsourakakis, and Faloutsos, 2009)
GraphStream	2008	Networks	Visualization	GraphStream is a dynamic graph library written in Java that allows for the presentation of dynamic graphs.	Command-line	No	All Major	(Pigné et al. 2008)
Gephi	2009	Networks	A + V	Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.	Graphical	Yes	All Major	(Bastian et al., 2009)

VOSviewer	2009	Networks	Visualization	VOSviewer is a visualization tool useful for analyzing bibliometric networks.	Graphical	Yes	All Major	(Eck and Waltman 2011)
GraphInsight	2009	Networks	Visualization	GraphInsight is a freely available program that allows for visualization of complex data structures into graphs to provide business intelligence for organizations.	Graphical	No	All Major	(Dallachiesa and Nicolini 2009)
NodeXL	2010	SocSci	A + V	NodeXL is a free, open-source template for Excel 2007 and 2010 that lets you enter a network edge list, click a button, and see the network graph, all in the Excel window.	Graphical	No	Windows	(Hansen et al. 2010)

TINA	2011	Scientom, Networks	A + V	Tool for interactive assessment of projects portfolio and visualization of scientific landscapes.	Graphical	Yes	All Major	(Cointet 2008)
Spotfire 4.0	2011	Statistics	A + V	Spotfire allows users to analyze and visualize their data using complex and predictive statistics.	Graphical	No	All Major	(Shneiderman 2007)

Many of these tools are very specialized and capable. For instance, BibExcel and Publish or Perish are great tools for bibliometric data acquisition and analysis. HistCite and CiteSpace each support very specific insight needs – from studying the history of science to the identification of scientific research frontiers. The S&T Dynamics Toolbox provides many algorithms commonly used in scientometrics research and it provided bridges to more general tools. Pajek and UCINET are very versatile, powerful network analysis tools that are widely used in social network analysis. Cytoscape is excellent for working with biological data and visualizing networks.

The Network Workbench Tool has fewer analysis algorithms than Pajek and UCINET, and less flexible visualizations than Cytoscape. Network Workbench, however, makes it much easier for researchers and algorithm authors to integrate new and existing algorithms and tools that take in diverse data formats. The OSGi (http://www.osgi.org) component architecture and CIShell algorithm architecture (http://cishell.org) built on top of OSGi make this possible. Cytoscape is also adopting an architecture based on OSGi, though it will still have a specified internal data model and will not use CIShell in the core. Moving to OSGi will make it possible for the tools to share many algorithms, including adding Cytoscape's visualization capabilities to Network Workbench.

Several of the tools listed in the table above are also libraries. Unfortunately, it is often difficult to use multiple libraries, or sometimes any outside library, even in tools that allow the integration of outside code. Network Workbench, however, was built to integrate code from multiple libraries (including multiple versions of the same library). For instance, two different versions of Prefuse are currently in use, and many algorithms use JUNG (the Java Universal Network/Graph Framework). We feel that the ability to adopt new and cutting edge libraries from diverse sources will help create a vibrant ecology of algorithms.

Although it is hard to discern trends for tools which come from such diverse backgrounds, it is clear that over time the visualization capabilities of scientometrics tools have become more and more sophisticated. Scientometrics tools have also in many cases become more user friendly, reducing the difficulty of common scientometrics tasks as well as allowing scientometrics functionality to be exposed to non-experts. Network Workbench embodies both of these

trends, providing an environment for algorithms from a variety of sources to seamlessly interact in a user-friendly interface, as well as providing significant visualization functionality through the integrated GUESS tool.

Many other tools are available outside the scope of network analysis that are still useful for studying the data of science. One such tool is the web-based Data Science Toolkit, a web-based collection of open-source data sets and tools which allows the user to query for geographical data, parse text, and run named entity recognition.

9 References

- Adar, E. (2007, August 13). *Guess: The graph exploration system*. Retrieved from http://graphexploration.cond.org/.
- Andrienko, G., Andrienko, N., Voss, H. (2003). GIS for everyone: The CommonGIS project and beyond. In M. Peterson (Ed.), *Maps and the Internet* (pp. 131-146). Oxford, UK: Elsevier Science.
- Anselin, L., Syabri, I., & Kho, Y. (2005). GeoDa: An introduction to spatial data analysis. *Geographical Analysis, 38*(1), 5-22.
- AT&T-Research-Group. (2008). *Graphviz-Graph visualization software*. Retrieved from http://www.graphviz.org/Credits.php.
- Auber, D. (2003). Tulip: A huge graph visualization framework. In Mutzel, P. & Jünger, M (Eds.), *Graph drawing softwares: Mathematics and visualization* (105-126). Berlin: Springer-Verlag.
- Barabási, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509-512.
- Barabási, A. L. & Alber, R. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics,74* , 47-97.
- 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.
- Batagelj, V. & Brandes, U. Efficient generation of large random networks. *Physical Review E, 71*(3), 036113-036118.
- Batagelj, V. & Mrvar, A. (1998). Pajek-program for large network analysis. *Connections, 21*(2), 47-57.
- Borgatti, S. P., Everett M.G., et al. (2002). Ucinet for windows: Software for social network analysis. Retrieved from http://www.analytictech.com/ucinet6/ucinet_5_description.htm.
- Börner, Katy. 2011. "Plug-and-Play Macroscopes". *Communications of the ACM*. Vol. 54(3), 60-69, ACM Press.
- Börner, Katy, Luca Dall'Asta, Weimao Ke & Alessandro Vespignani. (2005). Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *Complexity: Special Issue on Understanding Complex Systems*. *10*(4), 57-67. http://ivl.cns.iu.edu/km/pub/2005-borner-studglob.pdf
- Börner, K., Huang, W., Linnemeier, M., Duhon, R.J., Phillips, P., Ma, N., Zoss, A., Guo, H., & Price, Mark. (2009). Rete-Netzwerk-Red: Analyzing and visualizing scholarly networks using the scholarly database and the Network Workbench tool *Eds. Proceedings of ISSI 2009: 12th International Conference on Scientometrics and Informetrics*, Rio de Janeiro, Brazil, July 14-17 . Vol. 2, Bireme/PAHO/WHO and the Federal University of Rio de Janeiro, 619-630. http://ivl.cns.iu.edu/km/pub/2009-borner-issi.pdf
- Börner, Katy, Maru, Jeegar & Goldstone, Robert. (2004). The Simultaneous Evolution of Author and Paper Networks. *Proceedings of the National Academy of Sciences of the United States of America. 101*(*Suppl. 1*), 5266-5273. http://ivl.cns.iu.edu/km/pub/2004-borner-tarl.pdf
- Börner, K., Penumarthy, S., et al. (2006). Mapping the diffusion of information among major U.S. research institutions. *Scientometrics: Dedicated issue on the 10th International Conference of the International Society for Scientometrics and Informetrics, 68*(3), 415-426.
- Boyack, K.W. & Börner, K. (2003). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society of Information Science and Technology, Special Topic Issue on Visualizing Scientific Paradigms*, *54*(5), 447-461. http://ivl.cns.iu.edu/km/pub/2003-boyack-indasst.pdf
- Boyack, Kevin W., Börner, Katy & Klavans, Richard. (2009). Mapping the Structure and Evolution of Chemistry Research. *Scientometrics*. *79*(1), 45-60. http://ivl.cns.iu.edu/km/pub/2009-boyack-mapchem.pdf
- Boyack, Kevin W., Klavans, Richard & Börner, Katy. (2005). Mapping the Backbone of Science. *Scientometrics*. *64*(3), 351-374. http://ivl.cns.iu.edu/km/pub/2005-boyack-mapbckbn.pdf

- Brandes, U. & Wagner, D. (2008). *Analysis and visualization of social networks*. Retrieved from http://visone.info/.
- Carr, D.B., Wegman, E.J., & Luo, Q. (1997). ExplorN: Design considerations past and present. Technical Report 137, Center for Computational Statistics, George Mason University, Fairfax, VA.
- Chen, C. (2006). Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST*, *54*(5), 359-377.
- Csárdi, G. & Nepusz,T. (2006). The igraph software package for complex network research. Retrieved from http://necsi.org/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf.
- Cyberinfrastructure for Network Science Center (2008). *Cyberinfrastructure Shell*.
- Cytoscape-Consortium. (2008). *Cytoscape*. Retrieved from http://www.cytoscape.org/index.php.
- Davidson, G.S., Wylie, B.N., et al. (2001). Cluster stability and the use of noise in interpretation of clustering. *IEEE Information Visualization*. San Diego, CA, IEEE Computer Society, 23-30.
- De Roure, D., Goble, C., et al. (2009). The design and realization of the myExperiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems, 25*, 561-567.
- Duhon, R. J. (2009). Understanding outside collaborations of the Chinese Academy of Sciences using Jensen-Shannon Divergence. *Proceedings of SPIE-IS&T Visualization and Data Analysis Conference*, San Jose . *7243*, pp. 72430C. http://ivl.cns.iu.edu/km/pub/2009-duhon-cas.pdf
- Elnashai, A., Spencer, B. et al. (2008). Architectural overview of MAEviz - HAZTURK. *Journal of Earthquake Engineering*,*12*(S2), 92-99.
- Erdös, P. & Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae Debrecen, 6*, 290-297.
- Fekete, J.D. & Börner, K. (Eds.). (2004). *Workshop on Information Visualization Software Infrastructures*. Austin, TX.
- Garfield, E. (2008). HistCite: Bibliometric Analysis and Visualization Software. Bala Cynwyd, PA: HistCite Software LLC.
- Gilbert, E.N. (1959). Random graphs. *The Annals of Mathematical Statistics, 30*, 1141.
- Härdle, W., Klinke, S., and Müller, M. (2000). *XploRe learning guide.* Amsterdam, Netherlands: Springer.
- Harzing, A.W. (2008). Publish or Perish: A citation analysis software program. Retrieved from http://www.harzing.com/resources.htm.
- Heer, J., Card, S.K., et al. (2005). Prefuse: A toolkit for interactive information visualization. *Conference on Human Factors in Computing Systems*, Portland, OR.
- Herr II, B.W., Talley, E.M., Burns, G., Newman, D., & La Rowe, G. (2009). The NIH Visual Browser: An interactive visualization of biomedical research. *Proceedings of the 13th International Conference on Information Visualization (IV09)*, Barcelona, Spain, *IEEE Computer Society*, 505-509. http://ivl.cns.iu.edu/km/pub/2009-herr-iv-visual-browser.pdf
- Huang, W. B., Herr II, B.W., Duhon, R.J., & Börner, K. (2007). Network Workbench--Using service-oriented architecture and component-based development to build a tool for network scientists. *International Workshop and Conference on Network Science,* Queens, NY.
- Herr II, B.W., Duhon, R.J., Börner, K., Hardy, E.F. & Penumarthy, S. (2008). 113 years of Physical Review: Using flow maps to show temporal and topical citation patterns. *Proceedings of the 12th Information Visualization Conference (IV 2008)*, London, UK. http://ivl.cns.iu.edu/km/pub/2008-herr-phys-rev.pdf
- Hull, D., K. Wolstencroft, et al. (2006). Taverna: A Tool for Building and Running Workflows of Services. *Nucleic Acids Research*, *34*(Web Server Issue): W729-W732.
- Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314.
- Jaro, M. A. (1989). Advances in record linking methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Society*, *64*, 1183-1210.
- Jaro, M. A. (1995). Probabilistic linkage of large public health data file. *Statistics in Medicine*, *14*, 491-498.
- Kampis, G., Gulyas, L., et al. (2009). Dynamic social networks and the TEXTrend / CIShell framework. *Applications of Social Network Analysis*. University of Zurich, ETH Zurich.
- Ke, W., Börner, K., & Viswanath, L. (2004). Analysis and visualization of the IV 2004 contest dataset. *IEEE Information Visualization Conference Poster Compendium*, 49-50. http://iv.cns.iu.edu/ref/iv04contest.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1),10-25.

- Klavans, R., Boyack, K.W. (2007). Is there a convergent structure to science? In Torres-Salinas, D. & Moed, H.F. (Eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics*, 437-448. Madrid: CSIC.
- Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(*3*), 455-476. http://sci.cns.iu.edu/klavans_2009_JASIST_60_455.pdf
- Kleinberg, J. M. (2002). Bursty and hierarchical structure in streams. *8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*, 91-101.
- Krebs, V. (2008). *Orgnet.com: Software for social network analysis and organizational network analysis*. Retrieved from http://www.orgnet.com/inflow3.html.
- Leydesdorff, L. (2008). *Software and data of Loet Leydesdorff*. Retrieved from http://www.leydesdorff.net/software.htm.
- Marshakova, I.V. (1973.) Co-citation in scientific literature: A new measure of the relationship between publications. *Scientific and Technical Information Serial of VINITI, 6*, 3-8.
- Martin, S., Brown, W.M., et al. (in preparation). DrL: Distributed Recursive (Graph) Layout. *Journal of Graph Algorithms and Applications*.
- Mane, K.K. & Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America*. *101*(Suppl. 1), 5287-5290.http://ivl.cns.iu.edu/km/pub/2004-mane-burstpnas.pdfNicolaisen,
- Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, *41,* 609-641.
- OSGi-Alliance. (2008). *OSGi alliance*. Retrieved from http://www.osgi.org/Main/HomePage.
- Persson, O. (2008). Bibexcel (Software). Umeå, Sweden, Umeå University.
- Shannon, P., Markiel, A. et al. (2002). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498-2504.
- Shneiderman, B. (2007, July 26). Dynamic queries, starfield displays, and the path to Spotfire. Retrieved from: http://www.cs.umd.edu/hcil/spotfire/.
- Siek, J.,Lee, L.Q., et al. (2002). *The boost graph library: User guide and reference manua*l. New York: Addison-Wesley.
- Small, H. (1973). Co-citation in scientific literature: A new measure of the relationship between publications. *J ASIS*, *24*, 265-269.
- Small, H. & Greenlee, E. (1986). Collagen research in the 1970's. *Scientometrics, 10*, 95-117.
- Swayne, D.F., Cook, D., & Buja, A. (1998). XGobi: Interactive dynamic data visualization in the X Window system. *Journal of Computational and Graphical Statistics, 7*(1), 113-130.
- Takatsuka, M. & Gahegan, M. (2002). GeoVISTA studio: A codeless visual programming environment for geoscientific data analysis and visualization. *The Journal of Computers & Geosciences*, *28*(10), 1131-1144.
- Theus, M. (2002). Interactive data visualization using Mondrian. *Journal of Statistical Software, 7*(11). Retrieved from http://www.jstatsoft.org/v07/i11.
- Unwin, A., Hawkins, G., Hofmann, H., & Siegl, B. (1996). Interactive graphics for data sets with missing values - MANET. *Journal of Computational and Graphical Statistics, 5*(2), 113-122.
- Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*, 440.
- Wegman, E.J. and Dorfman, A. (2003). Visualizing cereal world. *Computational Statistics & Data Analysis: Special Issue on Data Visualization, 43*(4), 633-649.
- Wellman, B., White, H.D., et al. (2004). Does citation reflect social structure? Longitudinal evidence from the "Globenet" interdisciplinary research group. *JASIST, 55*, 111-126.
- Wikimedia Foundation. (2009). Poisson distribution. *Wikipedia: The Free Encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/Poisson_distribution.
- Wilkinson, L., Rope, D., Rubin, M., & Norton, A. (2000). *nViZn: An algebra-based visualization system*. Department of Computer Science, College of Engineering, University of Illinois at Chicago. Retrieved from http://www.cs.uic.edu/~wilkinson/Publications/ibm.pdf.
- Zoss, A. & Börner, K. (forthcoming). Mapping transdisciplinary tobacco use research centers publications. *American Journal of Public Health - special issue on "Modeling in Tobacco Control."*
- Zoss, A., Conover, M., & Börner, K. (2010). Where are the academic jobs? Interactive exploration of job advertisements in geospatial and topical space. Sun-Ki,C. & Salerno, J. (Eds.). *Proceedings of the 2010*

International Conference on Social Computing, Behavioral Modeling and Prediction (SBP10), Springer. http://ivl.cns.iu.edu/km/pub/zoss-et-al-jobmaps.pdf

Appendix 1 Glossary

The following definitions are taken from: Gross, Jonathan L., and Jay Yellen. Handbook of Graph Theory. New York: CRC, 2004

unless otherwise noted.

- **Weakly Connected**
  - A directed graph is said to be ***weakly connected*** if its underlying undirected graph is ***connected***.
- **Connected**
  - An undirected graph is said to be ***connected*** "if there exists a walk between every pair of its vertices."
- **Mutually Reachable**
  - "Let *u* and *v* be vertices in a digraph *G*. Then *u* and *v* are said to be ***mutually reachable*** in *G* if *G* cont ains both a directed *u* - *v* walk and a directed *v* - *u* walk. Every vertex is regarded as reachable from itself (by the trivial walk)."
- **Strongly Connected**
  - "A digraph is ***strongly connected*** if every two vertices are ***mutually reachable***.
- **Strong Component**
  - "A ***strong component*** of a digraph *G* is a maximal strongly connected subgraph of *G*. Equivalently, a ***strong component*** is a subdigraph induced on a maximal set of ***mutually reachable*** vertices.
- **Component**
  - "The subgraphs of *G* which are maximal with respect to the property of being ***connected*** are called the components of *G*."
- **Graph Density**
  - "The density of a graph is the ratio of the number of edges and the number of possible edges." (from igraph library documentation).

Appendix 2 CIShell Algorithms

Below is a list of the algorithms created so far from each CNS CIShell-powered tool. There are two lists, one organizes the algorithms by their function in the tools and the other organizes the algorithms alphabetically. Some pages are currently somewhat sparse, as of March 2011.

If you would like to view an index of algorithms in a particular CNS CIShell-powered tool, see the Sci2 Algorithm Index, the NWB Manual, or the EpiC Manual.

Source code for these algorithms is browsable in the CNS Center repository, see Document & Developer Resources (direct link to repository is currently down. Should be available soon as of March 2011).

Algorithms organized by function

Data Preparation

- Remove ISI Duplicate Records

- Remove Rows with Multitudinous Fields
- Extract Directed Network
- Extract Bipartite Network
- Extract Paper Citation Network
- Extract Author Paper Network
- Extract Co-Occurrence Network
- Extract Word Co-Occurrence Network
- Extract Co-Author Network
- Extract Reference Co-Occurrence (Bibliographic Coupling) Network
- Extract Document Co-Citation Network
- Detect Duplicate Nodes
- Update Network by Merging Nodes

Database

ISI

- Merge Identical ISI People
- Suggest ISI People Merges
- Merge Document Sources
- Match References to Papers
- Extract Authors
- Extract Documents
- Extract Keywords
- Extract Document Sources
- Extract Authors by Year
- Extract References by Year
- Extract Original Author Keywords by Year
- Extract New ISI Keywords by Year
- Extract Authors by Year for Burst Detection
- Extract Documents by Year for Burst Detection
- Extract Original Author Keywords by Year for Burst Detection
- Extract New ISI Keywords by Year for Burst Detection
- Extract References by Year for Burst Detection
- Extract Longitudinal Summary
- Extract Co-Author Network
- Extract Author Citation Network
- Extract Document Citation Network (Core Only)
- Extract Document Citation Network (Core and References)
- Extract Document Source Citation Network (Core Only)
- Extract Document Source Citation Network (Core and References)
- Extract Document Co-Citation Network (Core Only)
- Extract Document Co-Citation Network (Core and References)
- Extract Document Source Co-Citation Network (Core Only)
- Extract Document Source Co-Citation Network (Core and References)
- Extract Author Co-Citation Network
- Extract Author Bibliographic Coupling Network
- Extract Document Bibliographic Coupling Network
- Extract Document Source Bibliographic Coupling Network
- Extract Weighted Document-Document Network

NSF

- Merge Identical NSF People
- Extract Investigators

- Extract Awards
- Extract Organizations
- Extract Co-PI Network

General

- Create Merging Table
- Merge Entities
- Custom Table Query
- Custom Graph Query
- Extract Raw Tables from Database

Generic-CSV

- Extract Co-Occurrence Network
- Extract Bipartite Network
- Extract Co-Entity Occurrence Network
- Extract Table

Preprocessing

General

- Extract Top N% Records
- Extract Top N Records
- Aggregate Data

Temporal

- Slice Table by Time

Geospatial

- Extract ZIP Code

Topical

- Lowercase, Tokenize, Stem, and Stopword Text

Networks

- Extract Timestep
- Extract Top Nodes
- Extract Nodes Above or Below Value
- Remove Node Attributes
- Delete High Degree Nodes
- Delete Random Nodes
- Delete Isolates
- Extract Top Edges
- Extract Edges Above or Below Value
- Remove Edge Attributes
- Remove Self Loops
- Trim by Degree
- Snowball Sampling (N nodes)
- MST-Pathfinder Network Scaling
- Fast Pathfinder Network Scaling
- Pathfinder Network Scaling
- Node Sampling

- Edge Sampling
- Symmetrize
- Dichotomize
- Multipartite Joining
- Lowercase, Tokenize, Stem, and Stopword Text
- Slice Table by Time
- Merge 2 Networks

Analysis

Temporal

- Burst Detection

Geospatial

- Geocoder
- Congressional District Geocoder
- Yahoo Geocoder

Topical

- Burst Detection

Networks

- Network Analysis Toolkit (NAT)

Unweighted & Undirected

- Node Degree
- Degree Distribution
- K-Nearest Neighbor (Java)
- Watts-Strogatz Clustering Coefficient
- Watts Strogatz Clustering Coefficient over K
- Diameter
- Average Shortest Path
- Shortest Path Distribution
- Node Betweenness Centrality
- Weak Component Clustering
- Global Connected Components
- Extract K-Core
- Annotate K-Coreness
- Blondel Community Detection
- HITS

Weighted & Undirected

- Clustering Coefficient
- Nearest Neighbor Degree
- Strength vs Degree
- Degree & Strength
- Average Weight vs End-point Degree
- K-Nearest Neighbor (Java)
- Strength Distribution
- Weight Distribution
- Randomize Weights

- PageRank
- Node Betweenness Centrality
- Blondel Community Detection
- HITS
- MST-Pathfinder Network Scaling
- Fast Pathfinder Network Scaling

Unweighted & Directed

- Node Indegree
- Node Outdegree
- Indegree Distribution
- Outdegree Distribution
- K-Nearest Neighbor
- Single Node In-Out Degree Correlations
- Dyad Reciprocity
- Arc Reciprocity
- Adjacency Transitivity
- Node Betweenness Centrality
- Weak Component Clustering
- Strong Component Clustering
- Extract K-Core
- Annotate K-Coreness
- HITS
- PageRank

Weighted & Directed

- HITS
- PageRank
- Fast Pathfinder Network Scaling
- Blondel Community Detection

Search

- Can
- Chord
- K Random-Walk
- Random Breadth First

Discrete Network Dynamics

- Extract and Annotate Attractors

Modeling

Networks

- Random Graph
- Watts-Strogatz Small World
- Barabási-Albert Scale-Free
- Can
- Chord
- Hypergrid
- PRU
- TARL (Topics, Aging and Recursive Linking)
- Discrete Network Dynamics (DND)

- Evolving Network (Weighted)

Visualization

General

- GnuPlot

Temporal

- Horizontal Bar Graph
- Horizontal Bar Graph (not included version)
- Temporal Bar Graph

Geospatial

- Proportional Symbol Map
- Choropleth Map

Networks

- GUESS
- Gephi
- GnuPlot
- Radial Tree/Graph (prefuse alpha)
- Radial Tree/Graph with Annotations (prefuse beta)
- Tree View (prefuse beta)
- Tree Map (prefuse beta)
- Balloon Graph (prefuse beta)
- Force Directed with Annotation (prefuse beta)
- Kamanda-Kawai (JUNG)
- Fruchterman-Reingold (JUNG)
- Fruchterman-Reingold with Annotation (prefuse beta)
- Spring (JUNG)
- Small World (prefuse alpha)
- Parallel Coordinates (demo)
- LaNet
- DrL(VxOrd)
- Specified (Prefuse beta)
- Circular Hierarchy
- Circular (JUNG)
- Bipartite Network Graph

Simulation

Single-Population

- Stochastic
- Exact

Network

- Stochastic
- Bipartite Network Graph

R

- Create an R Instance
- Run Rgui
- Send a Table to R
- Get a Table from R

Scientometrics

- Remove ISI Duplicate Records
- Remove Rows with Multitudinous Fields
- Detect Duplicate Nodes
- Update Network by Merging Nodes
- Extract Directed Network
- Extract Paper Citation Network
- Extract Author Paper Network
- Extract Co-Occurrence Network
- Extract Word Co-Occurrence Network
- Extract Co-Author Network
- Extract Reference Co-Occurrence (Bibliographic Coupling) Network
- Extract Document Co-Citation Network

Google Citation Reader

- Google Citation User ID Search Algorithm
- Attach Citation BibTex File from Google Scholar
- Attach Citation Indices from Google Scholar
- Attach Citation Table from Google Scholar

XML to CSV Converter

- XML to CSV Converter Algorithm

Algorithms organized alphabetically

- Adjacency Transitivity
- Aggregate Data
- Annotate K-Coreness
- Annotate K-Coreness (Unweighted & Directed)
- Arc Reciprocity
- Attach Citation BibTex File from Google Scholar
- Attach Citation Indices from Google Scholar
- Attach Citation Table from Google Scholar
- Average Shortest Path
- Average Weight vs End-point Degree
- Barabási-Albert Scale-Free
- Bipartite Network Graph
- Blondel Community Detection
- Burst Detection
- Can
- Chord
- Circular Hierarchy
- Clustering Coefficient
- Congressional District Geocoder
- Create an R Instance
- Create Document Source Merging Table

- Create Merging Table
- Custom Graph Query
- Custom Table Query
- Degree & Strength
- Degree Distribution
- Delete Isolates
- Detect Duplicate Nodes
- Diameter
- Dichotomize
- DrL (VxOrd)
- Dyad Reciprocity
- Edge Sampling
- Extract Author Bibliographic Coupling Network
- Extract Author Citation Network
- Extract Author Co-Citation Network
- Extract Author Paper Network
- Extract Authors
- Extract Authors by Year
- Extract Authors by Year for Burst Detection
- Extract Awards
- Extract Bipartite Network
- Extract Co-Author Network
- Extract Co-Author Network (Text Files)
- Extract Co-Occurrence Network
- Extract Co-PI Network
- Extract Directed Network
- Extract Document Bibliographic Coupling Network
- Extract Document Citation Network (Core and References)
- Extract Document Citation Network (Core Only)
- Extract Document Co-Citation Network
- Extract Document Co-Citation Network (Core and References)
- Extract Document Co-Citation Network (Core Only)
- Extract Documents
- Extract Documents by Year for Burst Detection
- Extract Document Source Bibliographic Coupling Network
- Extract Document Source Citation Network (Core and References)
- Extract Document Source Citation Network (Core Only)
- Extract Document Source Co-Citation Network (Core and References)
- Extract Document Source Co-Citation Network (Core Only)
- Extract Document Sources
- Extract Edges Above or Below Value
- Extract Investigators
- Extract K-Core
- Extract K-Core (Unweighted & Directed)
- Extract Keywords
- Extract Longitudinal Summary
- Extract New ISI Keywords by Year
- Extract New ISI Keywords by Year for Burst Detection
- Extract Nodes Above or Below Value
- Extract Organizations
- Extract Original Author Keywords by Year
- Extract Original Author Keywords by Year for Burst Detection
- Extract Paper Citation Network

- Extract Raw Tables from Database
- Extract Reference Co-Occurrence (Bibliographic Coupling) Network
- Extract References by Year
- Extract References by Year for Burst Detection
- Extract Top Edges
- Extract Top N% Records
- Extract Top Nodes
- Extract Top N Records
- Extract Weighted Document-Document Network
- Extract Word Co-Occurrence Network
- Extract ZIP Code
- Fast Pathfinder Network Scaling
- Force Directed with Annotation (prefuse beta)
- Fruchterman-Reingold with Annotation (prefuse beta)
- Geocoder
- Geospatial Visualization
- Gephi
- Get a Table from R
- Global Connected Components
- Google Citation User ID Search Algorithm
- GUESS
- HITS
- Horizontal Bar Graph
- Horizontal Bar Graph (not included version)
- Hypergrid
- Indegree Distribution
- K-Nearest Neighbor
- K-Nearest Neighbor (Java)
- Load Existing Database
- Load ISI File Into Database
- Load ISI File with 'ISI flat format'
- Load NSF File Into Database
- Lowercase, Tokenize, Stem, and Stopword Text
- Map of Science via 554 Fields
- Map of Science via Journals
- Match References to Papers
- Merge 2 Networks
- Merge Document Sources
- Merge Entities
- Merge Identical ISI People
- Merge Identical NSF People
- MST-Pathfinder Network Scaling
- Multipartite Joining
- Nearest Neighbor Degree
- Network Analysis Toolkit (NAT)
- Node Betweenness Centrality
- Node Degree
- Node Indegree
- Node Outdegree
- Node Sampling
- Outdegree Distribution
- PageRank
- Pathfinder Network Scaling

- PRU
- Radial Tree or Graph (prefuse alpha)
- Radial Tree or Graph with Annotations (prefuse beta)
- Random Graph
- Randomize Weights
- Read citation indices for a given set of authors
- Remove Edge Attributes
- Remove ISI Duplicate Records
- Remove Node Attributes
- Remove Rows with Multitudinous Fields
- Remove Self Loops
- Run Rgui
- Send a Table to R
- Shortest Path Distribution
- Single Node In-Out Degree Correlations
- Slice Table by Time
- Snowball Sampling (N nodes)
- Specified (Prefuse beta)
- Strength Distribution
- Strength vs Degree
- Strong Component Clustering
- Suggest ISI People Merges
- Symmetrize
- TARL (Topics, Aging and Recursive Linking)
- Temporal Bar Graph
- Tree Map (prefuse beta)
- Tree View (prefuse beta)
- Trim by Degree
- Update Network by Merging Nodes
- Watts-Strogatz Clustering Coefficient
- Watts Strogatz Clustering Coefficient over K
- Watts-Strogatz Small World
- Weak Component Clustering
- Weight Distribution
- Workflow Module
- XML to CSV Algorithm
- Yahoo Geocoder

Appendix 3 Date Format

The date format is the format that represents the date system used by the date data. The Sci2 algorithms support date format schema that is defined in the Java documentation at here. The following are some guides to identify the date formats used in your data set.

How to view the date string in the data set

Most of the editor will convert the date string to the human readable format. For example, a date 'January 3, 1999 16:49' is displayed as 3-Jan-99 16:49 in Excel worksheet. So it is recommended you use a plaint text editor. The following is some recommentation:
i) Windows' system: WordPad, Notepad, Microsoft Word.
ii) Mac's system: TextEdit, Emacs, TextWrangler.
iii) Linux and Unix system: gEdit, Text Editor,Emacs, or any command line commands such as VIM, Nano, etc.

The following are some quick guides about ISI and NSF data's date format.

Common date format for NSF and ISI data

NSF

A standard NSF file contains the following date columns: Start Date, Last Amendment Date, and Expiration Date. An example of the date string and the relative date format is shown below.
Example date: January 3, 1999
Date format: MMMMM d, yyyy

ISI

A standard ISI file contains the following date columns: Cited Year, Publication Date and Publication Year. The Cited Year and Publication Year columns only contains year information while the publication date only contains month and day of month information. The following is the examples of the date string with the relative date format.

The Cited Year and Publication Year columns
Example date: 1923
Date format: yyyy

The Publication Date column
The Publication Date column might contains vary formats such as JAN, JAN-APR and JAN 03. However, most data only contains month value. We suggest you to change all date to the following example form and use the provided date format.
Example date: JAN
Date format: MMM

Other data

If your data is not ISI nor NSF data. You can use on of the following guides to retrieve the date format of your data.

Convert with Excel worksheet and use the following setting

This only applicable if your data file is compatible with Excel worksheet
i) Open with Excel worksheet and save the file as different name
ii) Find out your date string by following the 'How to view the date string in the data set' guidance
iii) Load the new data file into Sci2 tool and using the following matched example date string

Example date: 3/12/2011 13:17
Date format: M/dd/yyyy HH:mm

Example date: 3/12/2011
Date format: MM/dd/yyyy

Example date: 1-Apr
Date format: d-MMM

Example date: 1-Jan-11
Date format: d-MMM-yy

Do it your self with Java Documentation

i) Find out your date string by following the 'How to view the date string in the data set' guidance

ii) Retrieve your date format by combining the fields format that are defined at here

Appendix 4 Sci2 Release Notes

4.1 Sci2 Release Notes v0.5 alpha

Announcement

The Cyberinfrastructure for Network Science Center is pleased to announce the release of the Sci^2 (Science of Science) Tool v0.5 alpha, which contains much of the functionality of the Network Workbench tool as well as a variety of new general purpose and scientometrics-focused algorithms. Sci^2 is still undergoing significant changes, but already has many new features that are worth exploring. Please visit the new Sci^2 homepage to download and learn more about Sci^2 .

Sci^2 is well documented with over 12 tutorials prepared for NIH in Summer 2010 and an online user manual with more than 100 pages of content, including sample Sci^2 workflows and documentation for every algorithm in Sci^2. New features in Sci^2 so far include web-based Yahoo! and desktop Geocoders, two different ways to overlaying geographic information on U.S. and World Maps, customizable stop word lists, and an algorithm for combining two networks into one. We are also developing a new scalable database-oriented scientometrics pipeline capable of producing many new types of networks and tables based on ISI, NSF, and now generic CSV data. This new pipeline can be downloaded as a supplemental preview package from the download section of the Sci^2 site, and will be available by default in Sci^2 for future releases. We have also fixed many bugs from Network Workbench and previous Sci^2 versions. See below for a full list of changes since Sci^2 v0.3.

Network Workbench and Sci^2 are both based on the CIShell framework, which makes it easy to add new algorithms to each tool. See our recent video featured in the Communications of the ACM or the CIShell wiki for more information.

Documentation

Download an offline copy of this manual (87.6 MB) (Updated May 31, 2011)
Download an offline copy of the CIShell Manual, which contains algorithm documentation for Sci2 (19.9 MB) (Updated May 31, 2011)

Tool Change Log

New Features

- Yahoo! Geocoder.
- Combine 2 Networks algorithm.
- Added support for custom stop word lists (the stop word algorithm now refers to an external file, which can be copied and edited to contain different stop words).
- Added GUESS support for saving out node positions. This can be used in combination with network timeslicing to help in the creation of an animation or a series of network images where the network grows over time.
- GUESS zoom level can now be adjusted in absolute terms. This helps when trying to compare to GUESS visualizations (set the zoom levels to agree).
- Added support for loading files by dragging them from the desktop to the data manager, and saving files via

dragging them out of the Data Manager.
- Added support for loading multiple files simultaneously via File -> Load.
- New website, including the 'Ask an Expert' feature.
- In supplemental database package:
    - New 'Extract X by Year' algorithms that work as input for Burst detection.
    - database dump algorithm
    - bibliographic coupling network extractions
    - longitudinal summary extraction
    - Added support for generic-csv databases, including the loader GUI and the extraction GUI.

Improvements

- Improved support for UTF-8 data.
- Shortened file names for loaded files.
- improved zip code parsing.
- Fixed circle sizes in GeoMaps so they could be adjusted by a scaling factor.
- Improved GeoMaps so it only shows projections that make sense in the context of the map you want to show (either World or U.S. map).
- Enhanced Geocoder so it can resolve congressional districts to geographic coordinates.
- Improved CSV reader so it could handle some new variations in CSV input (including data from the NIH RePORTER site).
- GUESS now shows in its title bar the name of the network it is displaying.
- Blondel Community Detection now handles non-integer weights.
- In supplemental database package:
    - Expanded database default list of journal names to merge.
    - Improved database performance
    - Improved performance with bibliographic coupling extraction.
    - Database loaders now show % complete when loading data.
    - Improved database loading performance.
    - Improved database co-authorship extraction to include earliest & most recent publication date for each author node.
    - Added 'total # works' and 'total times cited' counts to the author extraction table.
    - Changed python scripts and/or database and old-pipeline attribute names so our custom GUESS python scripts work for both database and old-pipeline output.

Bug fixes

- Fixed a text display error on Macs.
- Fixed an issue where GUESS couldn't display certain pajek.net files.
- Fixed an issue where filenames could start or end with a space.
- Fixed an error with creating directed networks from some NSF files.
- Fixed an issue where text-based pipeline co-authorship network extraction no longer merged names with different cases.
- Fixed an issue in GUESS where certain drop-down boxes would not be populated.
- Fixed an issue with dates in the NSF database.
- Fixed an issue where Extract Authors by Year output would sometimes be invalid.
- Fixed issue with having multiple files with the same name in the Data Manager.
- Fixed an issue with the ISI paper -> reference matcher.
- Fixed a bug where Node Betweenness Centrality would return a graph with zero nodes.
- Fixed an error condition in Co-Word Occurrence algorithm.
- Fixed a confusing package naming issue between two versions of delete isolates.
- Fixed merging capabilities in non-db pipeline.
- Fixed broken links under Help -> About.
- Fixed error handling in Geo Maps when an input file had no numeric columns (and hence no basis for size or color coding).

- Fixed an issue with 'delete isolates' and 'weak component clustering'.
- Fixed an issue with NWB formatting that was causing issues in GUESS.
- Fixed an error with handling EndNote data.
- Fixed a number of errors in GUESS python scripts.
- Fixed an error in 'Extract Nodes Above or Below Value'.
- Fixed an error in 'normalize text'.
- Fixed an issue in GUESS where edges in UTF-8 files would not be shown.
- Fixed old references in Sci^2 that pointed to the old bug tracker to point to our new bug tracker.
- Fixed an issue with the output of document citation network where it wouldn't work with Weighted Page Rank.
- Fixed issue where some pajek.net files would not load properly in pajek.
- Changed some text that still said 'NWB' to 'Sci^2 '.
- Fixed an issue with viewing pajek.net files immediately after they were loaded into Sci^2 .
- Fixed an issue with GUESS colorizing.
- Fixed a merging error produced by one of our Sci^2 workflows.
- Fixed an issue with Node Betweenness Centrality.
- Fixed an issue with normalizing author names.
- Fixed an issue where Aggregate Data would not have a good error message when columns have null values in them.
- Fixed an issue with viewing the CSV output of Extract ZIP code.
- Fixed an issue with GUESS displaying XGMML input strangely.
- Fixed an issue with the Network Analysis Toolkit where it was sensitive to whitespace in pajek.net files.
- Fixed an issue with Delete Isolates where the number of nodes deleted differed from what it was reporting on the Console.
- In supplemental database package:
    - Fixed a database error with Core journal co-citation output producing strange results in NAT and GUESS.
    - Fixed a database error where Extract Authors would sometimes return zero results.
    - Fixed an issue with loading certain unusual ISI files into the ISI database.
    - Changed text so that it always used the more inclusive term 'Document Sources' rather than 'Journals' where appropriate.
    - Fixed a bug where database produced bad output for GUESS.

4.2 Sci2 Release Notes v0.5.1 alpha

Announcement

This release includes improvements and bug fixes for high-demand algorithms such as the Burst Detection algorithm and the Guess visualization algorithm. See the change log below for full details. See also the Sci2 release v0.5 alpha release notes (from our last major release).

Documentation

Download an offline copy of this manual (87.6 MB) (Updated May 31, 2011)
Download an offline copy of the CIShell Manual, which contains algorithm documentation for Sci2 (19.9 MB) (Updated May 31, 2011)

Tool Change Log

New Features

- Added support for different burst lengths and burst length units (minutes, hours, days, months and years) on the Burst Detection algorithm. This will allow user to divide the date range into a number of burstable periods (batches).

Improvements

- Extra information and suggestions will be printed to the Console when using burst detection in the following cases: no burst is detected, the given date format does not match the given date values, and when there is not enough memory to handle a large number of batches.

Bug fixes

- Fixed an issue where the Node Betweenness Centrality algorithm failed when processing certain NWB files.
- Fixed a GUESS issue where the Graph Modifier tab was missing on Mac machines.
- Fixed a null pointer exception on the Slice-by-time algorithm which occurred when the user-specified time range was larger than the time range of the provided data.

Patches

- WOS-plugins.zip - These plugins provide the support for the newest version of the **ISI Web of Science file format**. Download the WOS-plugin.zip file and uncompress the plugins into your sci2/plugins/ directory. Note: This plugins will be included in the future Sci2 release. See ISI (*.isi) for more information.
- edu.iu.nwb.analysis.extractnetfromtable_1.0.0.jar - Support long data type for aggregating data through extract network algorithms

4.3 Sci2 Release Notes v0.5.2 alpha

Announcement

This release includes support for the new Web of Science format, network overlays on the Geographical map and Prefuse's visualizations on Mac's OS X. See the change log below for full details.

Documentation

Download an offline copy of this manual (87.6 MB) (Updated May 31, 2011)
Download an offline copy of the CIShell Manual, which contains algorithm documentation for Sci2 (19.9 MB) (Updated May 31, 2011)

Tool Change Log

New Features

- Adds support for the new Web of Science format from ISI. This will allow the user to load both old and new ISI data downloaded from the ISI site.
- Adds support for network overlays on the Geographical map. This will allow Sci2 users to generate a geographical network through Sci2, Gephi and Adobe Photoshop. See tutorial here
- Adds support for Prefuse's visualizations on Mac's OS X. This will allow Mac's OS X users to be able to access all the visualization algorithms in Sci2, including Tree Map, Tree View, Radial Graph, etc.

Improvements

- Triples the speed and reduces memory usage by 70% for the Extract Top N Nodes and Extract Top N Edges algorithms. This will allow Sci2 to support rendering for bigger networks.
- Unified merging algorithms used by database

Bug fixes

- Fixed legend boundary issue for geographical maps
- Fixed a typo on the output data label
- Fixed slice by year bug for databases

4.4 Sci2 Release Notes v1.0 alpha

Announcement

This release is built with CIShell 2.0 framework. It includes support for Mac OS X Cocoa platform, and new features such as new visualizations, Google Scholar Citation reader, R bridging, Gephi bridging, and many more. See the change log below for full details

Note:

The database plugin at 3.2 Additional Plugins is not compatible with this version. A new database plugin will be available soon. Please follow us at http://sci2.cns.iu.edu

Tool Change Log

New Features

- Support Mac OS X Cocoa frameworks platform
- Temporal Bar Graph visualization replaced Horizontal bar graph visualization
- Map of Science visualization via Journals and Map of Science visualization via 554 Fields
- Bipartite Network visualization
- Geospatial visualizations with new map layout
- Reader for Google Scholar Citations Web data
- Added support for exporting from Sci2 to Gephi.
- New Gephi workflow to design network overlays on geomaps (requires that Gephi is pre-installed)
- R bridge that supports starting R and send/get a data table to/from R (requires that R is pre-installed)
- Writer for PDF, PNG, JPEG data format using Ghostscript (requires that Ghostscript is pre-installed)
- Novel CIShell 2.0 framework

Improvements

- Support overflow and underflow check for aggregate functions
- Unified output attribute names for network algorithms
- Unified visualization parameters and console print out
- Standardized visualization Postscript layout with How to Read text, legends and title
- Reposition map for Geographical Network Algorithm
- Improved Error message handling for Geographical Network algorithm
- Updated online tutorial to explain Author Ages issue
- Updated online tutorial regarding ISI total time cited default -1 value
- Improved bipartite network edges drawing
- Improved bipartite sorting for numerical numbers that are string type
- Updated developer list
- Improved bipartite sorting for numerical numbers that are string type
- Updated new color schema for all visualizations to follow color blind standard
- Updated UCSD map of science with 10 years of ISI and Scopus journal mapping, discipline colors, and added unclassified journal listing.
- Added download utility to standardize network handling for algorithms that use internet service
- Moved R working directory into system temporary directory
- Updated FreeHEP (a Java2D object to PostScript conversion library) to 2.1.1
- Updated Geolibs (an open source Java GIS toolkit) to 2.7.4
- Updated Menu items
    - Update menu item order and structure
    - Rename Circle Geographical Map to Proportional Symbol Map
    - Rename Region Geographical Map to Choropleth Map
    - Rename Yahoo Geocoder name to Yahoo! Geocoder
    - Rename K-Nearest Neighbor (java) to K-Nearest Neighbor

Bug fixes

- Fixed overflow issue of dropdown box in GUESS
- Fixed missing model.jpg issue
- Fixed menu path for all algorithms
- Fixed NULL pointer exception on Aggregate function
- Fixed Proportional Symbol Map failure to handle float value
- Fixed Proportional Symbol Map Null Pointers issue
- Fixed weighted PageRank failure to handle weight error
- Fixed duplicate items of GUESS dropdown box
- Removed the misleading author name normalization message from ISI file load algorithm