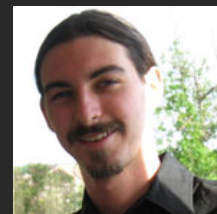


Preparing, Analyzing, and Visualizing Humanities Data

Dr. Katy Börner and Scott Weingart

Cyberinfrastructure for Network Science Center
Information Visualization Laboratory
School of Library and Information Science
Indiana University, Bloomington, IN
<http://cns.slis.indiana.edu>



With special thanks to Kevin W. Boyack, Micah Linnemeier, Russell J. Duhon, Patrick Phillips, Joseph Biberstine, Chintan Tank Nianli Ma, Hanning Guo, Mark A. Price, Angela M. Zoss, and Sean Lind



Networks and Network Analysis for the Humanities Workshop
Rice Room, Institute for Pure and Applied Mathematics
UCLA, Los Angeles, CA

9:00-12:00 on August 17, 2010



Workshop Overview

9:00 Marcoscope Design and Usage & CShell Powered Tools

9:15 Sci2 Tool Basics

- Download and run the tool.

9:30 Sci2 Workflow Design: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

10:00 Sci2 Tool – Using Text and Database Support

- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 *Break*

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance



Workshop Overview

9:00 Marcoscope Design and Usage & CShell Powered Tools

9:15 Sci2 Tool Basics

- Download and run the tool.

9:30 Sci2 Workflow Design: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

10:00 Sci2 Tool – Using Text and Database Support

- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance

3



The Changing Scientific Landscape

Star Scientist -> Research Teams: In former times, science was driven by key scientists.

Today, science is driven by effectively collaborating co-author teams often comprising expertise from multiple disciplines and several geospatial locations (Börner, Dall'Asta, Ke, & Vespignani, 2005; Shneiderman, 2008).

Users -> Contributors: Web 2.0 technologies empower anybody to contribute to Wikipedia or to exchange images and videos via Flickr and YouTube. WikiSpecies, WikiProfessionals, or WikiProteins combine wiki and semantic technology in support of real time community annotation of scientific datasets (Mons et al., 2008).

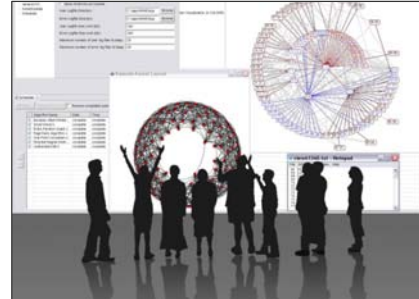
Cross-disciplinary: The best tools frequently borrow and synergistically combine methods and techniques from different disciplines of science and empower interdisciplinary and/or international teams of researchers, practitioners, or educators to fine-tune and interpret results collectively.

One Specimen -> Data Streams: Microscopes and telescopes were originally used to study one specimen at a time. Today, many researchers must make sense of massive streams of multiple types of data with different formats, dynamics, and origin.

Static Instrument -> Evolving Cyberinfrastructure (CI): The importance of hardware instruments that are rather static and expensive decreases relative to software infrastructures that are highly flexible and continuously evolving according to the needs of different sciences. Some of the most successful services and tools are decentralized increasing scalability and fault tolerance.



Microscopes, Telescopes, and Macroscopes



Just as the **microscope** empowered our naked eyes to see cells, microbes, and viruses thereby advancing the progress of biology and medicine or the **telescope** opened our minds to the immensity of the cosmos and has prepared mankind for the conquest of space, **macroscopes** promise to help us cope with another infinite: the infinitely complex. Macroscopes give us a ‘vision of the whole’ and help us ‘synthesize’. They let us detect patterns, trends, outliers, and access details in the landscape of science. Instead of making things larger or smaller, macroscopes let us observe what is at once too great, too slow, or too complex for our eyes.



Desirable Features of Macroscopes

Core Architecture & Plugins/Division of Labor: Computer scientists need to design the standardized, modular, easy to maintain and extend “core architecture”. Dataset and algorithm plugins, i.e., the “filling”, are provided by those that care and know most about the data and developed the algorithms: the domain experts.

Ease of Use: As most plugin contributions and usage will come from non-computer scientists it must be possible to contribute, share, and use new plugins without writing one line of code. Users need guidance for constructing effective workflows from 100+ continuously changing plugins.

Modularity: The design of software modules with well defined functionality that can be flexibly combined helps reduce costs, makes it possible to have many contribute, and increases flexibility in tool development, augmentation, and customization.

Standardization: Adoption of (industry) standards speeds up development as existing code can be leveraged. It helps pool resources, supports interoperability, but also eases the migration from research code to production code and hence the transfer of research results into industry applications and products.

Open Data and Open Code: Lets anybody check, improve, or repurpose code and eases the replication of scientific studies.

Macroscopes are similar to Flickr and YouTube and but instead of sharing images or videos, you freely share datasets and algorithms with scholars around the globe.



Macroscopic Design



Custom Tools for Different Scientific Communities

Information Visualization Cyberinfrastructure

<http://iv.slis.indiana.edu>

Network Workbench Tool + Community Wiki

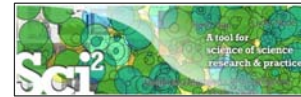
<http://nwb.slis.indiana.edu>

Science of Science (Sci²) Tool and Portal

<http://sci.slis.indiana.edu>

Epidemics Cyberinfrastructure

<http://epic.slis.indiana.edu/>



180+ Algorithm Plugins and Branded GUIs

+

Core Architecture

Open Services Gateway Initiative (OSGi) Framework.

<http://orgi.org>

Cyberinfrastructure Shell (CIShell)

<http://cishell.org>



7

CIShell Powered Tools: Network Workbench (NWB)

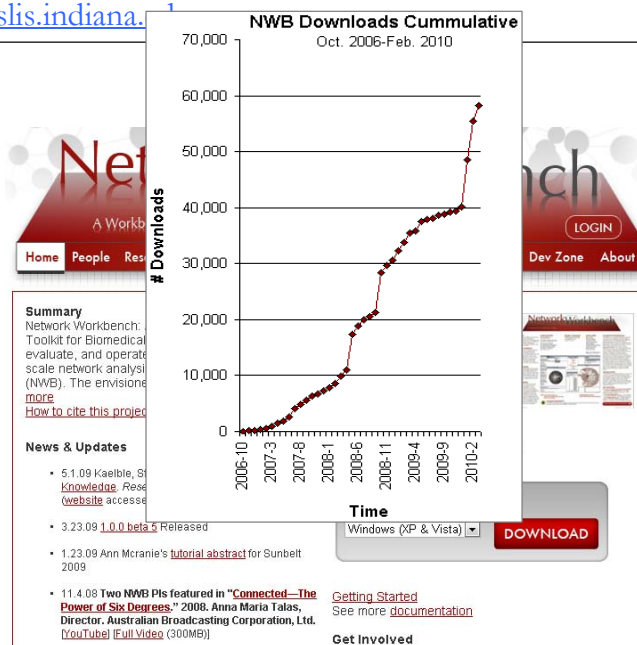
8

The Network Workbench (NWB) tool supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks.

In February 2009, the tool provides more than 169 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

More than 50 of these plugins can be applied or were specifically designed for S&T studies.

It has been downloaded more than 65,000 times since December 2006.



Börner, Katy, Huang, Weixia (Bonnie), Linnemeier, Micah, Dubon, Russell Jackson, Phillips, Patrick, Ma, Nianli, Zoss, Angela, Guo, Hanning & Price, Mark. (2010). Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Network Workbench Tool. *Scientometrics*. Vol. 83(3), 863-876.

Investigators: Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert



Software Team: Lead: Micah Linnemeier
Members: Patrick Phillips, Russell Duhon, Tim Kelley & Ann McCranie
Previous Developers: Weixia (Bonnie) Huang, Bruce Herr, Heng Zhang, Duygu Balcan, Bryan Hook, Ben Markines, Santo Fortunato, Felix Terkhorn, Ramya Sabbineni, Vivek S. Thakre & Cesar Hidalgo



Goal: Develop a large-scale network analysis, modeling and visualization toolkit for physics, biomedical, and social science research.

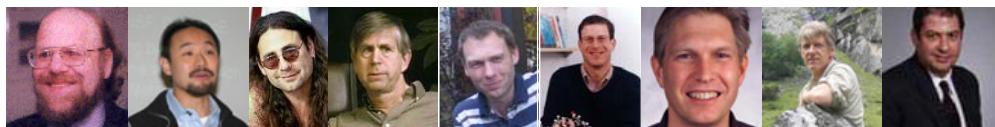
Amount: \$1,120,926, NSF IIS-0513650 award

Duration: Sept. 2005 - Aug. 2009

Website: <http://nwb.slis.indiana.edu>

NWB Advisory Board:

- James Hendler (Semantic Web) <http://www.cs.umd.edu/~hendler/>
- Jason Leigh (CI) <http://www.evl.uic.edu/spiff/>
- Neo Martinez (Biology) <http://online.sfsu.edu/~webhead/>
- Michael Macy, Cornell University (Sociology) <http://www.soc.cornell.edu/faculty/macy.shtml>
- Ulrik Brandes (Graph Theory) <http://www.inf.uni-konstanz.de/~brandes/>
- Mark Gerstein, Yale University (Bioinformatics) <http://bioinfo.mbb.yale.edu/>
- Stephen North (AT&T) <http://public.research.att.com/viewPage.cfm?PageID=81>
- Tom Snijders, University of Groningen <http://stat.gamma.rug.nl/snijders/>
- Noshir Contractor, Northwestern University <http://www.spcomm.uiuc.edu/nosh/>



Computational Proteomics

What relationships exist between protein targets of all drugs and all disease-gene products in the human protein–protein interaction network?

*Yildirim, Muhammed
A., Kwan-II Goh,
Michael E. Cusick,
Albert-László Barabási,
and Marc Vidal. (2007).
Drug-target Network.
Nature Biotechnology
25 no. 10: 1119-1126.*

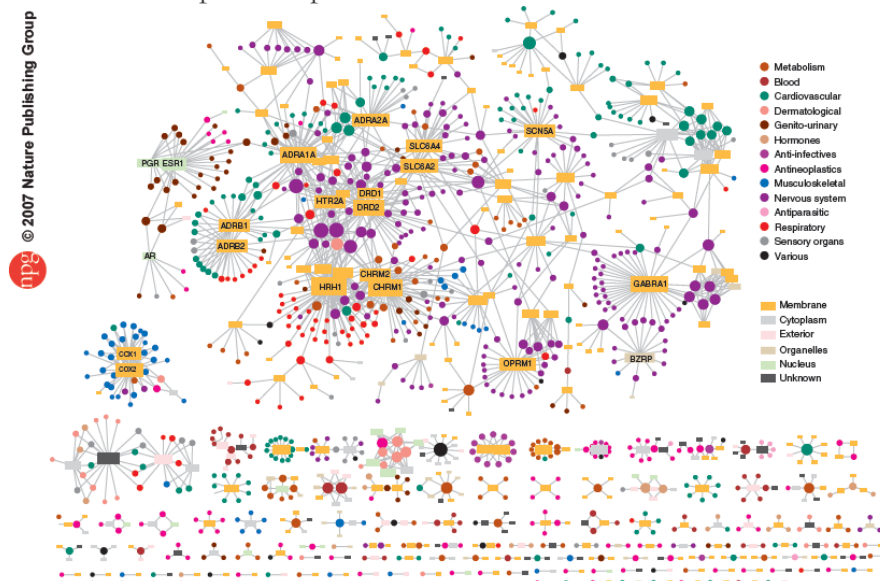


Figure 2 Drug-target network (DT network). The DT network is generated by using the known associations between FDA-approved drugs and their target proteins. Circles and rectangles correspond to drugs and target proteins, respectively. A link is placed between a drug node and a target node if the protein is a known target of that drug. The area of the drug (protein) node is proportional to the number of targets that the drug has (the number of drugs targeting the protein). Color codes are given in the legend. Drug nodes (circles) are colored according to their Anatomical Therapeutic Chemical Classification, and the target proteins (rectangular boxes) are colored according to their cellular component obtained from the Gene Ontology database.

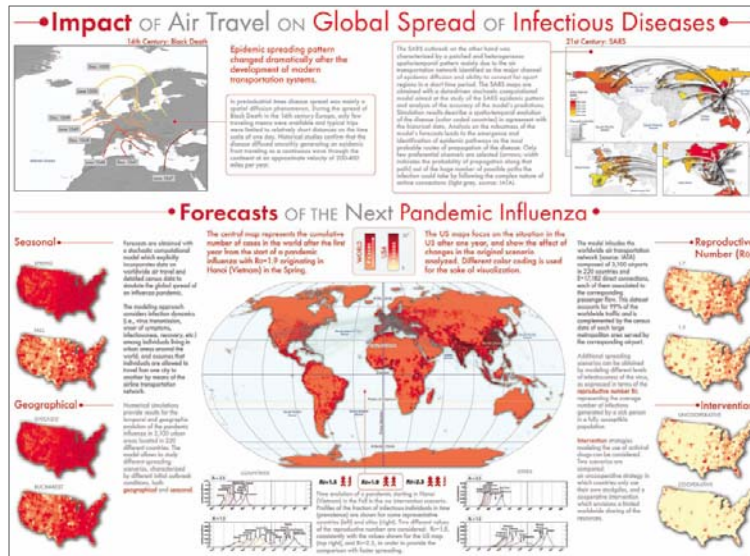
Computational Epidemics

Forecasting (and preventing the effects of) the next pandemic.

Epidemic Modeling in Complex realities, V. Colizza, A. Barrat, M. Barthelemy, A. Vespignani, *Comptes Rendus Biologie*, 330, 364-374 (2007).

Reaction-diffusion processes and metapopulation models in heterogeneous networks, V. Colizza, R. Pastor-Satorras, A. Vespignani, *Nature Physics* 3, 276-282 (2007).

Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions, V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, A. Vespignani, *PLoS-Medicine* 4, e13, 95-110 (2007).



NWB Tool Download, Install, and Run

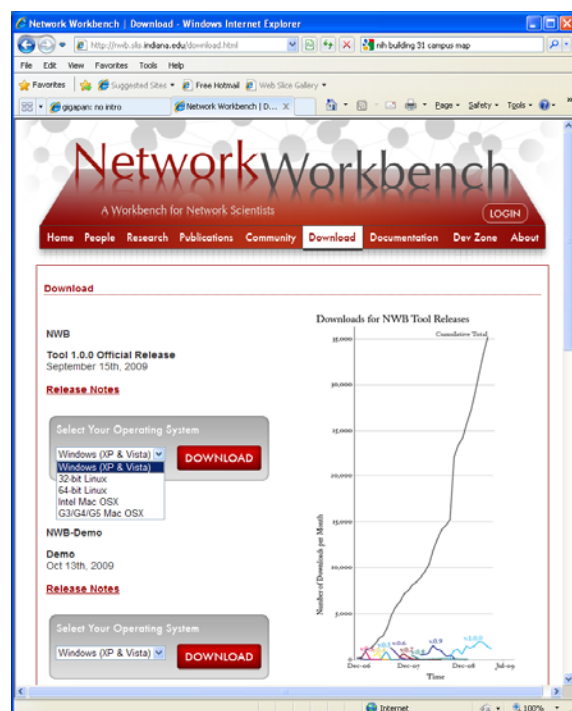
NWB Tool 1.0.0

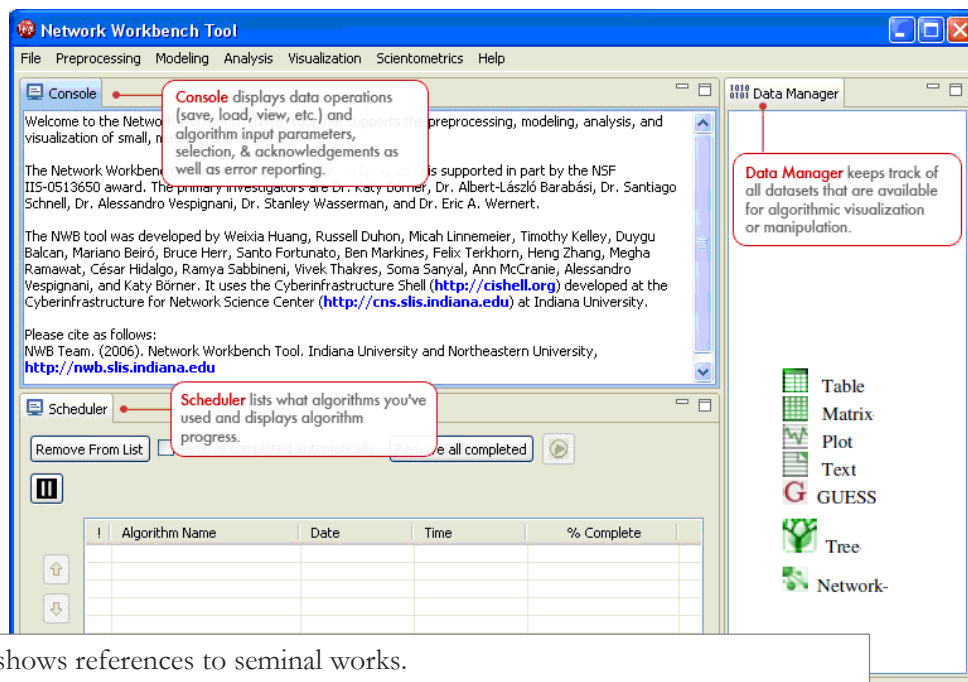
Can be freely downloaded for all major operating systems from <http://nwb.slis.indiana.edu>
Select your operating system from the pull down menu and download.
Unpack into a /nwb directory.
Run /nwb/nwb.exe

Session log files are stored in '*yournwbdirectory*/logs' directory.

Cite as

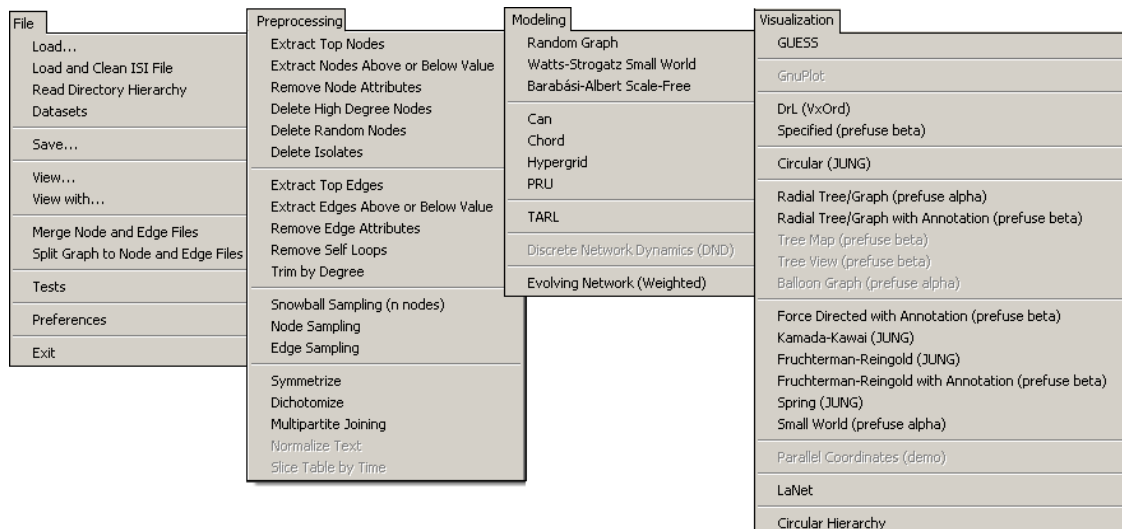
NWB Team. (2006). *Network Workbench Tool*. Indiana University, Northeastern University, and University of Michigan, <http://nwb.slis.indiana.edu>.





Console shows references to seminal works.
Workflows are recorded into a log file, and soon can be re-run for easy replication.
All algorithms are documented online; workflows are given in tutorials.

17



Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (2007). **Network Science**. In Blaise Cronin (Ed.), *ARIST*, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, Volume 41, Chapter 12, pp. 537-607.

<http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf>

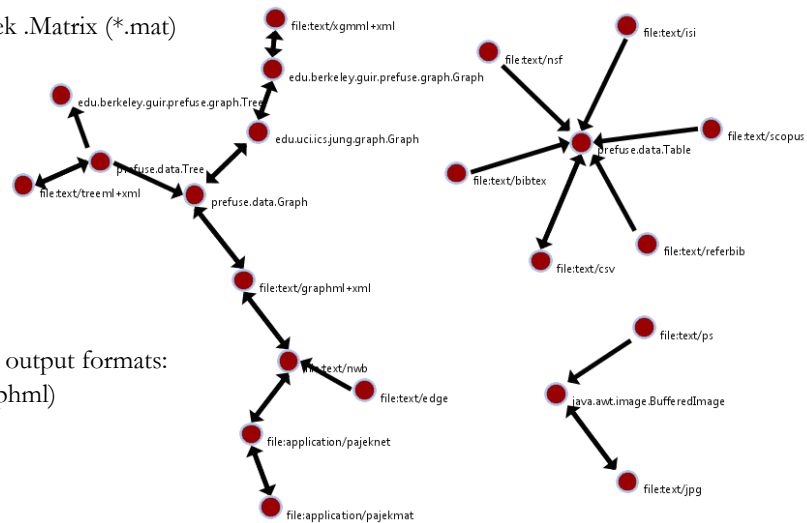
18

The NWB tool supports loading the following input file formats:

- GraphML (*.xml or *.graphml)
- XGMML (*.xml)
- Pajek .NET (*.net) & Pajek .Matrix (*.mat)
- NWB (*.nwb)
- TreeML (*.xml)
- Edge list (*.edge)
- CSV (*.csv)
- ISI (*.isi)
- Scopus (*.scopus)
- NSF (*.nsf)
- Bibtex (*.bib)
- Endnote (*.enw)

and the following network file output formats:

- GraphML (*.xml or *.graphml)
- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)
- CSV (*.csv)



Formats are documented at <https://nwb.slis.indiana.edu/community/?n=DataFormats.HomePage>.



Workshop Overview

9:00 Marcoscope Design and Usage & CShell Powered Tools

9:15 Sci2 Tool Basics

- **Download and run the tool.**

9:30 Sci2 Workflow Design: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

10:00 Sci2 Tool – Using Text and Database Support

- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance

CIShell Powered Tools: Science of Science (Sci2) Tool

23



Science of Science (Sci2) Tool

<http://sci.slis.indiana.edu>

- Explicitly designed for SoS research and practice, well documented, easy to use.
- Empowers many to run common studies while making it easy for exports to perform novel research.
- Advanced algorithms, effective visualizations, and many (standard) workflows.
- Supports micro-level documentation and replication of studies.
- Is open source—anybody can review and extend the code, or use it for commercial purposes.

nature

OPINION

SUMMARY

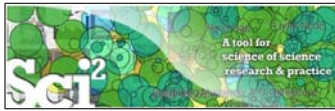
- Existing metrics have known flaws
- A reliable, open, joined-up data infrastructure is needed
- Data should be collected on the full range of scientists' work
- Social scientists and economists should be involved

Vol 464|25 March 2010

Let's make science metrics more scientific

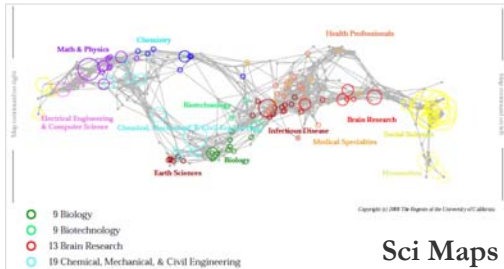
To capture the essence of good science, stakeholders must combine forces to create an open, sound and consistent system for measuring all the activities that make up academic productivity, says **Julia Lane**.

24

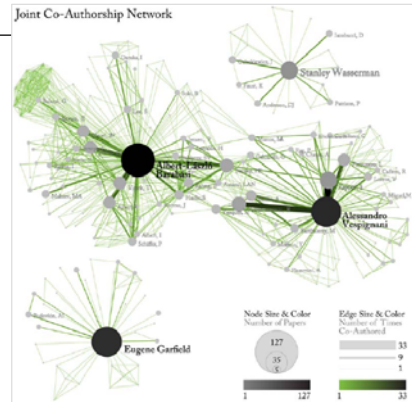


Sci² Tool – “Open Code for S&T Assessment”

OSGi/CIShell powered tool with NWB plugins and many new scientometrics and visualizations plugins.

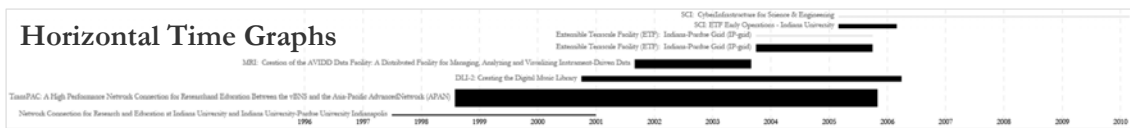


Sci Maps



GUESS Network Vis

Horizontal Time Graphs



Börner, Katy, Huang, Weixia (Bonnie), Linnemeier, Micab, Dubon, Russell Jackson, Phillips, Patrick, Ma, Nianli, Zoss, Angela, Guo, Hanning & Price, Mark. (2009). *Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Scholarly Database and the Network Workbench Tool. Proceedings of ISSI 2009: 12th International Conference on Scientometrics and Informetrics, Rio de Janeiro, Brazil, July 14-17. Vol. 2, pp. 619-630.*

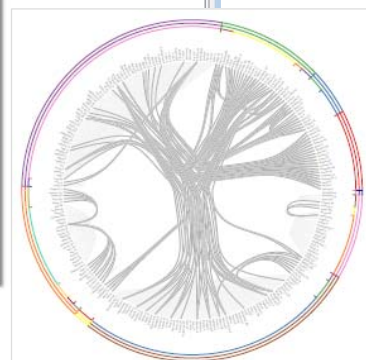
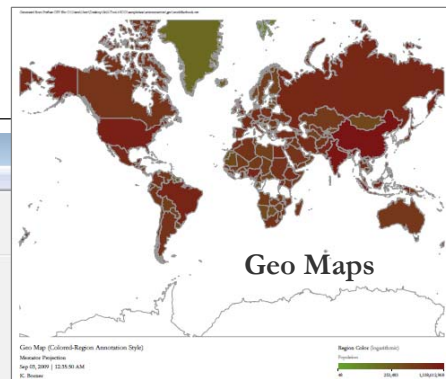


Sci² Tool

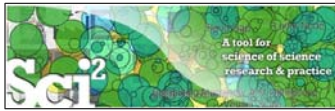
Visualization Menu:

- GUESS
- GnuPlot
- Radial Tree/Graph (prefuse alpha)
- Radial Tree/Graph with Annotation (prefuse beta)
- Tree View (prefuse beta)
- Tree Map (prefuse beta)
- Force Directed with Annotation (prefuse beta)
- Fruchterman-Reingold with Annotation (prefuse beta)
- DrL (VxOrd)
- Specified (prefuse beta)
- Horizontal Line Graph
- Circular Hierarchy
- Geo Map (circle annotations)
- Geo Map (region coloring annotations)
- Image Viewer
- RefMapper

!	Algorithm Name	Date	Time	% Cor
✓	Extract Co-Author Netw...	09/03/2009	00:15:20 AM	100%
✓	Load and Clean ISI File	09/03/2009	00:15:05 AM	100%



Circular Hierarchy



Sci² Tool: Download, Install, and Run

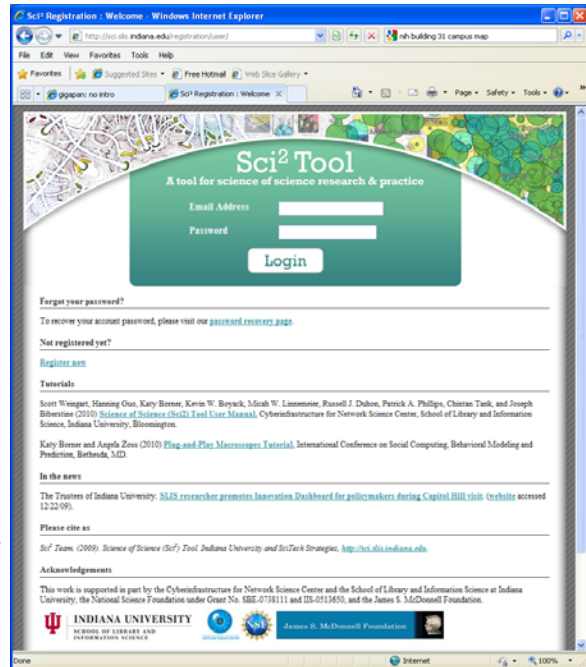
Sci² Tool Alpha 3 (March 2010)

Can be freely downloaded for all major operating systems from <http://sci.slis.indiana.edu/sci2>
 Select your operating system from the pull down menu and download.
 Unpack into a /sci2 directory.
 Run /sci2/sci2.exe

Tutorial is linked from web page.

Cite as

Sci² Team. (2009). Science of Science (Sci²) Tool. Indiana University and SciTech Strategies, <http://sci.slis.indiana.edu>



27



Sci² Tool: Download, Install, and Run

Sci² Tool Alpha 4 (Aug 2010)

Has new features such as

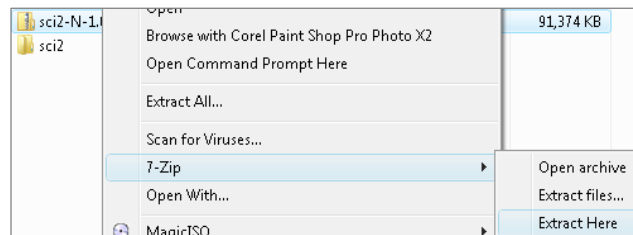
- STAR database
- Colored Horizontal Bar Graphs
- Supports ASCII UTF-8 characters
- Bug fixes, streamlined workflows

NEH Tutorial beta version was made available on DVD for Windows, Mac, and Linux. Unzip and run /sci2/sci2.exe

Name	Size
0 - 9 (1)	
2010-03-sci2-manual.pdf	10,947 KB
Q - Z (4)	
sci2-N-1.0.0.201008130505NGT-macosx.carbon.ppc.zip	91,415 KB
sci2-N-1.0.0.201008130505NGT-macosx.carbon.x86.zip	91,414 KB
sci2-N-1.0.0.201008130505NGT-macosx.cocoa.x86_64.zip	89,717 KB
sci2-N-1.0.0.201008130505NGT-win32.win32.x86.zip	91,374 KB

Pre-release of tutorial is at

<http://ella.slis.indiana.edu/~katy/outgoing/Sci2TutorialAlpha4-NEH-Workshop.pdf>



28

Contents

Contents.....	35
1 Introduction.....	35
2 Getting Started.....	
2.1 Download, Install, Uninstall.....	
2.2 User Interface.....	
2.2.1 Menus.....	
2.2.2 Console.....	
2.2.3 Data Manager.....	
2.2.4 Scheduler.....	
2.3 Data Formats.....	
2.4 Saving Visualizations for Publication.....	
2.5 Sample Datasets.....	
3 Algorithm and Tool Plugins.....	
3.1 Sci ² Tool Plugins.....	
3.2 Load, View, and Save Data.....	
3.3 Memory Allocation.....	
3.3.1 Windows and Linux.....	
3.3.2 Mac.....	
3.4 Memory Limits.....	
4 Workflow Design.....	
4.1 Overview.....	
4.2 Data Acquisition and Preparation.....	
4.2.1 Datasets: Publications.....	
4.2.2 Datasets: Funding.....	
4.2.3 Datasets: Scholarly Database.....	
4.3 Database Loading and Manipulation.....	
4.4 Summaries and Table Extractions.....	
4.5 Statistical Analysis/Profiling.....	
4.6 Temporal Analysis (When).....	
4.6.1 Burst Detection.....	
4.6.2 Slice Table by Time.....	
4.7 Geospatial Analysis (Where).....	
4.8 Topical Analysis (What).....	
4.8.1 Word Co-Occurrence Network.....	
4.9 Network Analysis (With Whom?).....	35
4.9.1 Network Extraction.....	35
4.9.2 Compute Basic Network Characteristics.....	
4.9.3 Network Analysis.....	
4.9.4 Network Visualization.....	
4.10 Modeling (Why?).....	
4.10.1 Random Graph Model.....	
4.10.2 Watts-Strogatz Small World.....	
4.10.3 Barabási-Albert Scale Free Model.....	
5 Sample Workflows.....	
5.1 Individual Level Studies - Micro.....	
5.1.1 Mapping Collaboration, Publication and Funding Profiles of One Rese.....	
5.1.2 Time Slicing of Co-Authorship Networks (ISI Data).....	
5.1.3 Funding Profiles of Three Researchers at Indiana University (NSF Data).....	
5.1.4 Studying Four Major NetSci Researchers (ISI Data).....	
5.1.5 Studying Four Major NetSci Researchers (ISI Data) using Database.....	
5.2 Institution Level Studies - Meso.....	
5.2.1 Funding Profiles of Three Universities (NSF Data).....	
5.2.2 Funding Profiles of Three Universities (NSF Data) Using Database.....	
5.2.3 Mapping CISA Centers (NIH RePORTER Data).....	
5.2.4 Biomedical Funding Profile of NSF (NSF Data).....	
5.2.5 Mapping Scientometrics (ISI Data).....	
5.2.6 Burst Detection in Scientometrics (ISI Data).....	
5.2.7 Mapping the Field of RNAi Research (SDB Data).....	
5.3 Global Level Studies - Macro.....	
5.3.1 Geo USPTO (SDB Data).....	
6 Sample Science Studies & Online Services.....	
6.1 Science Dynamics.....	
6.1.1 Mapping Topics and Topic Bursts in PNAS (2004).....	
6.2 Local Impact-Output / 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).....	87
6.2.2 Mapping Transdisciplinary Tobacco Use Research Centers Publications (forthcoming).....	88
6.3 Local and Global Science Studies.....	89
6.3.1 Mapping the Evolution of Co-Authorship Networks (2004).....	89
6.3.2 Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams (2005).....	90
6.3.4 Mapping the Diffusion of Information Among Major U.S. Research Institutions (2006).....	92
6.3.5 Research Collaborations by the Chinese Academy of Sciences (2009).....	93
6.3.6 Mapping the Structure and Evolution of Chemistry Research (2009).....	94
6.3.7 Science Map Applications: Identifying Core Competency (2007).....	95
6.4 Modeling Science.....	96
6.4.1 113 Years of Physical Review: Using Flow Maps to Show Temporal and Topical Citation (2008).....	96
6.4.2 The Simultaneous Evolution of Author and Paper Networks (2004).....	97
6.5 Accuracy Studies.....	98
6.5.1 Mapping the Backbone of Science (2005).....	98
6.5.2 Toward a Consensus Map of Science (2009).....	99
6.6 Databases and Tools.....	100
6.6.1 The Scholarly Database and Its Utility for Scientometrics Research (2009).....	100
6.6.2 Reference Mapper.....	101
6.6.3 Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Scholarly Database and the Network Workbench Tool (2009).....	102
6.7 Interactive Online Services.....	103
6.7.1 The NIH Visual Browser: An Interactive Visualization of Biomedical Research (2009).....	103
6.7.2 Interactive World and Science Map of S&T Jobs (2010).....	104
7 Extending the Sci ² Tool.....	105
7.1 CShell Basics.....	105
7.2 Read New Data.....	105
7.3 Creating and Sharing New Algorithm Plugins.....	105
7.4 Tools That Use OSGI and/or CShell.....	106
8 Relevant Datasets and Tools.....	107
8.1 Datasets.....	107
8.2 Network Analysis Tools.....	108
9 References.....	111

Sci2 Manual

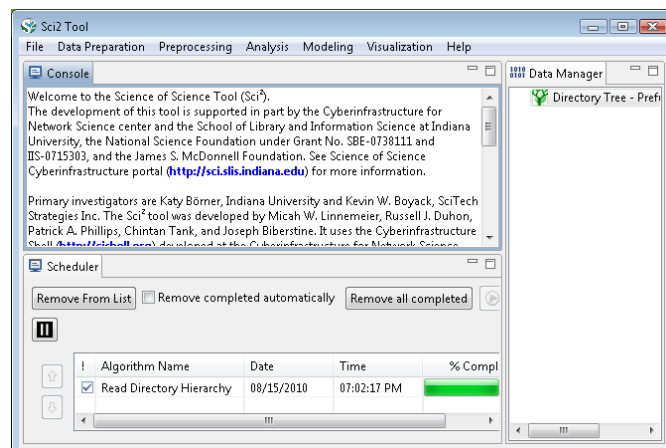
140+ pages



Sci2 Tool Interface Components

Use

- **Menu** to read data, run algorithms.
- **Console** to see work log, references to seminal works.
- **Data Manager** to select, view, save loaded, simulated, or derived datasets.
- **Scheduler** to see status of algorithm execution.



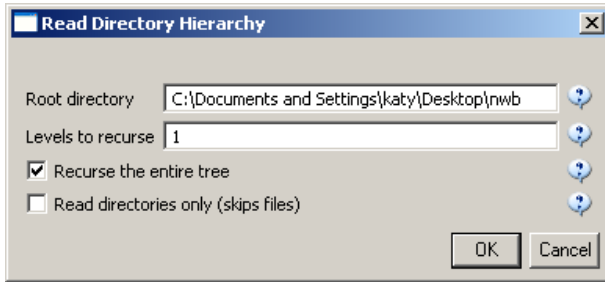
All workflows are recorded into a log file (see /sci2/logs/...), and soon can be re-run for easy replication. If errors occur, they are saved in an error log to ease bug reporting.

All algorithms are documented online; workflows are given in tutorials, see <http://sci.slis.indiana.edu/sci2> and <http://nwb.slis.indiana.edu> > Community



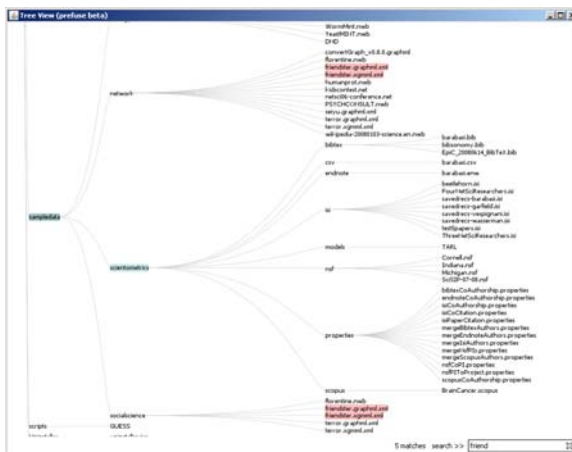
Sci2 Tool – Read+Visualize Sci2 Tool Directory Tree

Use 'File > Read Directory Hierarchy' with parameters



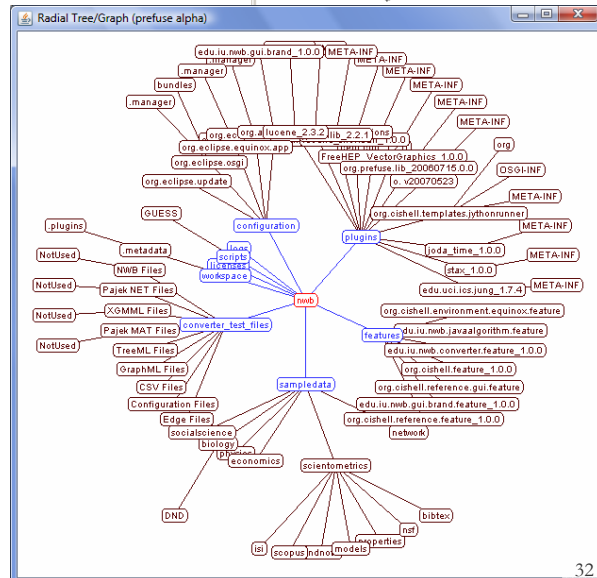
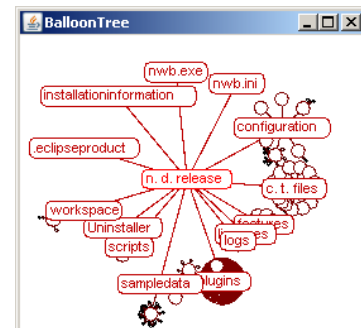
Visualize resulting 'Directory Tree - Prefuse (Beta) Graph' using

- 'Visualization > Tree View (prefuse beta)'
- 'Visualization > Tree Map (prefuse beta)'
- 'Visualization > Balloon Graph (prefuse alpha)'
- 'Visualization > Radial Tree/Graph (prefuse alpha)'



Different views of the /nw b directory hierarchy.

Note the size of the /plugin directory.





Workshop Overview

9:00 Marcoscope Design and Usage & CShell Powered Tools

9:15 Sci2 Tool Basics

- Download and run the tool.

9:30 Sci2 Workflow Design: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

10:00 Sci2 Tool – Using Text and Database Support

- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance

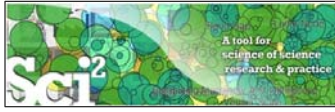
33



Padgett's Florentine Families - Compute Basic Network Properties & View in GUESS

- Florentine families related through business ties (specifically, recorded financial ties such as loans, credits and joint partnerships) and marriage alliances.
- Node attributes
 - Wealth: Each family's net wealth in 1427 (in thousands of lira)
 - Priorates: The number of priorates (seats on the civic council) held between 1282- 1344
 - Totalties: The total number of business or marriage ties in the total dataset of 116 families.
- “Substantively, the data include families who were locked in a struggle for political control of the city of Florence around 1430. Two factions were dominant in this struggle: one revolved around the infamous Medicis, the other around the powerful Strozziis.”
- <http://svitsrv25.epfl.ch/R-doc/library/ergm/html/florentine.html>

34



Padgett's Florentine Families - Compute Basic Network Properties & View in GUESS

- Load `*yoursci2directory*/sampledata/socialscience/florentine.mwb`
- Run 'Analysis > Network Analysis Toolkit (NAT)' to get basic properties.

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

marriage...T

business...F

Average degree: 3.375

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

- Optional: Run 'Analysis > Unweighted & Undirected > Node Betweenness Centrality' with default parameters.
- Select network and run 'Visualization > GUESS' to open GUESS with file loaded.
- Apply 'Layout > GEM'.

35

The screenshot shows the Network Workbench Tool interface. The main window displays a network graph with blue nodes and yellow edges. The graph is titled 'Medici-Acciaiuoli'. The 'Information Window' shows the following table:

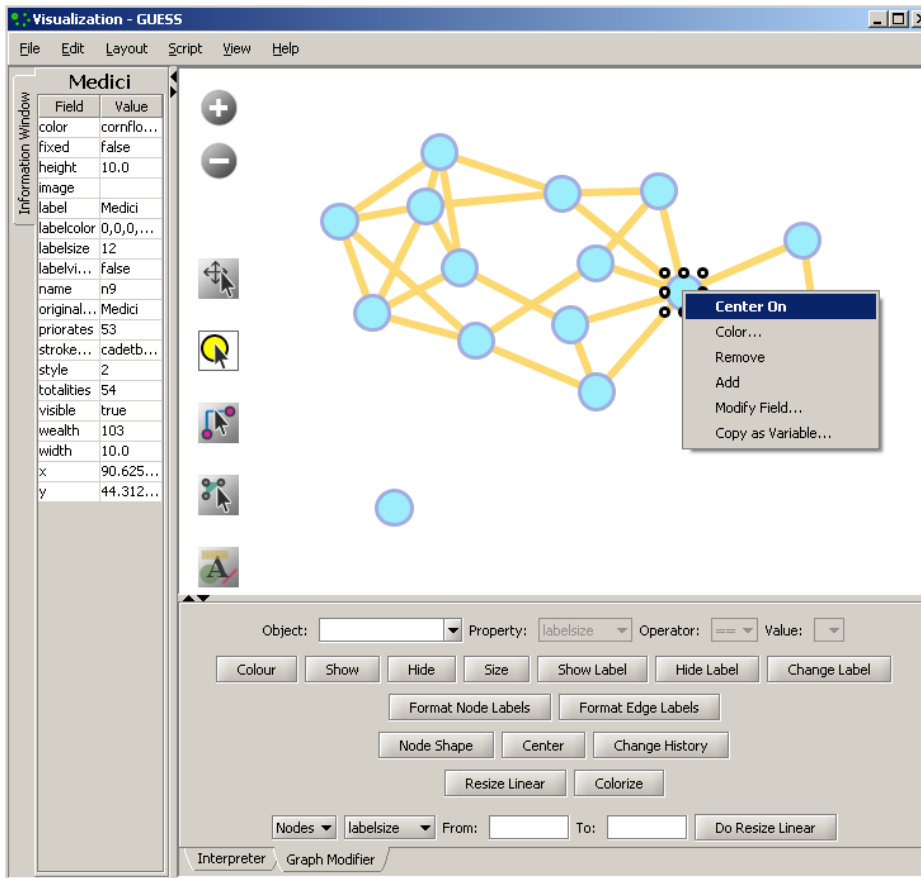
Field	Value
_edgeid	0
business	F
color	dandelion
directed	false
label	
labelcolor	0,0,0,255
labelsize	12
labelvisible	false
marriage	T
node1	n9
node2	n1
visible	true
weight	1.0
width	2.0

The 'Console' window shows the following output:

```
.....
GUESS was selected.
Author(s): Eytan Adar
Implementer(s): Eytan Adar (GUESS), Russell Duhon (resizeLinear, colorize fix)
Integrator(s): Russell Duhon
Reference: Adar, Eytan, "GUESS: A Language and Interface for Graph Exploration," CHI 2006 (http://graphexploration.cond.org/)
Documentation: https://nwb.sls.indiana.edu/community/?m=VisualizeData.GUESS
ECHO is off.
Starting GUESS...
ECHO is off.
The initial layout for your visualization is random. For a clearer visualization, please run a layout from the Layout menu. (We recommend GEM.)
ECHO is off.
GUESS log file for this session can be found in
C:\DOCUMENT~1\katey\LOCALS~1\Temp\CIShell-Session-55892\StaticExecutableRunner-55904\ALGORITHM\guesslog.txt
.....
GUESS was selected.
Author(s): Eytan Adar
Implementer(s): Eytan Adar
Integrator(s): Russell Duhon
Reference: Adar, Eytan, "GUESS: A Language and Interface for Graph Exploration," CHI 2006 (http://graphexploration.cond.org/)
Documentation: https://nwb.sls.indiana.edu/community/?m=VisualizeData.GUESS
ECHO is off.
Starting GUESS...
ECHO is off.
The initial layout for your \
ECHO is off.
GUESS log file for this sess
C:\DOCUMENT~1\katey\LOCA
```


The 'Data Manager' window shows a list of files, including 'NWB file: C:\Documents and Settings\katey\I...', 'Distribution of degree for network at st...', 'List of edges of network created through th...', 'Distribution of degree for network at st...', 'NWB file: C:\Documents and Settings\katey\I...', 'Graph and Network Analysis Log', 'Sequence of site betweennesses for no...', 'Distribution of site betweenness for net...', 'Distribution of site betweenness for net...', and 'NWB file with site betweenness attribut...'. The 'Scheduler' window shows a list of algorithms, including 'GUESS', 'Node Betw', 'Network Ar', 'Load...', 'GnuPlot', 'Degree Dis', 'Barabási-A', 'GnuPlot', 'Degree Dis', and 'GUESS'. The 'Visualization - GUESS' window shows a menu with 'File', 'Edit', 'Layout', 'Script', and 'View'. The 'Graph Modifier' window shows a table with 'Object', 'Property', 'Operator', and 'Value' columns, and buttons for 'Colour', 'Show', 'Hide', 'Size', 'Show Label', 'Hide Label', 'Change Label', 'Node Shape', 'Center', and 'Change History'.

36

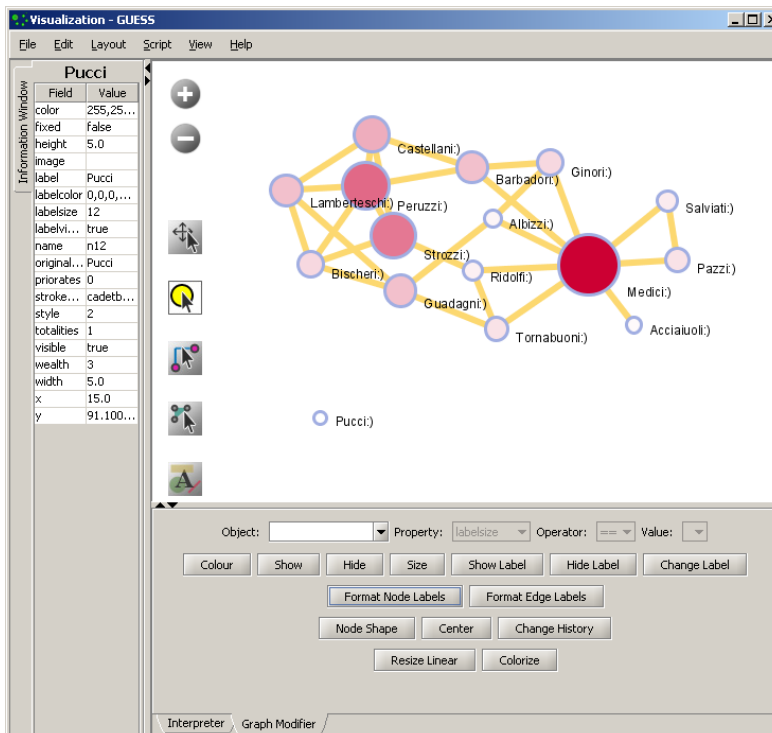


Pan:
“grab” the background by holding left-click and moving your mouse.

Zoom:
Using scroll wheel, press the “+” and “-” buttons in the upper-left hand corner, or right-click and move the mouse left or right. Center graph by selecting ‘View -> Center’.

Select  to select/move single nodes. Hold down ‘Shift’ to select multiple.

Right click to modify Color, etc.



Graph Modifier:

Select “all nodes” in the Object drop-down menu and click ‘Show Label’ button.

Select ‘Resize Linear > Nodes > totalities’ drop-down menu, then type “5” and “20” into the From” and To” Value box separately. Then select ‘Do Resize Linear’.

Select ‘Colorize> Nodes>totalities’, then select white and enter (204,0,51) in the pop-up color boxes on in the “From” and “To” buttons.

Select “Format Node Labels”, replace default text {originallabel} with your own label in the pop-up box ‘Enter a formatting string for node labels.’

Visualization - GUESS

File Edit Layout Script View Help

Information Window

Field	Value
color	255,24...
fixed	False
height	5.6635...
image	
label	Acciaiu...
labelcolor	0,0,0,...
labelsize	12
labelvi...	true
name	n1
original...	Acciaiuoli
priorates	53
stroke...	cadetb...
style	2
totalities	2
visible	true
wealth	10
width	5.6635...
x	112.01...
y	70.315...

Interpreter:
Uses Jython a combination of Java and Python.

Try
colorize(wealth, white, red)

```
resizeLinear(sitebetweenness, 5, 25)
```

Interpreter Graph Modifier

39



Workshop Overview

9:00 Marcoscope Design and Usage & CShell Powered Tools

9:15 Sci2 Tool Basics

- Download and run the tool.

9:30 Sci2 Workflow Design: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

10:00 Sci2 Tool – Using Text and Database Support

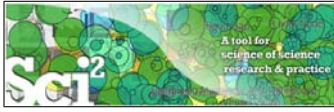
- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance



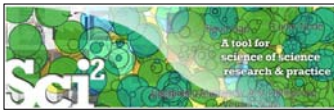
Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.4)

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

Thomson Reuter's Web of Knowledge (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, WoS provides some of the most comprehensive 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.

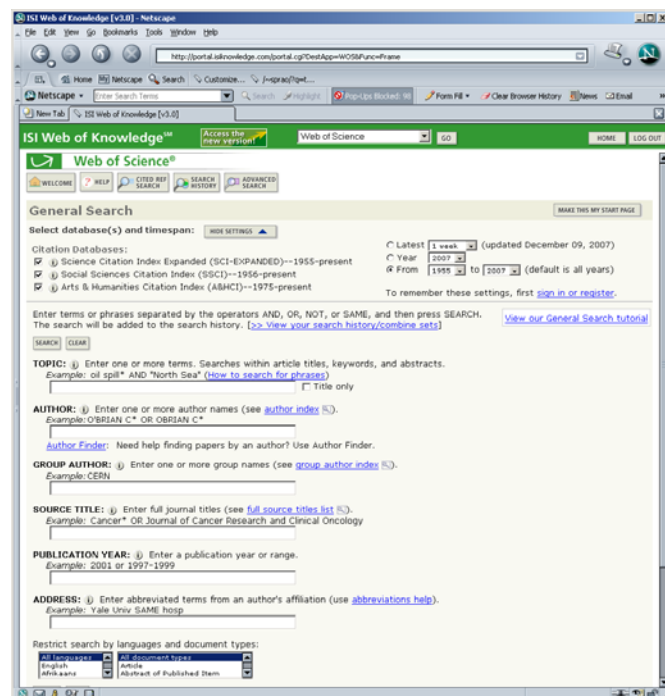
41



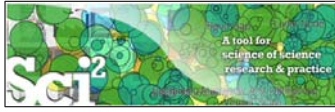
Data Acquisition from Web of Science

Download all papers by

- Eugene Garfield
 - Stanley Wasserman
 - Alessandro Vespignani
 - Albert-László Barabási
- from
- Science Citation Index Expanded (SCI-EXPANDED) --1955-present
 - Social Sciences Citation Index (SSCI)--1956-present
 - Arts & Humanities Citation Index (A&HCI)--1975-present



42

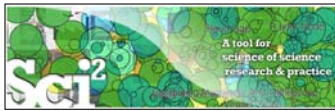


Comparison of Counts

No books and other non-WoS publications are covered.

	Age	Total # Cites	Total # Papers	H-Index
Eugene Garfield	82	1,525	672	31
Stanley Wasserman		122	35	17
Alessandro Vespignani	42	451	101	33
Albert-László Barabási	40	2,218	126	47 (Dec 2007)
	41	16,920	159	52 (Dec 2008)

43



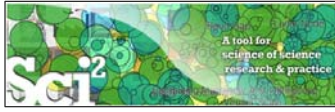
Extract Co-Author Network

Load **yoursci2directory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi'*
 using 'File > Load and Clean ISI File' and parameters

The screenshot shows the Sci2 Tool interface. A 'Load' dialog box is open, displaying a message: 'The file you have selected can be loaded using one or more of the following formats. Please select the format you would like to try.' Below the message, 'Load as...' is set to 'ISI scholarly format' and 'ISI database' is also visible. The 'Data Manager' window shows a list of records: 'ISI Data: C:\Users\User\Desktop\10-NEH-A&H-Workshop' with '361 Unique ISI Records'.

And file with 361 records
 appears in Data Manager.

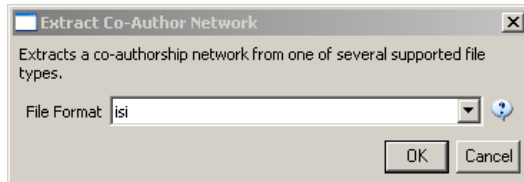
44



Extract Co-Author Network

(see section 5.1.4.2 on correcting duplicate/ misspelled author names)

To extract the co-author network, select the '361 Unique ISI Records' table and run *Data Preparation > Text Files > Extract Co-Author Network* using isi file format:



The result is an undirected but weighted network of co-authors in the Data Manager.

Run *Analysis > Network > Network Analysis Toolkit (NAT)* to calculate basic properties: the network has 247 nodes and 891 edges.

Use *Analysis > Network > Unweighted and Undirected > Node Degree* to calculate the number of neighbors for each node.

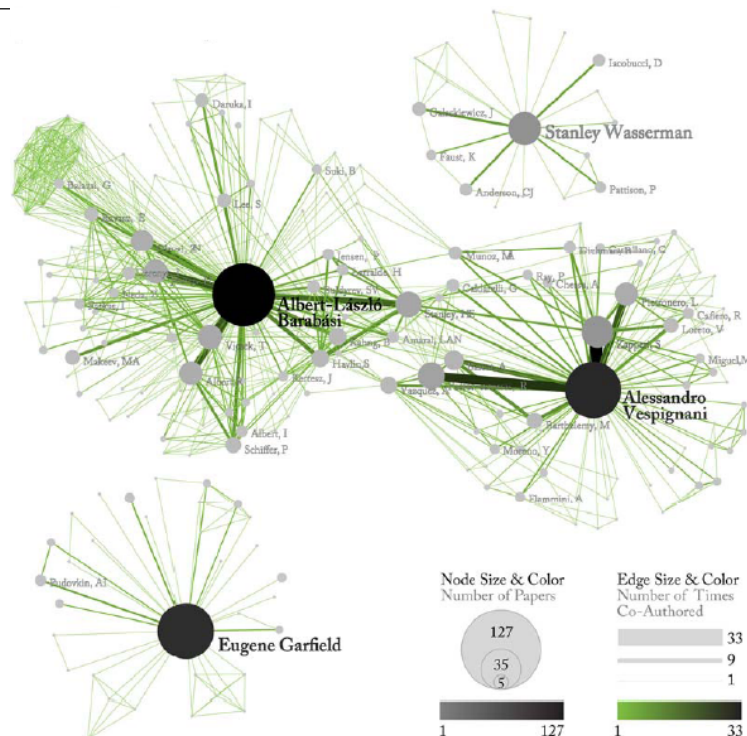
To view the complete network, select the *Extracted Co-Authorship Network* and run *Visualization > Networks > GUESS*.

Network is loaded with random layout. In GUESS, run *Layout > GEM* and *Layout > Bin Pack* to improve layout.

45



Co-Author Network of all Four NetsSci Researchers



46



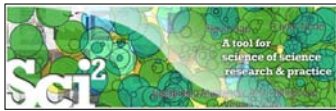
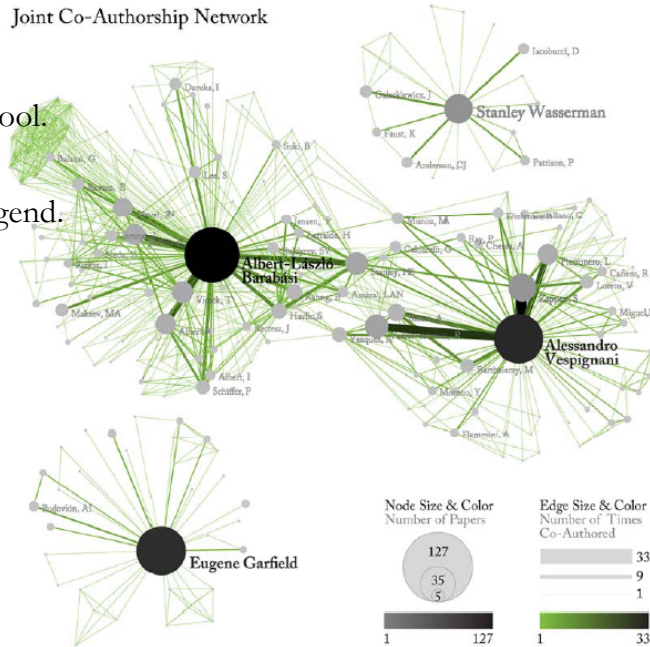
Co-Author Network of all Four NetsSci Researchers

Use the GUESS Graph Modifier to change color and size coding.

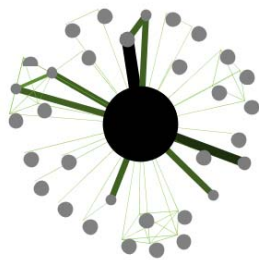
Joint Co-Authorship Network

Calculate node degrees in Sci2 Tool.

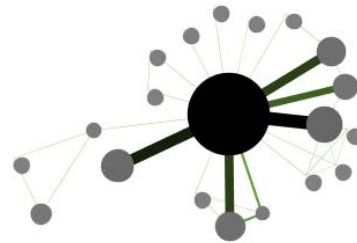
Use a graphic program to add legend.



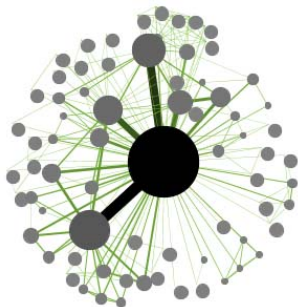
Individual Co-Author Networks (Read/map 4 files separately)



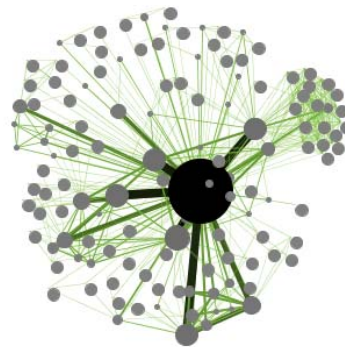
Eugene Garfield



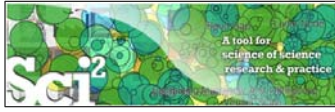
Stanley Wasserman



Alessandro Vespignani

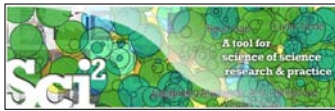
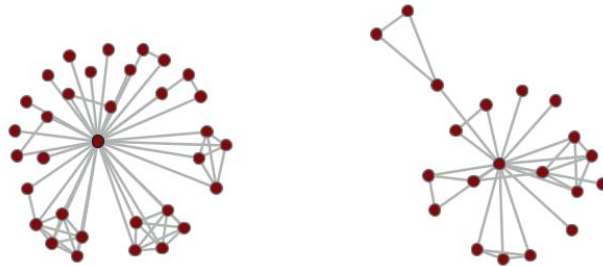
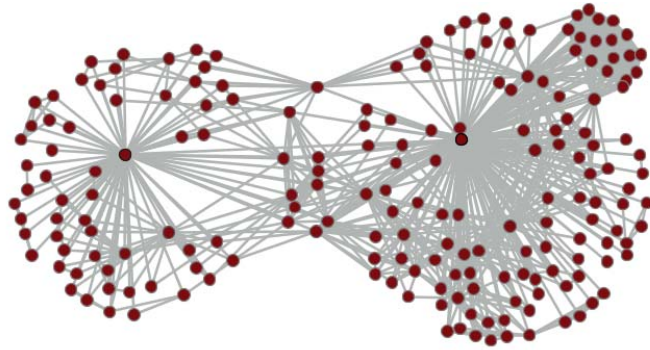


Albert-László Barabási

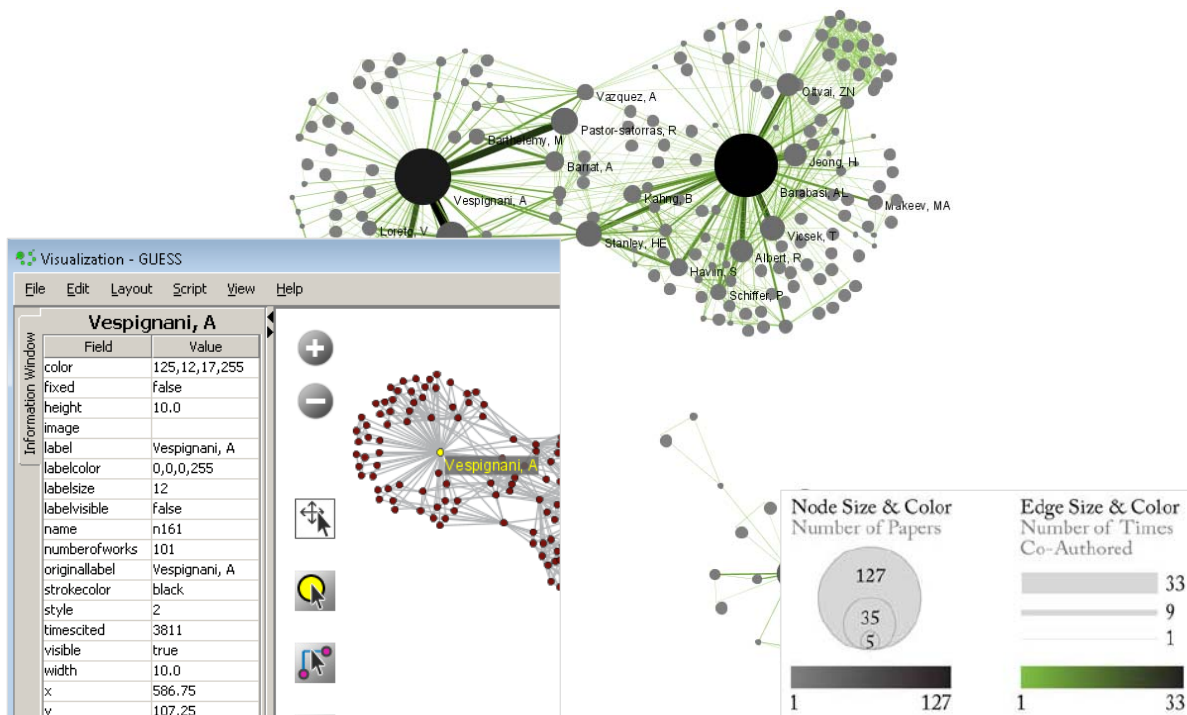


Network Visualization: Node Layout

Load and Clean ISI File was selected.
Loaded 361 records.
Removed 0 duplicate records.
Author names have been normalized.
361 records with unique ISI IDs are available
via Data Manager.
.....
Extract Co-Author Network was selected.
Input Parameters:
File Format: isi
.....
Network Analysis Toolkit (NAT) was selected.
Nodes: 247
Edges: 891
.....
GUESS was selected.

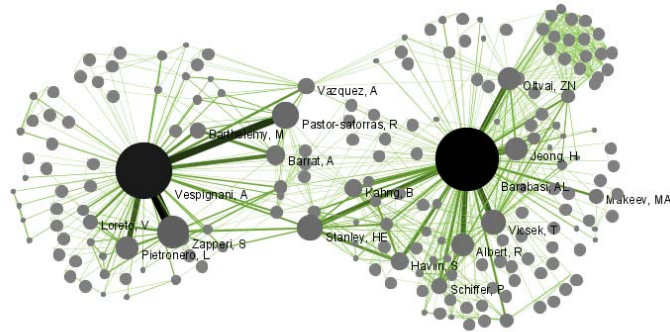


Network Visualization: Color/Size Coding by Data Attribute Values





Network Visualization: Giant Component



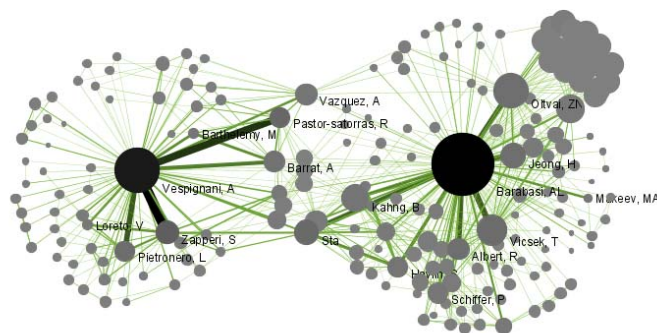
.....
Weak Component Clustering was selected.
Implementer(s): Russell Duhon
Integrator(s): Russell Duhon

Input Parameters:
Number of top clusters: 10
3 clusters found, generating graphs for the top 3 clusters.
.....

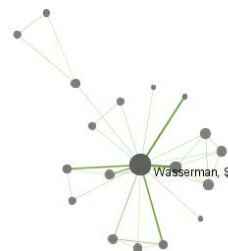
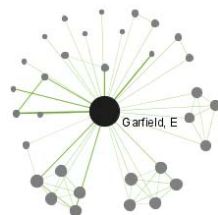
51



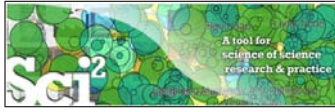
Network Visualization: Color/Size Coding by Degree



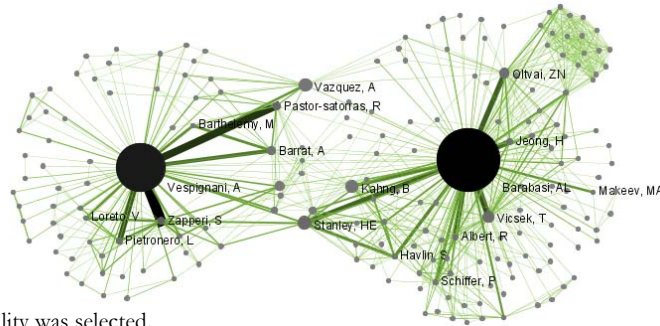
.....
Node Degree was selected.
Documentation:
<https://nwb.slis.indiana.edu/community/?n=AnalyzeData.No deDegree>
.....



52



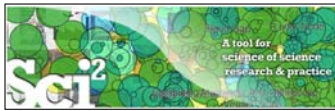
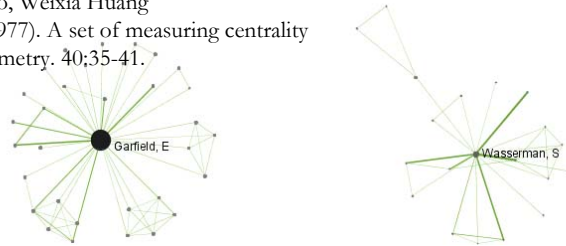
Network Visualization: Color/Size Coding by Betweenness Centrality



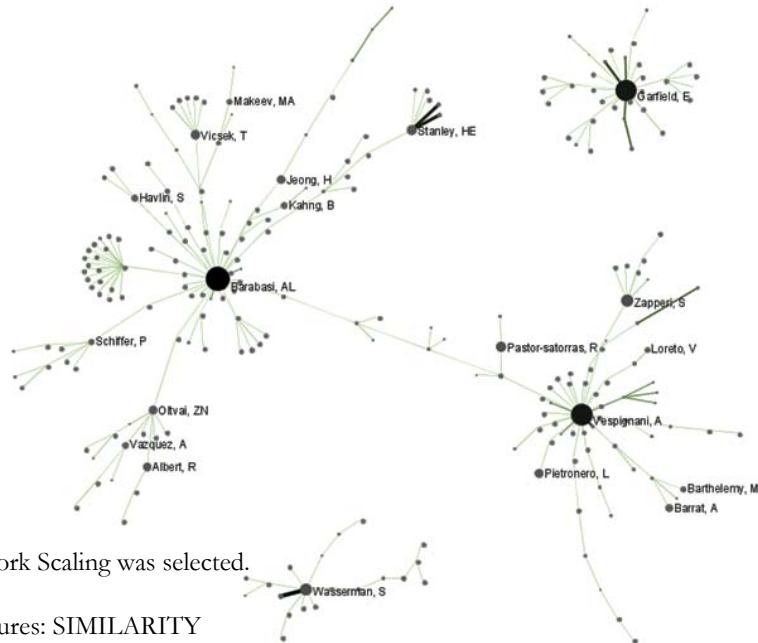
.....
Node Betweenness Centrality was selected.
Author(s): L. C. Freeman
Implementer(s): Santo Fortunato
Integrator(s): Santo Fortunato, Weixia Huang
Reference: Freeman, L. C. (1977). A set of measuring centrality based on betweenness. *Sociometry*. 40;35-41.

Input Parameters:
Number of bins: 10

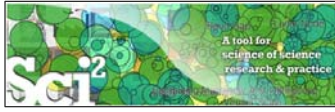
umber of bins: 10
.....



Network Visualization: Reduced Network After Pathfinder Network Scaling

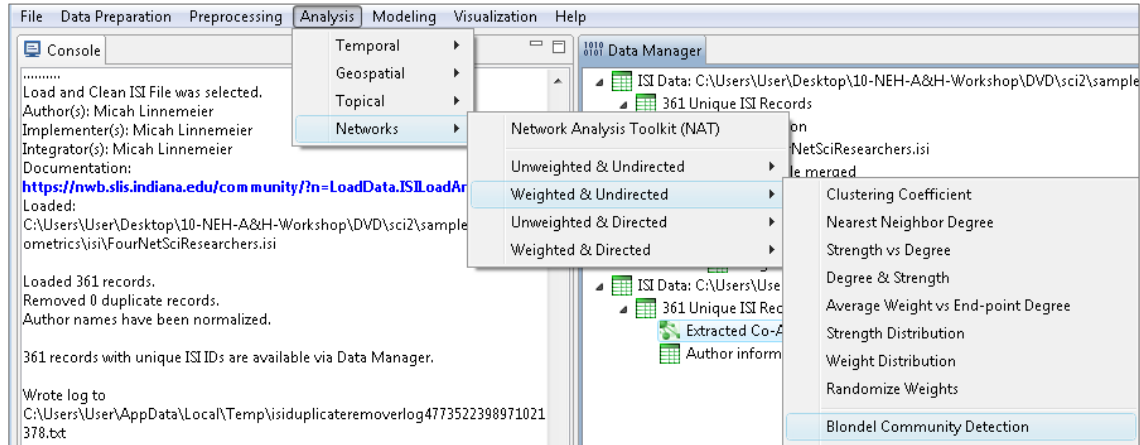


.....
MST-Pathfinder Network Scaling was selected.
Input Parameters:
Weight Attribute measures: SIMILARITY
Edge Weight Attribute: weight
.....

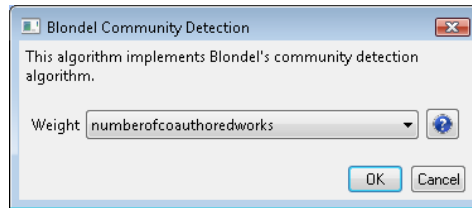


Network Visualization: Circular Hierarchy Visualization

Select Co-Author Network and run Blondel Community detection:



With parameter values

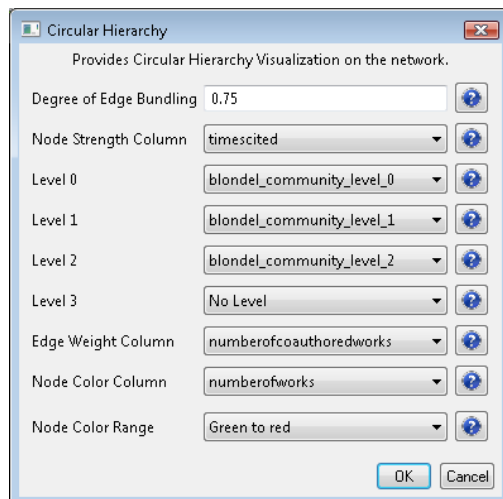


55

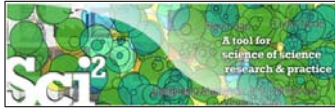


Network Visualization: Circular Hierarchy Visualization

Visualize resulting file using '*Visualization > Networks > Circular Hierarchy*' with parameter values



56



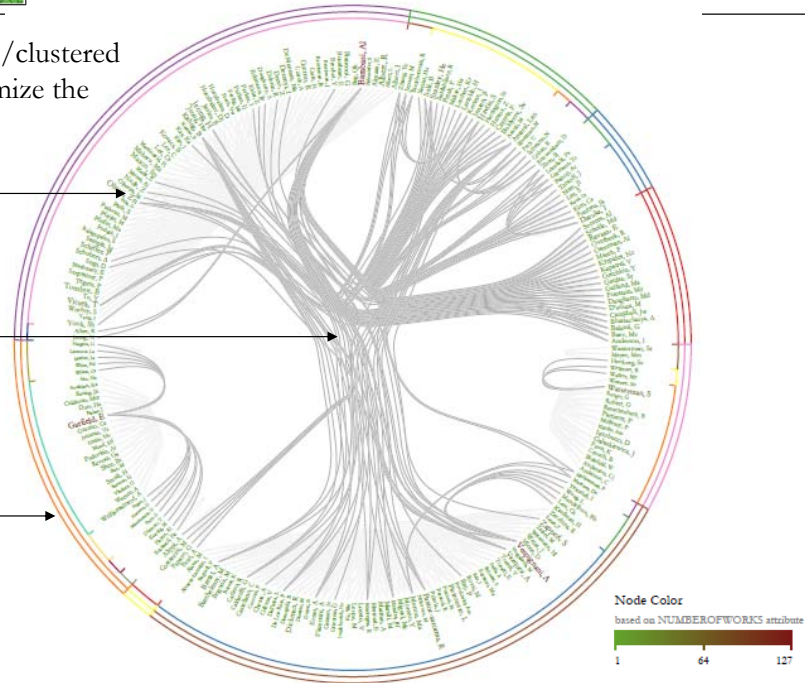
Network Visualization: Circular Hierarchy Visualization

Nodes that are interlinked/clustered are spatially close to minimize the number of edge crossings.

Node labels, e.g., author names.

Network structure using edge bundling.

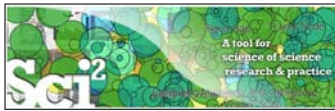
Color coded cluster hierarchy according to Blondel community detection algorithm.



Note:

Header/footer info, legend, and more meaningful color coding are under development.

57



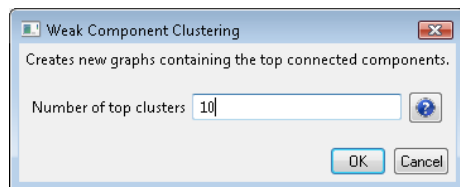
Paper-Citation Network Layout

To extract the paper-citation network, select the '361 Unique ISI Records' table and run *Data Preparation > Text Files > Extract Paper Citation Network.*

The result is a unweighted, directed network of papers linked by citations, named *Extracted paper-citation network* in the Data Manager.

Run *NAT* to calculate that the network has 5,342 nodes and 9,612 edges. There are 15 weakly connected components. (0 isolates)

Run *'Analysis > Networks > Unweighted and Directed > Weak Component Clustering'* with parameters

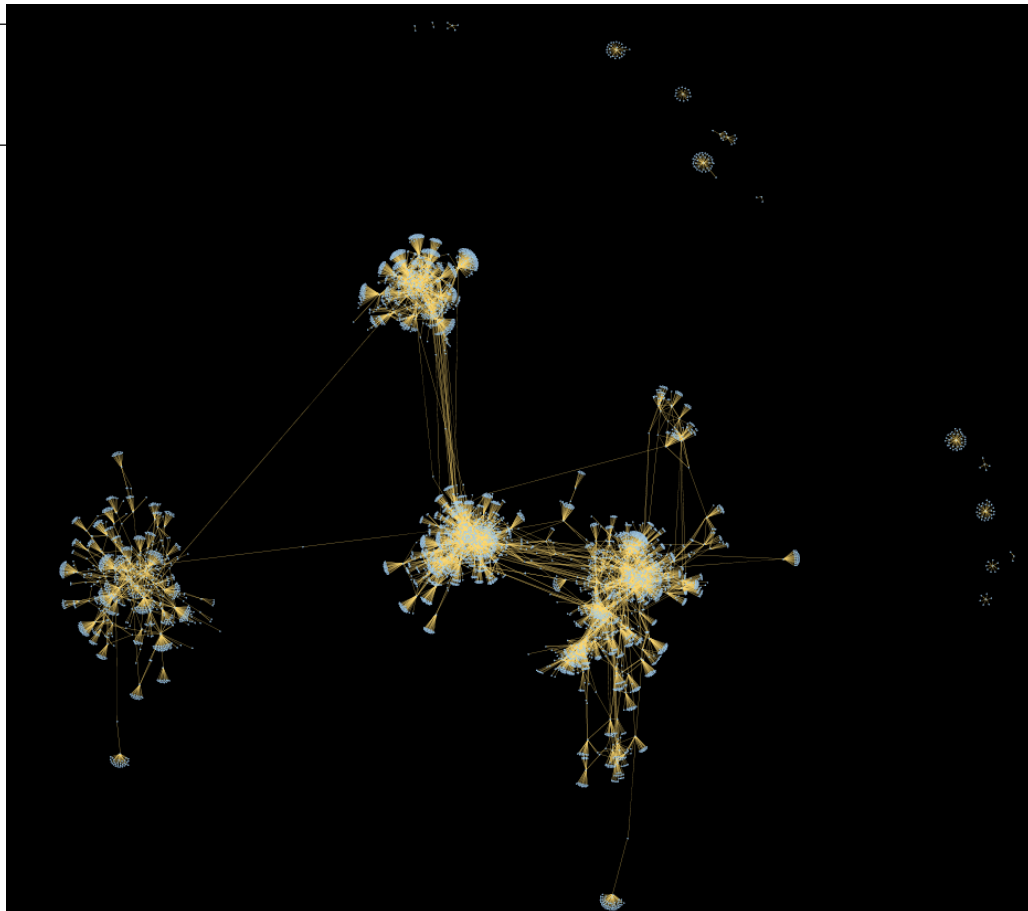


- Weak Component Cluster of 5151 nodes
- Weak Component Cluster of 38 nodes
- Weak Component Cluster of 35 nodes
- Weak Component Cluster of 27 nodes
- Weak Component Cluster of 27 nodes.2
- Weak Component Cluster of 15 nodes

to identify top-10 largest components. The largest (giant) component has 5,151 nodes.

To view the complete network, select the network and run *'Visualization > GUESS'*.

58

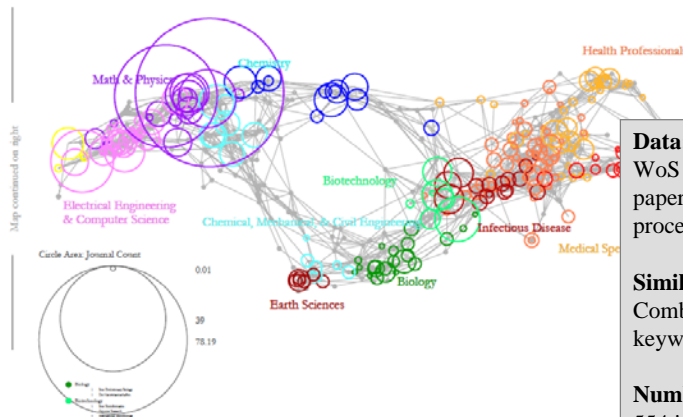


Topic Mapping: UCSD Science Map

Science Map via Journals for FourNetSciResearchers.isi

314 journal references matched out of 361 found.

These 314 references are associated with 13 of 13 disciplines of science and 255 of 554 research specialties in the UCSD Map of Science.



Data:

WoS and Scopus for 2001–2005, 7.2 million papers, more than 16,000 separate journals, proceedings, and series

Similarity Metric:

Combination of bibliographic coupling and keyword vectors

Number of Disciplines:

554 journal clusters further aggregated into 13 main scientific disciplines that are labeled and color coded in a metaphorical way, e.g., Medicine is blood red and Earth Sciences are brown as soil.

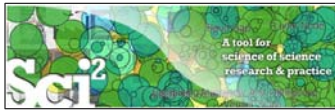
Science Map

Locate the journals from a table on the UCSD Map of Science

Journal column:

Dataset display name:

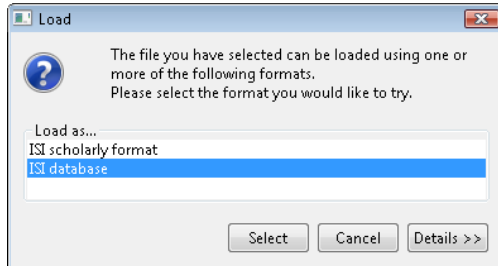
OK Cancel



Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – Load a File

Load **yoursci2directory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi*, using *File > Load* and select



It might take several minutes to load the data into the database.

Select the *ISI Database* in the Data Manager and run *Data Preparation > Database > ISI > ** to calculate statistics, e.g., annual counts, extract networks, prepare data for burst detection, etc.

61



Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – View Database Structure

View the database schema by right-clicking on the loaded database in the Data Manager and clicking “View”

The screenshot shows the Sci2 Data Manager window with the 'ISI Database' selected. A context menu is open over the database, with 'View' selected. The 'Console' window shows the following text:

```
Primary investigators are Katy Börner, Indiana University and Kevin W. Boyack, SciTech Strategies Inc. The Sci2 tool was developed by Micah W. Linnemeier, Russell J. Duhon, Patrick A. Phillips, Chintan Tank, and Joseph Biberstine. It uses the Cyberinfrastructure Shell (http://cishell.org) developed at the Cyberinfrastructure for Network Science Center (http://cns.slis.indiana.edu) at Indiana University. Many algorithm plugins were derived from the Network Workbench Tool http://nwb.slis.indiana.edu.
```

The 'Scheduler' window shows a table of tasks:

#	Algorithm Name	Date
1	Load ISI File Into Database	03/27/2010
2	Load...	03/27/2010

The 'Untitled - Notepad' window shows the database schema:

```
ADDRESS ( PK INTEGER, ADDRESS_CITY VARCHAR, ADDRESS_COUNTF
AUTHORS ( AUTHORS_DOCUMENT_FK INTEGER, AUTHORS_PERSON_FK I
AUTHORS_DOCUMENT_FK -----> DOCUMENT.PK
AUTHORS_PERSON_FK -----> PERSON.PK
CITED_PATENTS ( CITED_PATENTS_DOCUMENT_FK INTEGER, CITED_F
CITED_PATENTS_DOCUMENT_FK -----> DOCUMENT.PK
CITED_PATENTS_PATENT_FK -----> PATENT.PK
CITED_REFERENCES ( CITED_REFERENCES_DOCUMENT_FK INTEGER, C
CITED_REFERENCES_DOCUMENT_FK -----> DOCUMENT.PK
CITED_REFERENCES_REFERENCE_FK -----> REFERENCE.PK
DOCUMENT ( PK INTEGER, ABSTRACT_TEXT VARCHAR, ARTICLE_NUME
FIRST_AUTHOR_FK -----> PERSON.PK
DOCUMENT_SOURCE_FK -----> SOURCE.PK
DOCUMENT_KEYWORDS ( DOCUMENT_KEYWORDS_DOCUMENT_FK INTEGER,
DOCUMENT_KEYWORDS_DOCUMENT_FK -----> DOCUMENT.PK
DOCUMENT_KEYWORDS_KEYWORD_FK -----> KEYWORD.PK
DOCUMENT_OCCURRENCES ( DOCUMENT_OCCURRENCES_DOCUMENT_FK IN
DOCUMENT_OCCURRENCES_DOCUMENT_FK -----> DOCUMENT.PK
DOCUMENT_OCCURRENCES_ISI_FILE_FK -----> ISI_FILES.PK
EDITORS ( EDITORS_DOCUMENT_FK INTEGER, EDITORS_PERSON_FK I
```

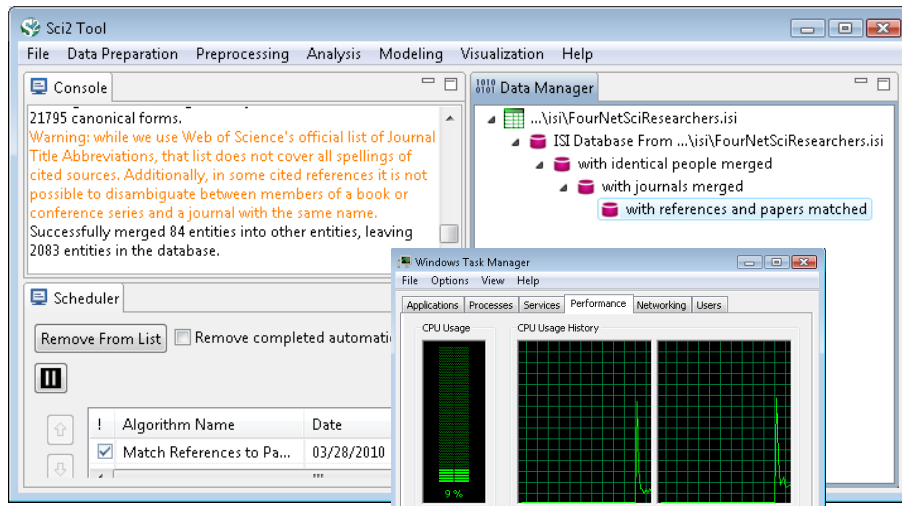
62



Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – Data Unification

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. Read red warnings.



63



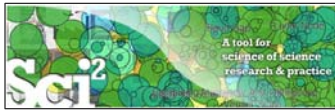
Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – Extract Basic Properties

Run ‘Data Preparation > Database > ISI > Extract Authors’ and right-click on the resulting table to view all the authors from FourNetSciResearchers.isi. The table also has columns with information on how many 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).

	A	B	C	D	E	F	G	H	I	J	K
1	UN_SPLIT_NAME	PAPERS	GLOBAL_CITATION_COUNT	LOCAL_CITATION_COUNT	ADDITIONAL_CITATION_COUNT	FAMILY_NAME	FIRST_INITIAL	FULL_NAME	MIDDLE_INITIAL	PERSONAL_NAME	
2	Barthelemy, M	9	454	12		Barthelemy	M				
3	Barrat, A	13	480	14		Barrat	A				
4	Pastor-satorras, R	24	1769	48		Pastor-satorras	R				
5	Vespignani, A	101	3811	213		Vespignani	A				
6	Wasserman, S	32	675	109		Wasserman	S				
7	Daruka, I	7	392	11		Daruka	I				
8	Makeev, MA	8	198	19		Makeev	M		A		
9	Sidoretti, S	1	1	1		Sidoretti	S				
10	Iacobucci, D	6	115	33		Iacobucci	D				
11	Vazquez, A	10	620	5		Vazquez	A				
12	Oliveira, JG	2	20	0		Oliveira	J		G		
13	Farkas, I	3	47	1		Farkas	I				
14	Jeong, H	17	4160	143		Jeong	H				
15	Oltvai, ZN	17	2961	59		Oltvai	Z		N		
16	Cuerno, R	2	267	11		Cuerno	R				
17	Dobrin, R	2	85	2		Dobrin	R				
18	Beg, GK	1	41	0		Beg	G		K		
19	Pudovkin, AI	5	32	6		Pudovkin	A		I		

64



Using the ISI Database (section 3.1)

Database > ISI

- Extract Authors – Outputs a table containing one row per author in the database, which includes Papers Authored in Dataset, Global Citation Count, and Local Citation Count.
- Extract Documents – Outputs a table containing one row per document in the database, with all bibliographic information related to the document (including abstract text, if available), as well as Times Cited, Cited Reference Count, Digital Object ID (if available), and Funding Information.
- Extract Keywords – Outputs a table containing one row per keyword in the database which includes occurrences of each keyword in the dataset.
- Extract Document Sources – Outputs a table containing one row per document source in the database, which includes the Full Title, Number of Papers Contained from the Dataset, ISSN, and other bibliographic information (if available).

65



Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – Records over time

Aggregate data by year by running ‘Data Preparation > Database > ISI > Extract Longitudinal Summary.’ Result is 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 amount of references made, the amount of distinct references, distinct authors, distinct sources, and distinct keywords per year.

F1		DISTINCT_AUTHORS										
A	B	C	D	E	F	G	H	I	J	K	L	
1	YR	DOCUMENTS	REFERENCES	TOTAL_REFERENCES	DISTINCT_REFERENCES	DISTINCT_AUTHORS	DISTINCT_SOURCES	DISTINCT_KEYWORDS	DISTINCT_OTHER_KEYWORDS			
83	1995	19	153	672	477	32	9	0	57	0		
84	1996	14	148	490	401	23	9	3	62	0		
85	1997	13	179	343	289	16	6	4	49	0		
86	1998	19	159	527	383	23	9	4	57	0		
87	1999	24	176	757	590	39	11	18	94	0		
88	2000	19	191	660	455	28	9	13	57	0		
89	2001	28	192	706	497	44	13	13	68	0		
90	2002	21	186	770	542	44	11	12	61	0		
91	2003	21	144	474	358	51	15	8	62	0		
92	2004	23	94	723	471	34	12	14	68	0		
93	2005	20	24	542	406	25	13	20	49	0		
94	2006	3	1	100	94	9	3	3	17	0		
95	2007	1	0	12	12	1	1	1	2	0		

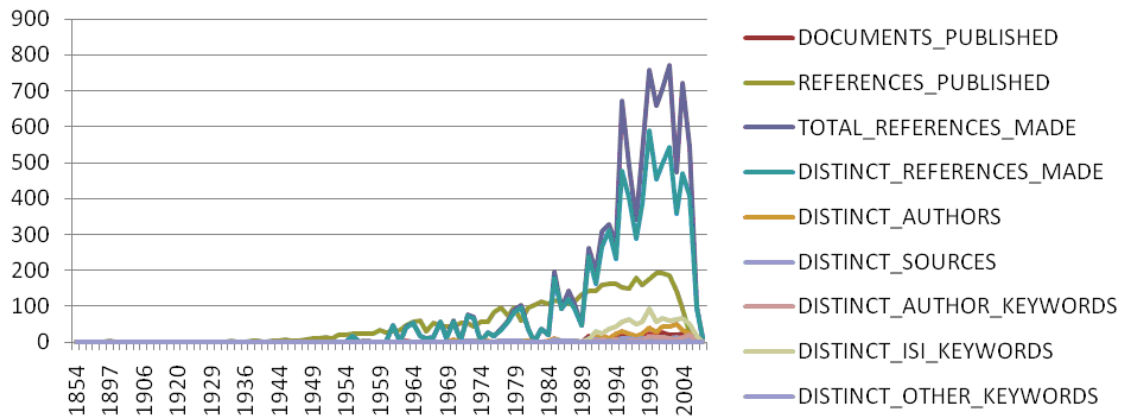
66



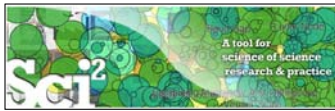
Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – Records over time

Aggregate data by year by running ‘Data Preparation > Database > ISI > Extract Longitudinal Study.’ Result is 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 amount of references made, the amount of distinct references, distinct authors, distinct sources, and distinct keywords per year.



67



Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – Burst Analysis for References

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 Authors > Extract References by Year for Burst Detection’ on the cleaned database followed by ‘Analysis > Topical > Burst Detection’ with parameters on left and then run ‘Visualize > Temporal > Horizontal Bar Graph’ with parameters on right.

Burst Detection

Perform Burst Detection on time-series textual data.

Gamma: 1.0

General Ratio: 2.0

First Ratio: 2.0

Bursting States: 1

Date Column: Year

Date Format: yyyy

Text Column: Reference

Text Separator: ||

OK Cancel

Watch those red warnings!

Horizontal Bar Graph

Takes tabular data and generates PostScript for a horizontal bar graph.

Label: Word

Start Date: Start

End Date: End

Size By: Strength

Date Format: Month-Day-Year Date Format (U.S., e.g. 10/31/2010)

Year Label Font Size: 20.0

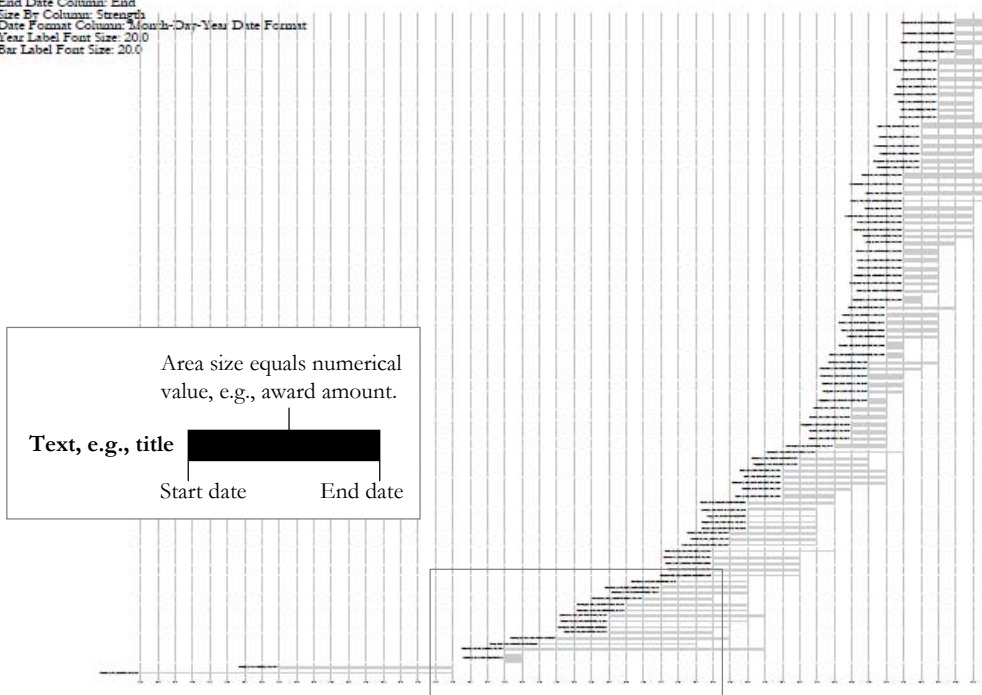
Bar Label Font Size: 20.0

OK Cancel

68

Horizontal Bar Graph for maximum burst level 1

Label Column: Word
 Start Date Column: Start
 End Date Column: End
 Size For Column: Strength
 Date Format Column: Month-Day-Year Date Format
 Year Label Font Size: 20.0
 Bar Label Font Size: 20.0

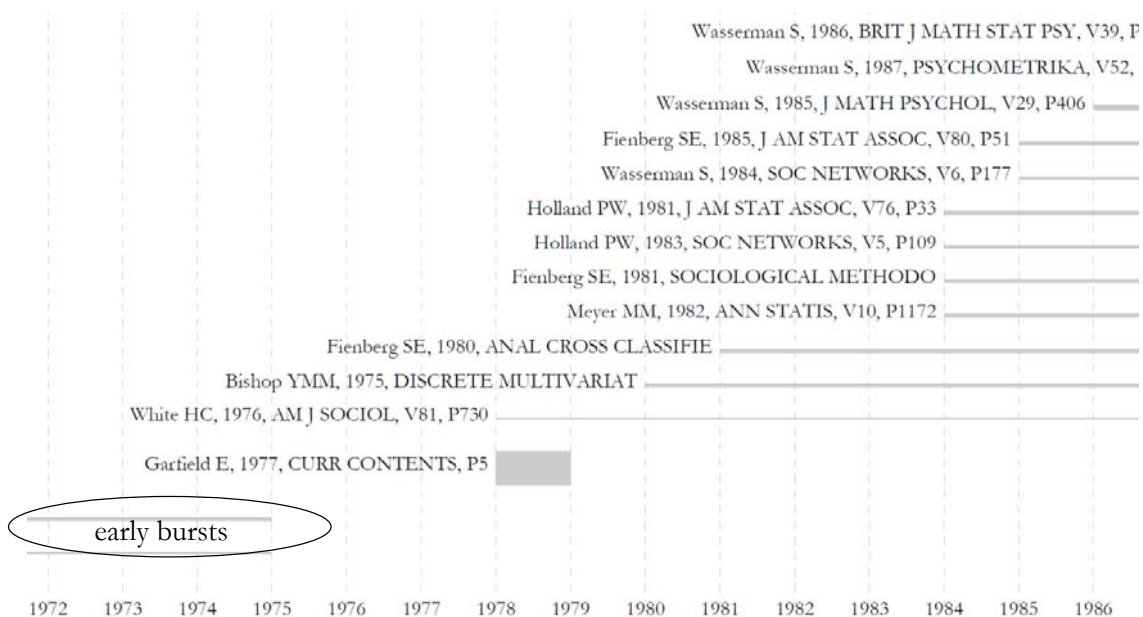


Cyberinfrastructure for Network Science Center (2009) Horizontal Bar Graph. Indiana University, <http://cisl.ohio.indiana.edu>



Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – Burst Analysis Result





Workshop Overview

9:00 Marcoscope Design and Usage & CShell Powered Tools

9:15 Sci2 Tool Basics

- Download and run the tool.

9:30 Sci2 Workflow Design: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

10:00 Sci2 Tool – Using Text and Database Support

- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

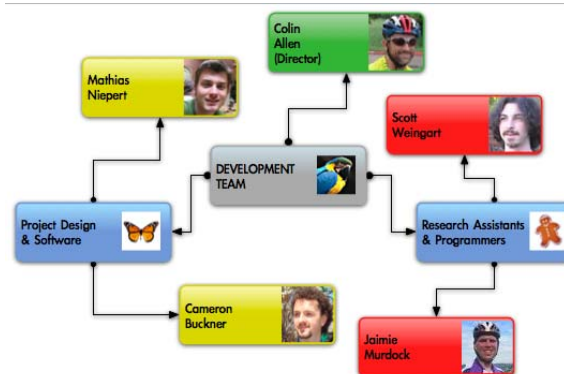
11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance

71

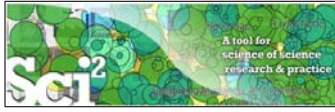


Indiana Philosophy Ontology Project – Mapping Concepts in Philosophy



The Indiana Philosophy Ontology ([InPhO](#)) project offers a variety of tools for students, researchers, programmers and scholars. The ontology currently covers philosophical [ideas](#), [thinkers](#), and [journals](#), and is powered by expert feedback and cutting-edge machine learning techniques over the 13 million word [Stanford Encyclopedia of Philosophy](#) (SEP).

72



Indiana Philosophy Ontology Project – Mapping Concepts in Philosophy

- SEPreRelatedness.nwb – Using the philosophy ontology to find related articles in the Stanford Encyclopedia of Philosophy

- Open *SEPreRelatedness.mwb* in a text editor.

```
*Nodesid*int label*string xpos*real ypos*real
1 "law-language" 0.648 0.2027
2 "constructive-empiricism" 0.7694 0.3133
3 "patrizi" 0.45 0.6043
4 "habermas" 0.1964 0.3122
5 "marcel" 0.5689 0.31996
...
*UndirectedEdgessource*int target*int weight*float
154 433 0.45
417 161 0.62
428 85 0.53
709 349 0.45
647 652 0.68
```

- NWB Network file includes nodes, edges, and attributes for each

73



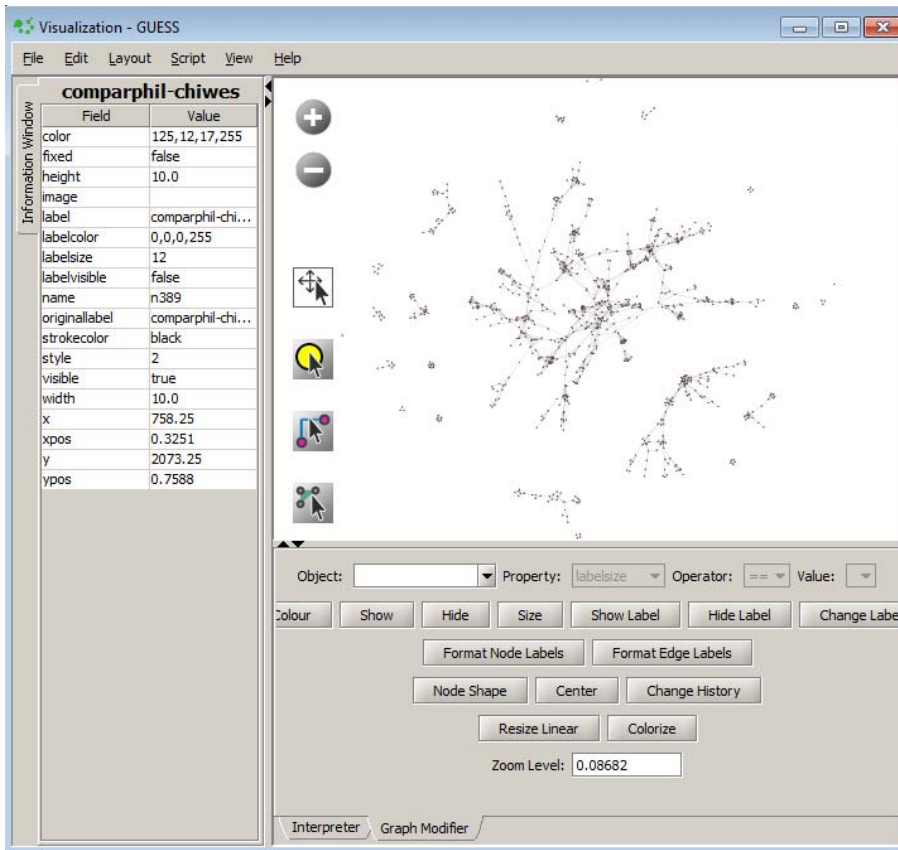
Indiana Philosophy Ontology Project – Mapping Concepts in Philosophy

- Load *SEPreRelatedness.mwb*
- Run '*Analysis > Network Analysis Toolkit (NAT)*' to get basic properties.


```
Nodes: 868
Isolated nodes: 4
Node attributes present: label, xpos, ypos
Edges: 2539
No self loops were discovered.
No parallel edges were discovered.
Edge attributes:
Did not detect any nonnumeric attributes
Numeric attributes:
minmaxmean
weight 0.430.850.53677
This network seems to be valued.
Average degree: 5.8502
...
```

- Select network and run '*Preprocessing > Networks > Delete Isolates*'
- Run '*Analysis > Networks > Weighted & Undirected > Degree & Strength*' selecting 'weight' as the Weight Attribute, followed by '*Analysis > Networks > Weighted & Undirected > Node Betweenness Centrality*' with default values.
- Run '*Visualization > GUESS*' to open GUESS with file loaded.
- Apply '*Layout > GEM*' followed by '*Layout > Bin Pack*'


74



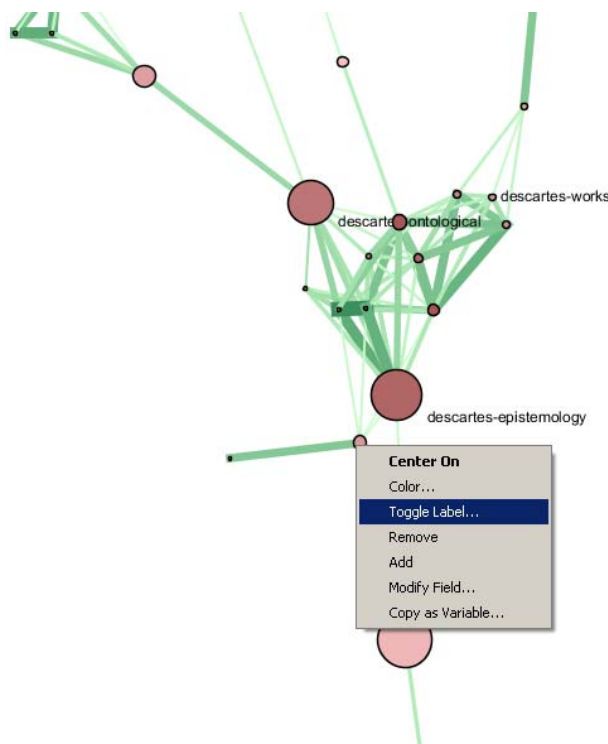
Resize Linear > Nodes > Site Betweenness > From 1 to 50 > Do Resize Linear.

Colorize > Nodes > strength >  > Do Colorize.


Resize Linear > Edges > Weight > 1 to 10 > Do Resize Linear.

Colorize > Edges > weight >  > Do Colorize.


Zoom in to view cliques & right-click to toggle labels.



Resize Linear > Nodes > Site Betweenness > From 1 to 50 > Do Resize Linear.

Colorize > Nodes > strength >  > Do Colorize.

Resize Linear > Edges > Weight > 1 to 10 > Do Resize Linear.

Colorize > Edges > weight >  > Do Colorize.

Zoom in to view cliques & right-click to toggle labels.



Workshop Overview

9:00 Marcoscope Design and Usage & CShell Powered Tools

9:15 Sci2 Tool Basics

- Download and run the tool.

9:30 Sci2 Workflow Design: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

10:00 Sci2 Tool – Using Text and Database Support

- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance

77



The Letters of Athanasius Kircher – Glen Worthey



Source: <http://kircher.stanford.edu/>

- *Kircher.csv* – 362 letters and documents from Stanford's Kircher collection
- Open *Kircher.csv*

```
ID|TITLE,NAMES,OCCUPATIONS,DAY,MONTH,DATE,CITY,LANGUAGE,Subs
```

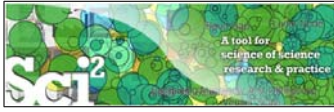
```
2487,Brahe @ Copernicus @ Appenzeller @ Bidermann @ Adriansens @
```

```
Clavius, astronomer, 27, June, 1600, Ingolstadt, Latin, calendar @ sundials @ astronomy
```

```
2541, Gellibrand, mathematician, 09, March, 1635, no place given, Latin, longitude @ magnetic declination @ magnetism
```

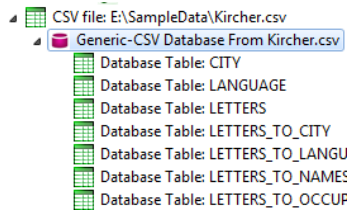
```
...
```

78

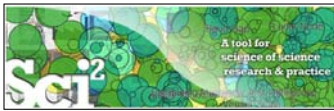


The Letters of Athanasius Kircher – Glen Worthey

- Load *Kircher.csv* in ‘Standard csv format’.
- Run ‘File > Load Generic CSV-File Into Database’ with the following parameters:
 - Core entity = LETTERS
 - IDTITLE=Integer
 - NAMES = String, Create a separate leaf table, merge if identical, contains multi-valued fields, type ‘@’ as the separator
 - OCCUPATIONS = String, Create a separate leaf table, merge if identical, contains multi-valued fields, type ‘@’ as the separator
 - DAY = Integer
 - MONTH = String
 - DATE = Integer
 - CITY = String, Create a separate leaf table, merge if identical
 - LANGUAGE = String, Create a separate leaf table, merge if identical
 - Subs = String, Create a separate leaf table, merge if identical, contains multi-valued fields, type ‘@’ as the separator
- These attributes can also be loaded from columnAttributes.gcl
- ‘I’m Finished!’
- Run ‘Data Preparation > Database > General > Extract Raw Tables From Database’ to view various database tables.



79



The Letters of Athanasius Kircher – Glen Worthey

- Run ‘Data Preparation > Database > Generic CSV > Extract Co-Occurrence Network’ with the following parameters:

Choose the Leaf column to extract the co-occurrence network on: NAME!

Node Attributes

Add Another Node Attribute Field

Remove All Node Fields

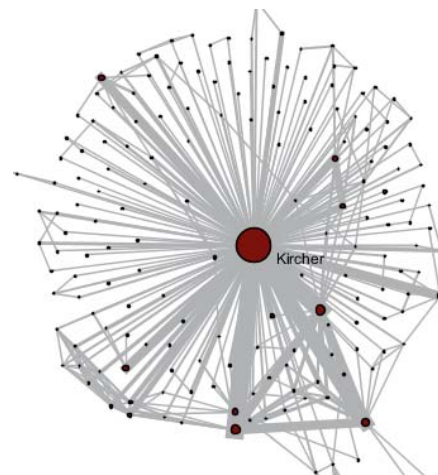
Summarize By: Count
Column to Summarize: IDTITLE
Attribute Name: NumLetters
Delete

Edge Attributes

Add Another Edge Attribute Field

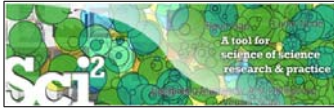
Remove All Edge Fields

Summarize By: Count
Column to Summarize: IDTITLE
Attribute Name: NumLetters
Delete



- Run ‘Visualization > GUESS’ to open GUESS with file loaded.
- Apply ‘Layout > GEM’ followed by ‘Layout > Bin Pack’
- Results not surprising for an egocentric network.

80



The Letters of Athanasius Kircher – Glen Worthey

- Run ‘Data Preparation > Database > Generic CSV > Extract Co-Occurrence Network’ with the following parameters:

Choose the Leaf column to extract the co-occurrence network on: Subs

Node Attributes

Add Another Node Attribute Field

Remove All Node Fields

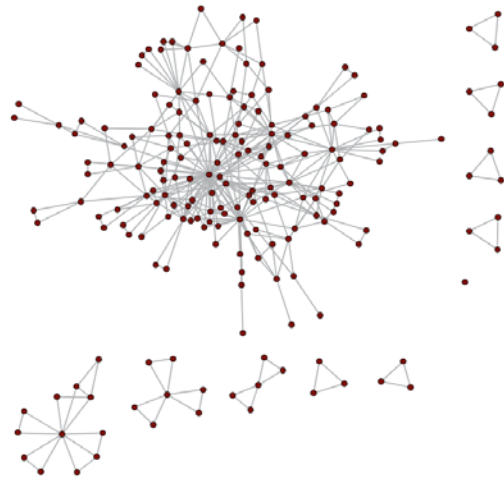
Summarize By: Count
 Column to Summarize: IDTITLE
 Attribute Name: COUNT_IDTITI [Delete]

Edge Attributes

Add Another Edge Attribute Field

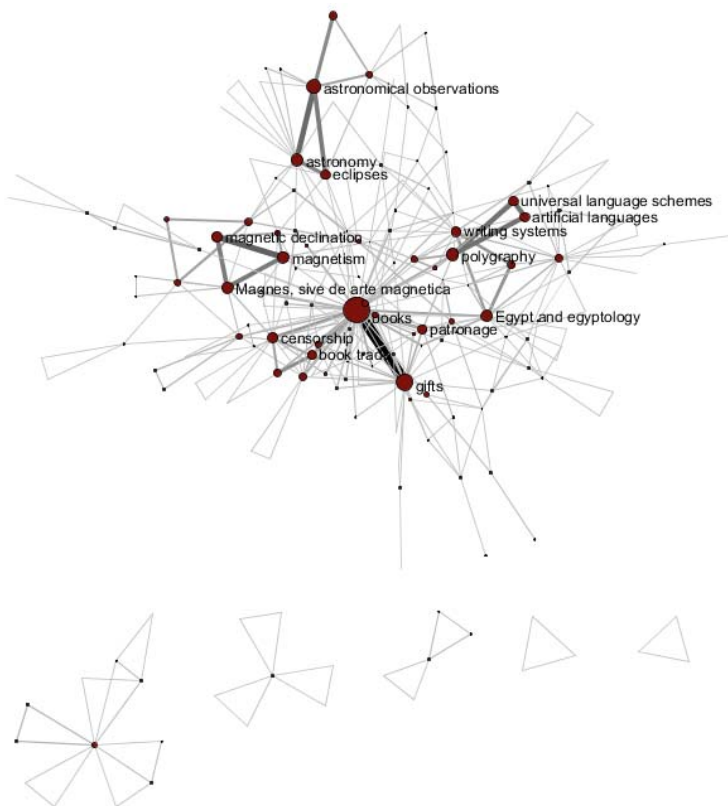
Remove All Edge Fields

Summarize By: Count
 Column to Summarize: IDTITLE
 Attribute Name: COUNT_IDTITI [Delete]



- Run ‘Visualization > GUESS’ to open GUESS with file loaded.
- Apply ‘Layout > GEM’ followed by ‘Layout > Bin Pack’

81



Resize Linear > Edges > edge_count_idtitle > From 1 to 10 > Do Resize Linear.

Colorize > Edges > edge_count_idtitle > From: [light gray] To: [dark gray] > Do Colorize.

Resize Linear > Nodes > node_count_idtitle > 0 to 25 > Do Resize Linear.

Object: ‘Nodes Based On’
 Property: ‘node_count_idtitle’
 ID: ‘>=’
 Value: ‘15’
 Show Label

82



- Run ‘*Data Preparation > Database > Generic CSV > Extract Bi-Partite Network*’ with the following parameters:

Choose the Source for your bipartite network extraction: LANGUAGE

Choose the Target for your bipartite network extraction: OCCUPATIONS

Node Attributes

Add Another Node Attribute Field

Remove All Node Fields

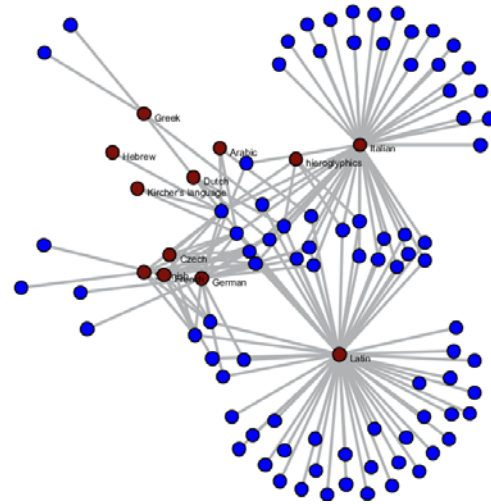
Summarize By: Count
Column to Summarize: IDTITLE
Attribute Name: COUNT_IDTITI [Delete]

Edge Attributes

Add Another Edge Attribute Field

Remove All Edge Fields

Summarize By: Count
Column to Summarize: IDTITLE
Attribute Name: COUNT_IDTITI [Delete]



- Run ‘*Visualization > GUESS*’ to open GUESS with file loaded.
- Apply ‘*Layout > GEM*’ followed by ‘*Layout > Bin Pack*’
- Show labels and color nodes based on entity_type Language

83



Workshop Overview

9:00 Marcoscope Design and Usage & CShell Powered Tools

9:15 Sci2 Tool Basics

- Download and run the tool.
- 9:30 Sci2 Workflow Design: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

10:00 Sci2 Tool – Using Text and Database Support

- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance

84

Extraneous Slides

- Adding Plugins to CShell Powered Tools
- OSGi/CShell Adoption

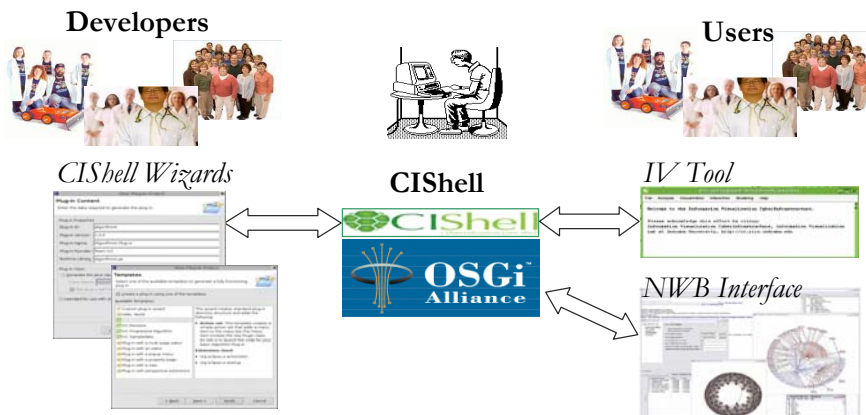
85



Cyberinfrastructure Shell (CIShell)

<http://cishell.org>

- CIShell is an open source software specification for the integration and utilization of datasets, algorithms, and tools.
- It extends the Open Services Gateway Initiative (OSGi) (<http://www.osgi.org>), a standardized, component oriented, computing environment for networked services widely used in industry since 10 years.
- Specifically, CIShell provides “sockets” into which existing and new datasets, algorithms, and tools can be plugged using a wizard-driven process.



86

CIShell is built upon the Open Services Gateway Initiative (OSGi) Framework.

OSGi (<http://www.osgi.org>) is

- A standardized, component oriented, computing environment for networked services.
- Successfully used in the industry from high-end servers to embedded mobile devices since 8 years.
- Alliance members include IBM (Eclipse), Sun, Intel, Oracle, Motorola, NEC and many others.
- Widely adopted in open source realm, especially since Eclipse 3.0 that uses OSGi R4 for its plugin model.

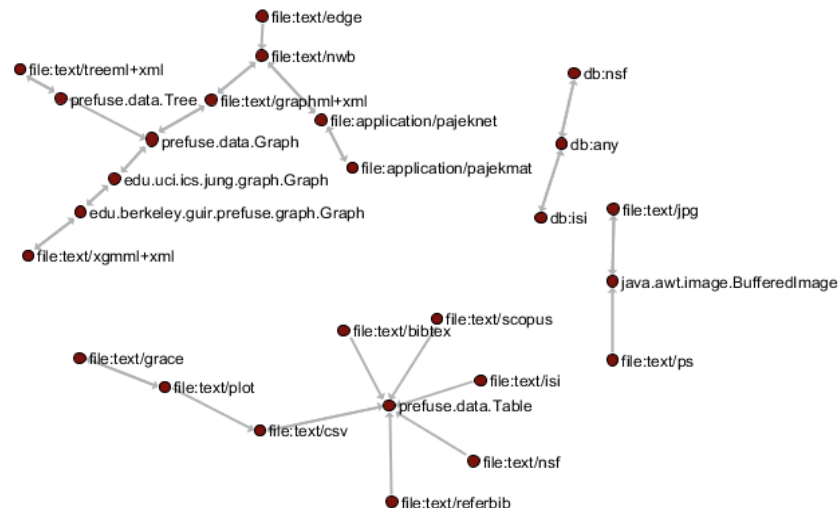
Advantages of Using OSGi

- Any CIShell algorithm is a service that can be used in any OSGi-framework based system.
- Using OSGi, running CIShells/tools can be connected via RPC/RMI supporting peer-to-peer sharing of data, algorithms, and computing power.

Ideally, CIShell becomes a standard for creating OSGi Services for algorithms.

87

- No central data format.
- Sci² Tool has 26 external and internal data formats and 35 converters.
- Their relationships can be derived by running ‘File > Converter Graph’ and plotted as shown here. Note that some conversions are symmetrical (double arrow) while others are one-directional (arrow).



88

- Not all code can be shared freely (yet).
- To make the UCSD Science Map and new geomaps available via the Sci² menu, simply add

Name	Size	Type	Date Modified
edu.iu.scipolicy.visualization.geomaps_0.0.1.jar	4,864 KB	Executable Jar File	6/24/2010 5:41 PM
edu.iu.scipolicy.visualization.scimaps_0.0.1.jar	1,507 KB	Executable Jar File	6/18/2010 3:17 PM
org.cishell.reference.gui.persistence_1.0.0.jar	61 KB	Executable Jar File	6/24/2010 5:41 PM
org.cishell.utilities_1.0.0.jar	72 KB	Executable Jar File	6/24/2010 5:41 PM

to the *'yourdirectory/plugin'* directory and restart the tool.

The rights to the UCSD map are owned by the Regents of UCSD. Usage does not require a separate, signed agreement or an additional request to our office if consistent with the permission. As a courtesy, please send information on how the map is being used to

William J. Decker, Ph.D., Associate Director, Technology Transfer Office
 University of California, San Diego, 9500 Gilman Drive Dept. 0910, La Jolla, CA 92093
 phone:858-822-5128, fax: 858-534-7345, e-mail: wjdecker@ucsd.edu

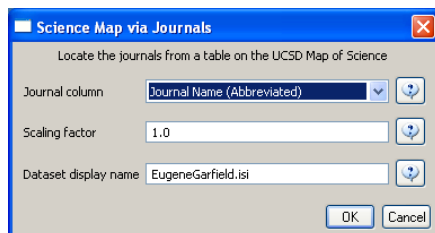
- To delete algorithms that you do not use, simply delete the corresponding *.jar files in the plugin directory.
- Customize your menu structure accordingly—see next slide.

89

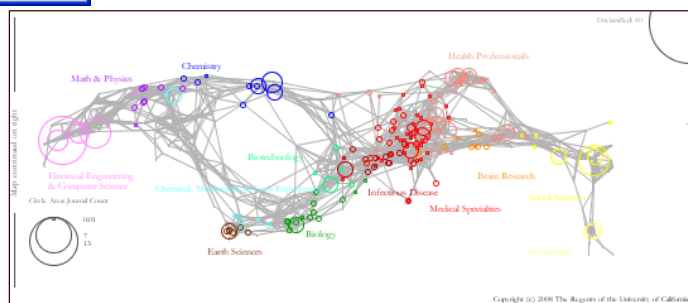
- After you added the new plugins, load an ISI file using 'File > Load and Clean ISI File > EugeneGarfield.isi.'

The file can be found in the */sampledata/scientometrics/isi* directory.

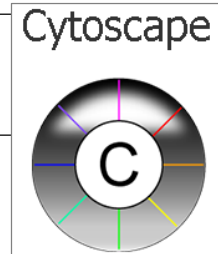
- Select '99 Unique ISI Records' file in Data Manger and run 'Visualization > Topical > Science Map via Journals' with parameters:



- The result is a science map overlay of Garfield's papers and a listing of journals in 13 fields of science below.



90

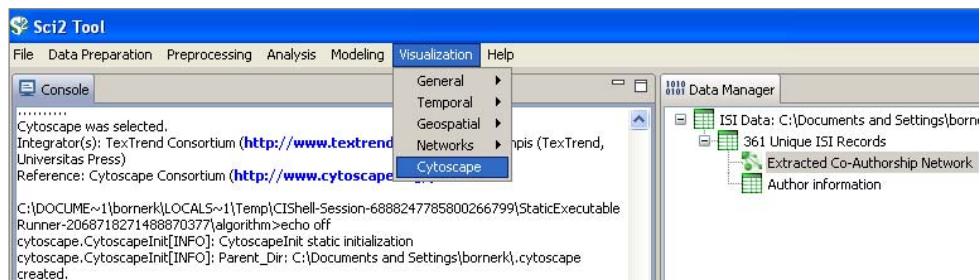


Adding more layout algorithms and network visualization interactivity via Cytoscape <http://www.cytoscape.org>.

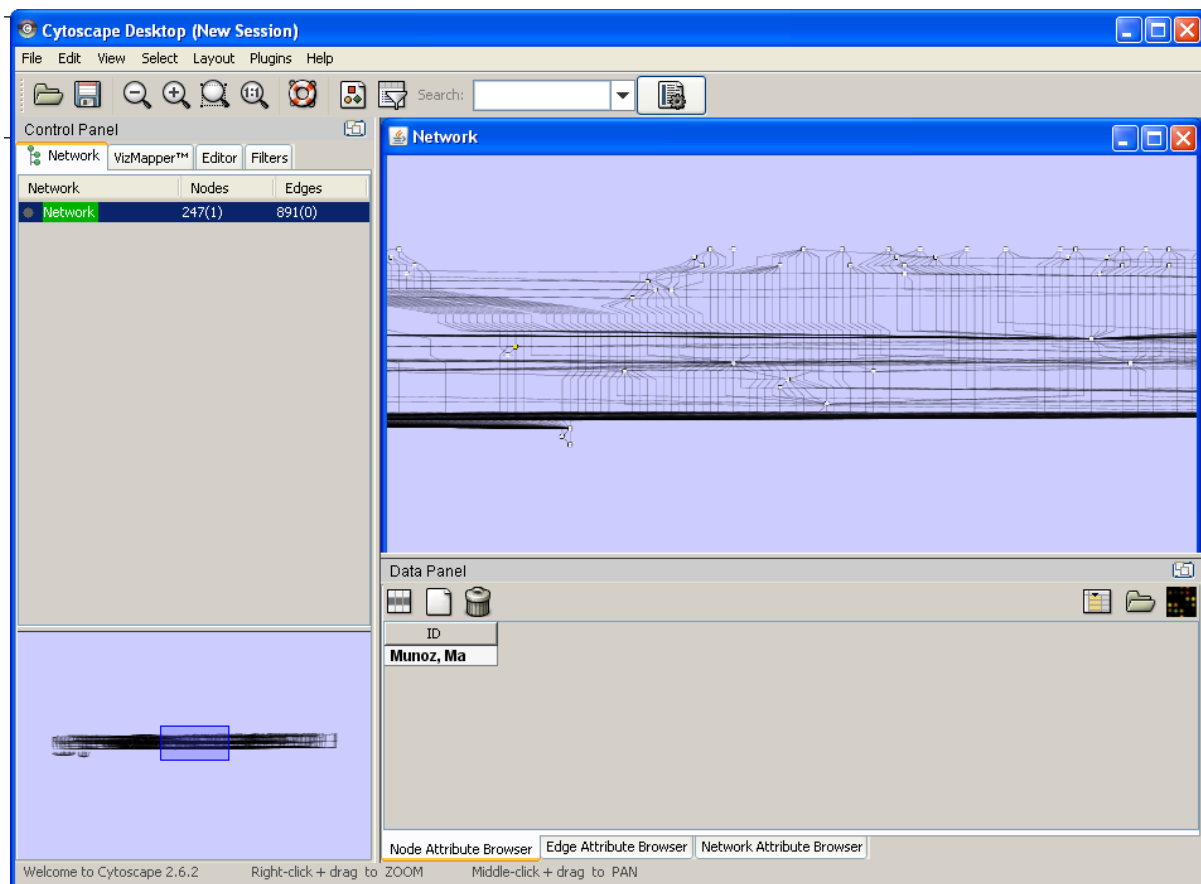
Simply add *org.textrend.visualization.cytoscape_0.0.3.jar* into your /plugin directory.

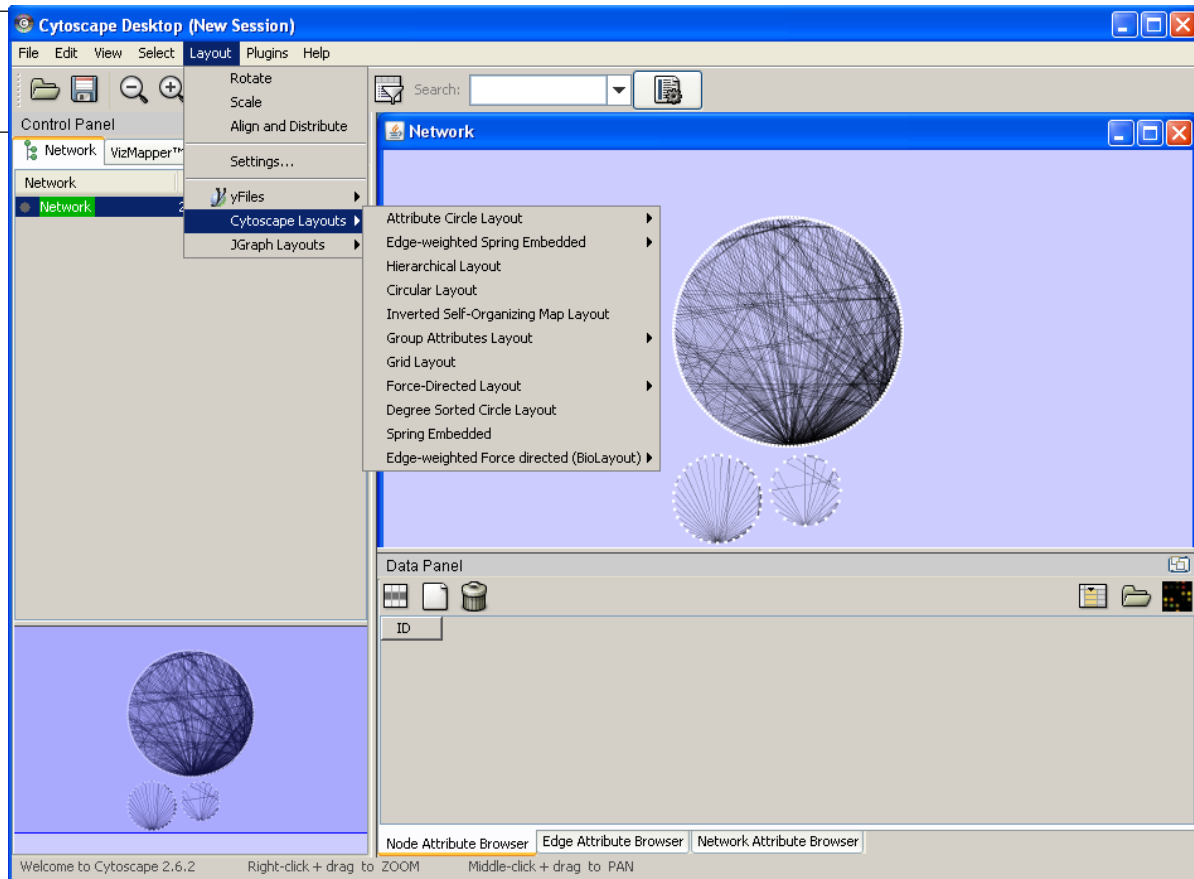
Restart Sci2 Tool.

Cytoscape now shows in the Visualization Menu.



Select a network in Data Manager, run Cytoscape and the tool will start with this network loaded.





CIShell – Integrate New Algorithms

Algorithm Developer's Guide

Overview

The Cyberinfrastructure Shell (CIShell) is an open source, community-driven platform for the integration and utilization of datasets, algorithms, tools, and computing resources. Algorithm integration support is built in for Java and most other programming languages. Being Java based, it will run on almost all platforms. The software and specification is released under an [Apache 2.0 License](#).

This guide attempts to aid algorithm developers in creating algorithms for CIShell (and applications built on CIShell).

This guide tries to contain all the information a new developer needs, but where necessary, it may cite the [CIShell 1.0 Specification \(API\)](#) or the [OSGi Service Platform Specification, Release 4 \(API\)](#). While the guide tries to make beginning algorithm development easier, the CIShell Specification has the last word on how the CIShell Platform works.

Table of Contents

1. [CIShell Basics](#)
2. Getting Started
 1. [Tutorial 0: Setting Up the Development Environment](#)
 2. [Tutorial 1: Creating a Hello World Java Algorithm](#)
 3. [Tutorial 2: Practical Java Algorithm Development](#)
 4. [Tutorial 3: Integrating a Non-Java Program As An Algorithm](#)
 5. [Mini-Tutorial: Integrating 3rd-party libraries](#)
 6. [Where to Learn More](#)
3. Reference
 1. [How Algorithms Work: A guide to algorithm plugins in CIShell](#)
 2. [Accessing the OSGi Console in CIShell tools](#)

<http://cishell.org/?n=DevGuide.NewGuide>

- The file `yourtooldirectory/configuration/default_menu.xml` encodes the structure of the menu system.
- In NWB Tool, the Modeling menu (left) is encoded by the following piece of xml code:

The screenshot shows the Network Workbench Tool interface with the 'Modeling' menu selected. The menu items listed are: Random Graph, Watts-Strogatz Small World, Barabási-Albert Scale-Free, Can, Chord, Hypergrid, PRU, TARL, Discrete Network Dynamics (DND), and Evolving Network (Weighted). Below the screenshot, the XML code for the 'Modeling' menu is displayed:

```
<top_menu name="Modeling">
<menu pid="edu.iu.nwb.modeling.erdosrandomgraph"/>
<menu pid="edu.iu.nwb.modeling.smallworld"/>
<menu pid="edu.iu.nwb.modeling.barabasi-albert"/>
<menu type="break"/>
<menu pid="edu.iu.iv.modeling.p2p.can.CanAlgorithm"/>
<menu pid="edu.iu.iv.modeling.p2p.chord.ChordAlgorithm"/>
<menu pid="edu.id.iv.modeling.p2p.hypergrid.Hypergrid"/>
<menu pid="edu.iu.iv.modeling.p2p.pru.PruAlgorithm"/>
<menu type="break"/>
<menu pid="edu.iu.iv.modeling.tarl.TarlAlgorithm"/>
<menu type="break"/>
<menu pid="edu.iu.nwb.modeling.discretenetworkdynamics.DNDAlgorithm"/>
<menu type="break"/>
<menu pid="edu.iu.nwb.modeling.weighted.evolvingnetwork"/>
</top_menu>
```



OSGi/CIShell Adoption

CIShell/OSGi is at the core of different CIs and a total of 169 unique plugins are used in the

- **Information Visualization** (<http://iv.slis.indiana.edu>),
- **Network Science (NWB Tool)** (<http://nwb.slis.indiana.edu>),
- **Scientometrics and Science Policy (Sc² Tool)** (<http://sci.slis.indiana.edu>), and
- **Epidemics** (<http://epic.slis.indiana.edu>) research communities.

Most interestingly, a number of other projects recently adopted OSGi and one adopted CIShell:

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 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 et al., 2006). Taverna allows users to integrate many different software tools, including over 30,000 web services.

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.

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.

As the functionality of OSGi-based software frameworks improves and the number and diversity of dataset and algorithm plugins increases, the capabilities of custom tools will expand.

TEXTrend adds R bridge, WEKA, Wordij, CFinder, and more.

See the latest versions of TEXTrend Toolkit modules at

http://textrend.org/index.php?option=com_content&view=article&id=47&Itemid=53



Epidemics Marketplace and EpiC Tool

<http://dev.epic.slis.indiana.edu>

