

Sci2: A Tool of Science of Science Research and Practice

Dr. Katy Börner

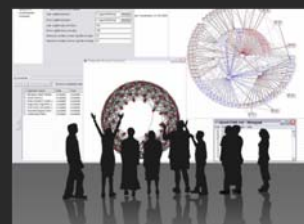
Cyberinfrastructure for Network Science Center
Information Visualization Laboratory
School of Library and Information Science
Indiana University, Bloomington, IN
<http://cns.iu.edu>



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

*13th International Society for Scientometrics and Informetrics (ISSI) Conference
Durban, South Africa*

Thursday, July 4, 2011 • 1.30pm – 5.00pm



Online Resources

- These slides
<http://sci2.cns.iu.edu/docs/2011-borner-ISSI-Tutorial.pdf>
- Sci2 Tool Manual v0.5.1 Alpha
<http://sci2.wiki.cns.iu.edu>
- Sci2 Tool v0.5.1 Alpha (May 4, 2011)
<http://sci2.cns.iu.edu>
- Additional Datasets
<http://sci2.wiki.cns.iu.edu/2.5+Sample+Datasets>
- Additional Plugins
<http://sci2.wiki.cns.iu.edu/3.2+Additional+Plugins>



Or copy them from the DVD or memory stick.



Workshop Overview

1:30 Macroscope Design and Usage & CISHell Powered Tools: NWB & Sci2

1:45 Sci2 Tool Basics

- Download and run the tool.

2:00 Sci2 Sample Workflow: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

2:30 Sci2 Sample Workflow: Studying Four Major NetSci Researchers.

- Load and clean a dataset as text file; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize as either a circular hierarchy or network

3:30 Break

4:00 Sci2 Demo I: Geospatial maps with congressional districts

4:30 Sci2 Demo II: Evolving collaboration networks

4:45 Outlook and Discussion

5:00 Adjourn



Workshop Overview

1:30 Macroscope Design and Usage & CISHell Powered Tools: NWB & Sci2

1:45 Sci2 Tool Basics

- Download and run the tool.

2:00 Sci2 Sample Workflow: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

2:30 Sci2 Sample Workflow: Studying Four Major NetSci Researchers.

- Load and clean a dataset as text file; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize as either a circular hierarchy or network

3:30 Break

4:00 Sci2 Demo I: Geospatial maps with congressional districts

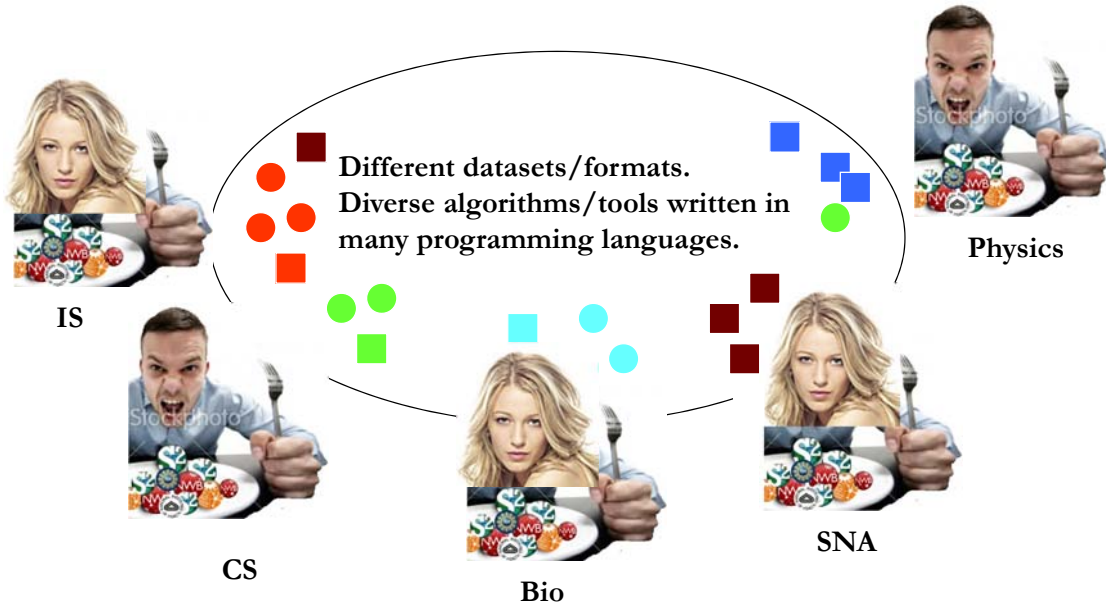
4:30 Sci2 Demo II: Evolving collaboration networks

4:45 Outlook and Discussion

5:00 Adjourn



Macrosopes Serve the Changing Scientific Landscape



5



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.

6



Macroscopic Design



Custom Tools for Different Scientific Communities

Information Visualization Cyberinfrastructure

<http://iv.cns.iu.edu>

Network Workbench Tool + Community Wiki

<http://nwb.cns.iu.edu>

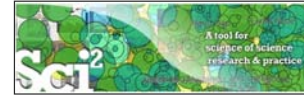
Science of Science (Sci²) Tool and Portal

<http://sci2.cns.iu.edu>



Epidemics Cyberinfrastructure

Coming soon



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



Network Workbench Tool

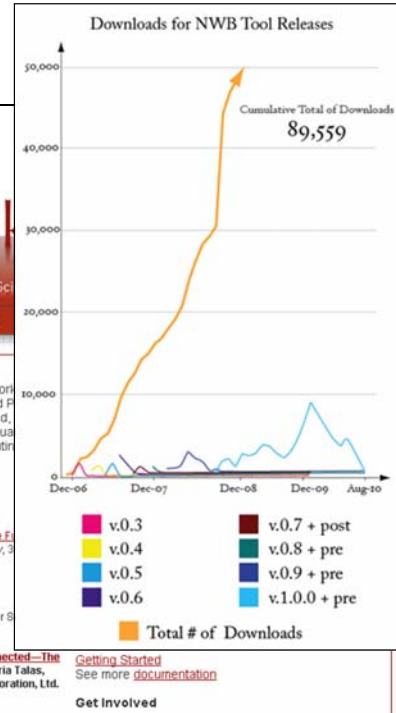
<http://nwb.slis.indiana.edu>

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.



Herr II, Bruce W., Huang, Weixia (Bonnie), Penumathy, Shashikant & Börner, Katy. (2007). Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. In Bainbridge, William S. & Roco, Mihail C. (Eds.), *Progress in Convergence - Technologies for Human Wellbeing* (Vol. 1093, pp. 161-179), *Annals of the New York Academy of Sciences*, Boston, MA.

9



Project Details

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>

10

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 Economics

Does the type of product that a country exports matter for subsequent economic performance?

C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann (2007) The Product Space Conditions the Development of Nations. Science 317, 482 (2007).

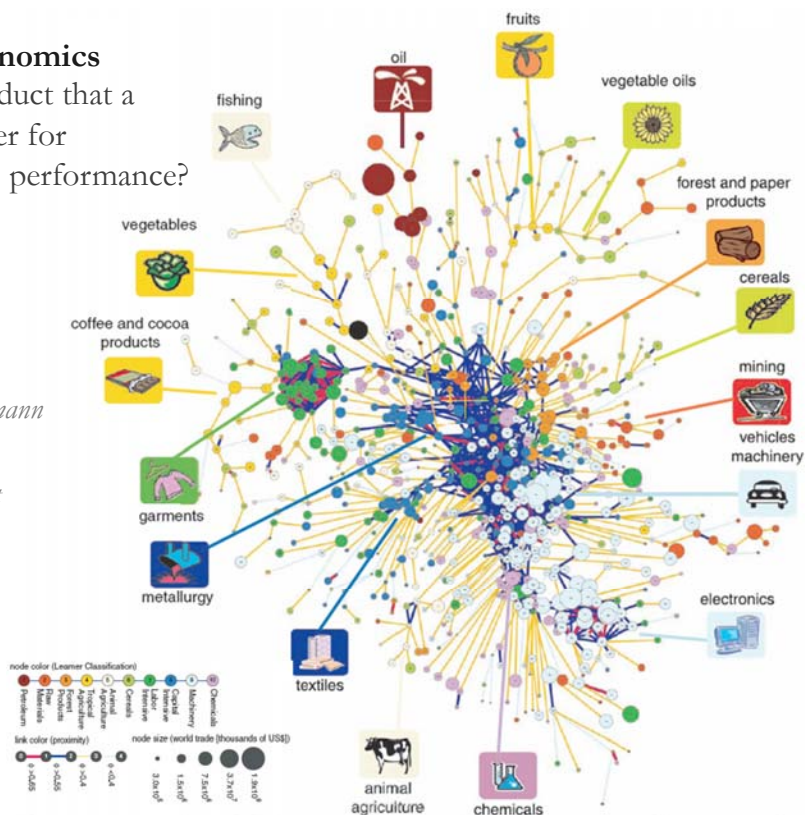


Fig. 1. The product space. (A) Hierarchically clustered proximity (a) matrix representing the 775 SITC-4 product classes exported in the 1998–2000 period. (B) Network representation of the product space. Links are color coded with their proximity value. The sizes of the nodes are proportional to world trade, and their colors are chosen according to the classification introduced by Leamer.

Computational Social Science

Studying large scale social networks such as Wikipedia

Second Sight: An Emergent Mosaic of Wikipedian Activity, The NewScientist, May 19, 2007



Second sight

Image: Bruce W. Hest and Todd M. Holloway

Power struggle

How do you keep track of the bubbling mass of information that is Wikipedia? This chaotic-looking mosaic is one attempt to show which topics are contained in the online encyclopedia.



...since the most likely topics pages at the time of writing include entries on Sheffield Wednesday football club, Mikhail Gorbachev and pigs). The mosaic has been commended in a competition for images that visualise network dynamics, coinciding with this week's International Workshop and Conference on Network Science in Bloomington.

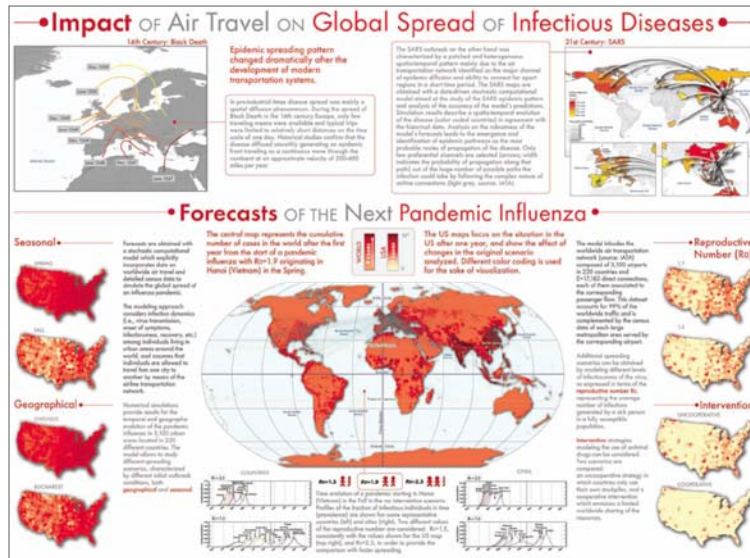
Computational Epidemics

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.cns.iu.edu>

Select your operating system from the pull down menu and download.

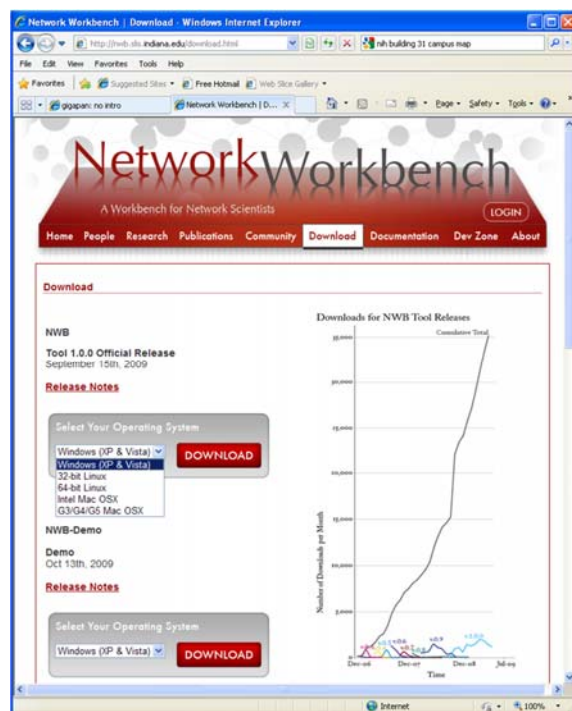
Unpack into a /nwb directory.

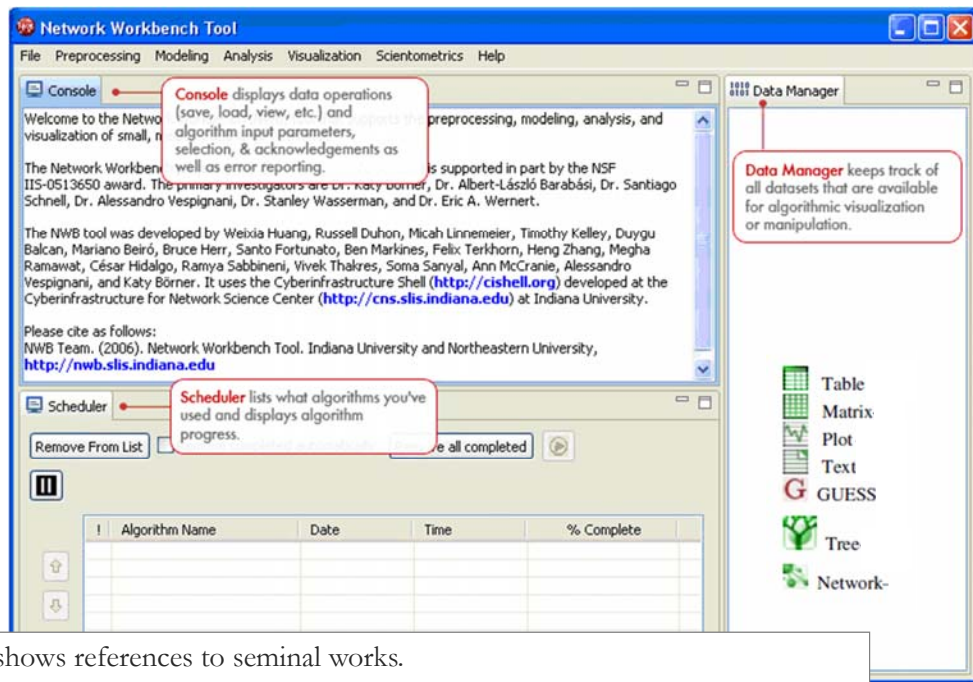
Run /nwb/nwb.exe

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

Cite as

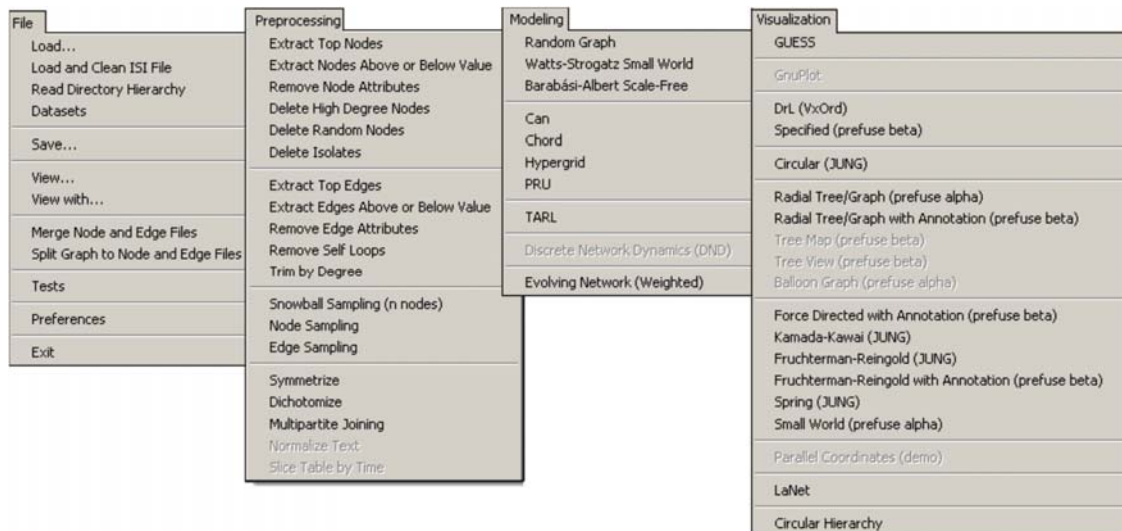
NWB Team. (2006). *Network Workbench Tool*. Indiana University, Northeastern University, and University of Michigan, <http://nwb.cns.iu.edu>.





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.

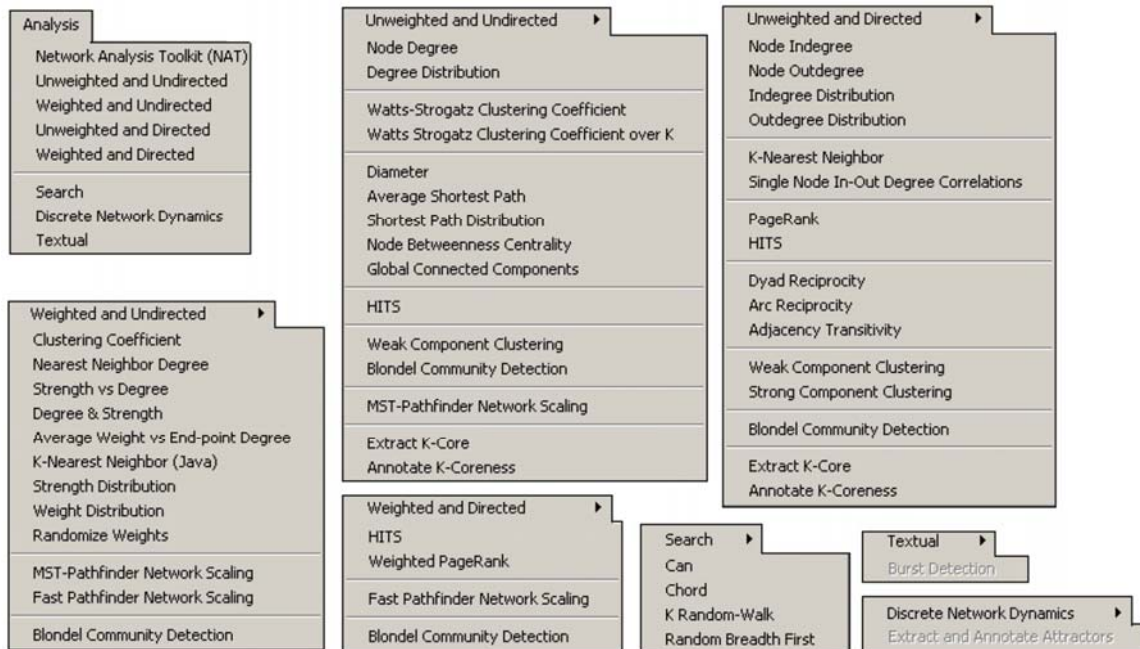
17



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



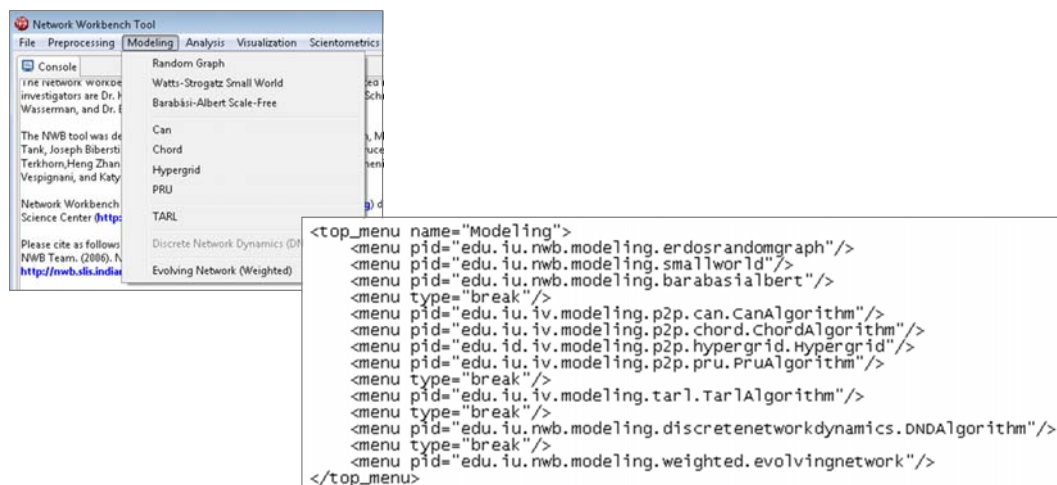
Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (2007). **Network Science**. In Blaise Cronin (Ed.), *ARIST*, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, Volume 41, Chapter 12, pp. 537-607. <http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf>

19

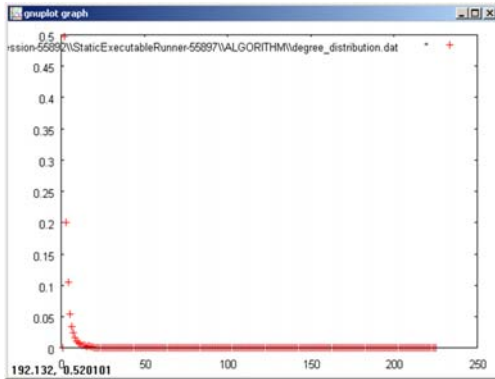


CIShell – Customize Menu

- 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:

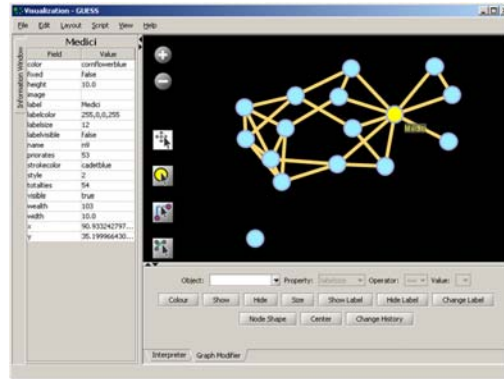


20



Gnuplot

portable command-line driven
interactive data and function plotting
utility <http://www.gnuplot.info/>.



GUESS

exploratory data analysis and visualization tool
for graphs and networks.

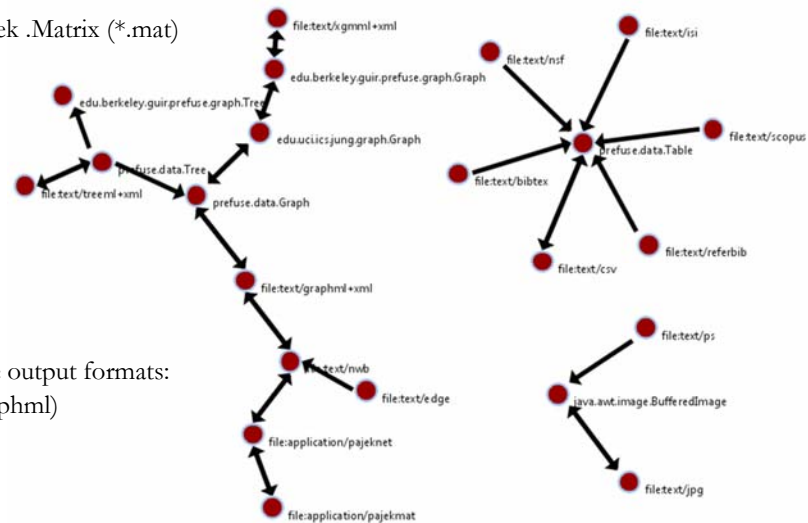
<https://nwb.slis.indiana.edu/community/?n=VisualizeData.GUESS>.

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>.

CIShell Powered Tools: Science of Science (Sci2) Tool

23



Science of Science (Sci2) Tool

<http://sci2.cns.iu.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

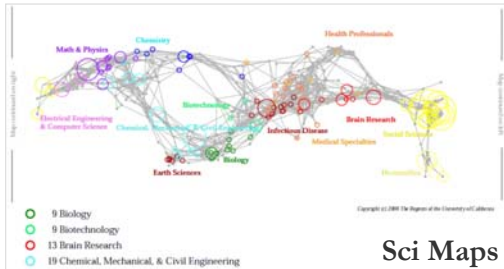
To capture the essence of good science, stakeholders must combine forces to create an open, sound and consistent system for measuring all the activities that make up academic productivity, says **Julia Lane**.

24

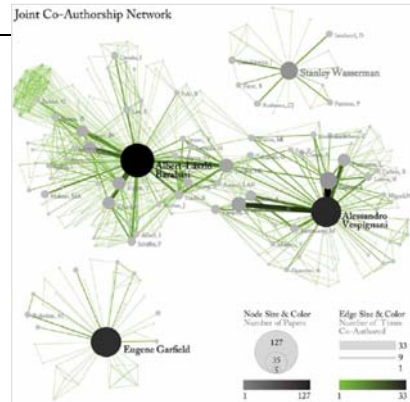


Sci² Tool – “Open Code for S&T Assessment”

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). *Reti-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.*

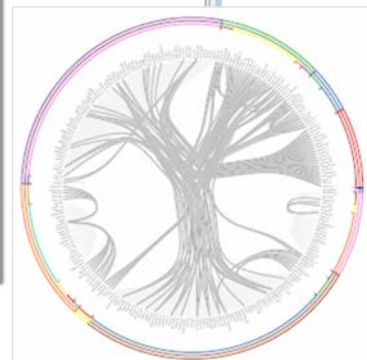
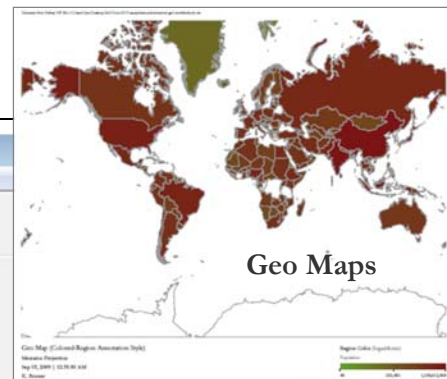


Sci² Tool

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	% Con
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%





Workshop Overview

1:30 Macroscopic Design and Usage & CShell Powered Tools: NWB & Sci2

1:45 Sci2 Tool Basics

➤ **Download and run the tool.**

2:00 Sci2 Sample Workflow: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

2:30 Sci2 Sample Workflow: Studying Four Major NetSci Researchers.

- Load and clean a dataset as text file; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize as either a circular hierarchy or network

3:30 *Break*

4:00 Sci2 Demo I: Geospatial maps with congressional districts

4:30 Sci2 Demo II: Evolving collaboration networks

4:45 Outlook and Discussion

5:00 *Adjourn*

27



Sci² Tool: Download, Install, and Run

Sci2 Tool v0.5 Alpha (April 4, 2011)

Can be freely downloaded for all major operating systems from

<http://sci2.cns.iu.edu>

Select your operating system from the pull down menu and download.

Unpack into a /sci2 directory.

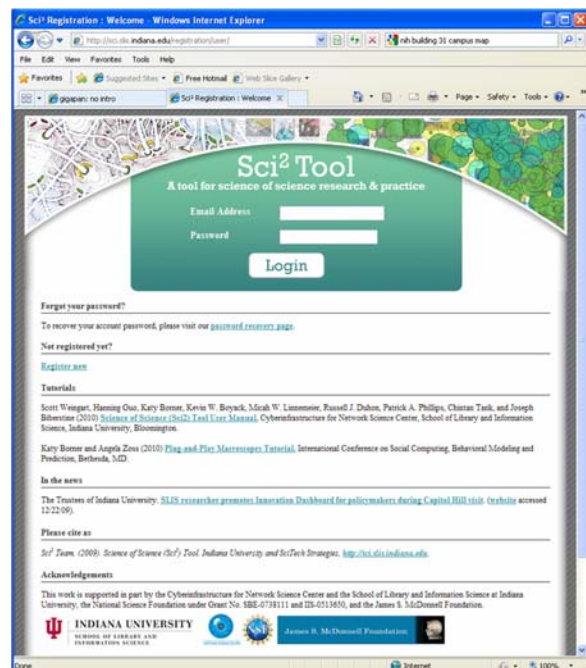
Run /sci2/sci2.exe

Sci2 Manual is at

<http://sci2.wiki.cns.iu.edu>

Cite as

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



28



Sci² Tool: Download, Install, and Run

Sci2 Tool v0.5 Alpha (April 4, 2011)

- Supports ASCII UTF-8 characters
- Web-based Yahoo! and desktop Geocoders
- U.S. and World geomapper
- Customizable stop word lists
- Merging of networks
- New home page, wiki-based tutorial
- Bug fixes, streamlined workflows



Sci2 Tool runs on Windows, Mac, and Linux.

Unzip.

Run /sci2/sci2.exe



29

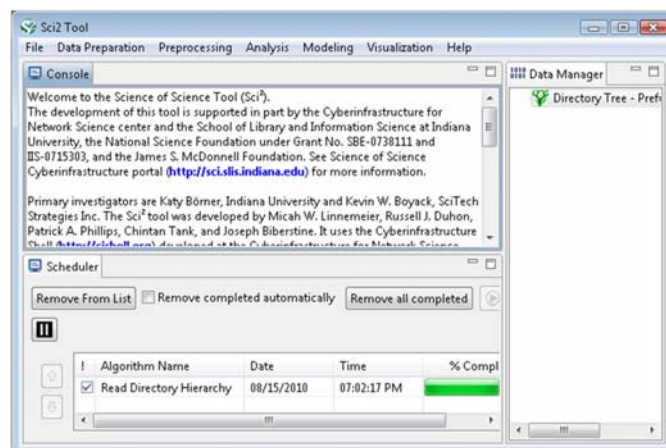


Sci² Tool Interface Components

See also <http://sci2.wiki.cns.iu.edu/2.2+User+Interface>

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 a error log to ease bug reporting.

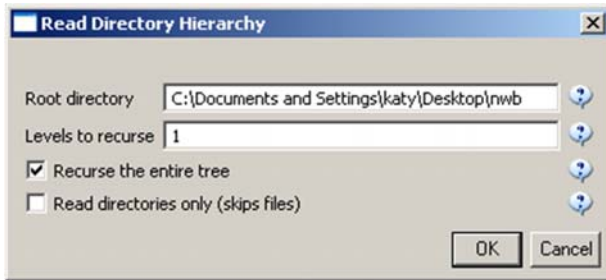
All algorithms are documented online; workflows are given in tutorials, see Sci2 Manual at <http://sci2.wiki.cns.iu.edu>

30



Sci2 Tool – Read+Visualize Sci2 Tool Directory Tree

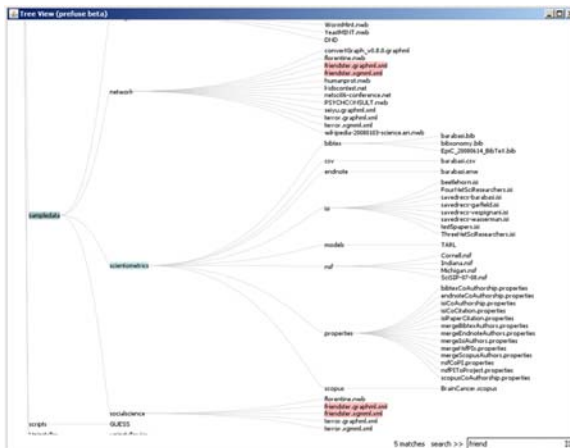
Use 'File > Read Directory Hierarchy' with parameters



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

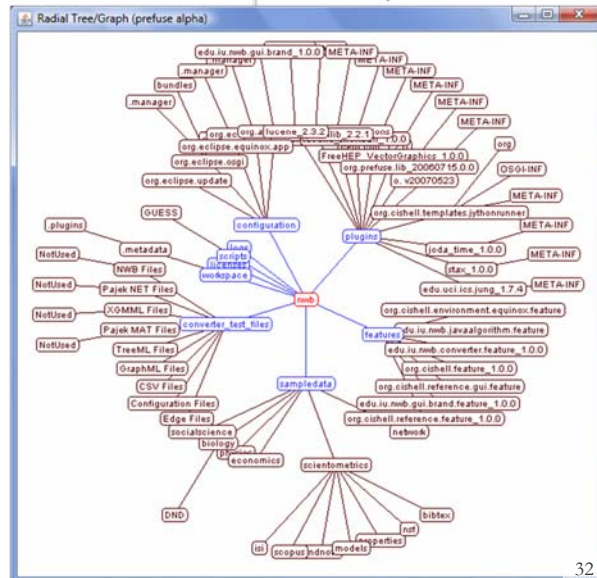
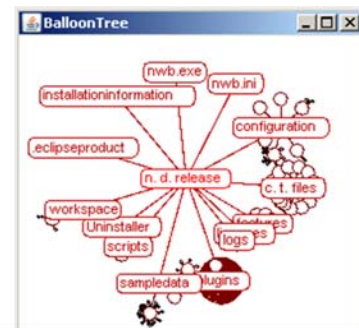
- 'Visualization > Tree View (prefuse beta)'
- 'Visualization > Tree Map (prefuse beta)'
- 'Visualization > Balloon Graph (prefuse alpha)'
- 'Visualization > Radial Tree/Graph (prefuse alpha)'

31



Different views of the /nwb directory hierarchy.

Note the size of the /plugin directory.



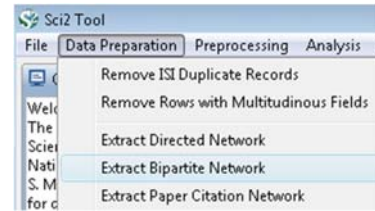
32



Sci2 Tool – Visualize Workshop Attendees

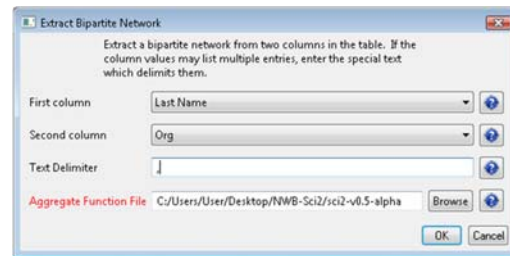
Use *File > Read* to load *SciTS Conf SNA Registrants report 4.10.11-clean.csv*

	A	B	C
1	Last Name	Org	Organization-Cleaned
2	Agoulnik	edu	Brigham and Women's Hospital, Harvard Medical School
3	Amaral	edu	Northwestern University
4	Bates	edu	University of Illinois at Chicago
5	Bennett	gov	NIH
6	Bietz	edu	University of California, Irvine
7	Bishop	edu	University of Tennessee
8	...		
9	Lotrecchiano	org.edu	George Washington University
10	Lusina	ca_edu	Centre for Hip Health & Mobility



Run *Data Preparation > Extract Bipartite Network*

With parameter values:



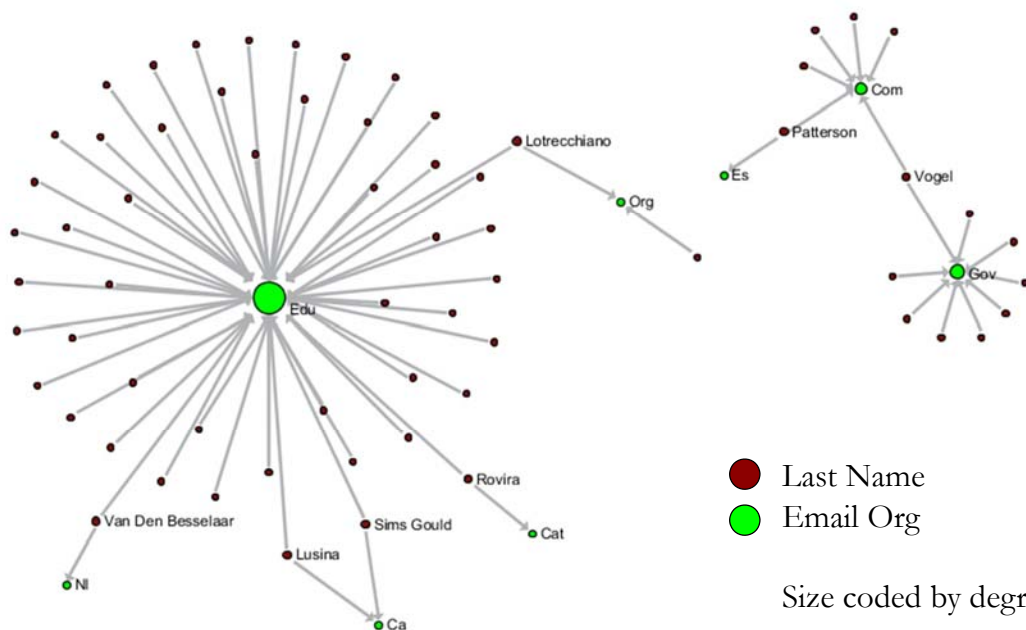
Optional: Calculate Node Degree

Visualize resulting *Bipartite network* from *Last Name and Org* using *Visualization > Network > GUESS* and *Layout > GEM*, *Layout > Bin Pack*

33



Sci2 Tool – Visualize Workshop Attendees



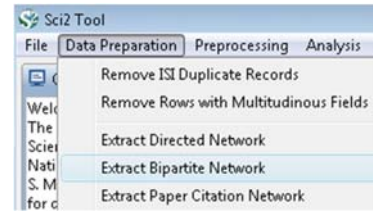
34



Sci2 Tool – Visualize Workshop Attendees

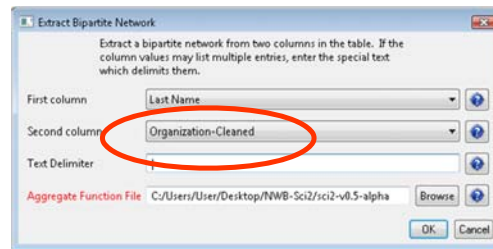
Use *File > Read* to load *SciTS Conf SNA Registrants report 4.10.11-clean.csv*

	A	B	C
1	Last Name	Org	Organization-Cleaned
2	Agoulnik	edu	Brigham and Women's Hospital, Harvard Medical School
3	Amaral	edu	Northwestern University
4	Bates	edu	University of Illinois at Chicago
5	Bennett	gov	NIH
6	Bietz	edu	University of California, Irvine
7	Bishop	edu	University of Tennessee
8	...		
9	Lotrecchiano	org.edu	George Washington University
10	Lusina	ca_edu	Centre for Hip Health & Mobility



Run *Data Preparation > Extract Bipartite Network*

With parameter values:



Optional: Calculate Node Degree

Visualize resulting *Bipartite network*

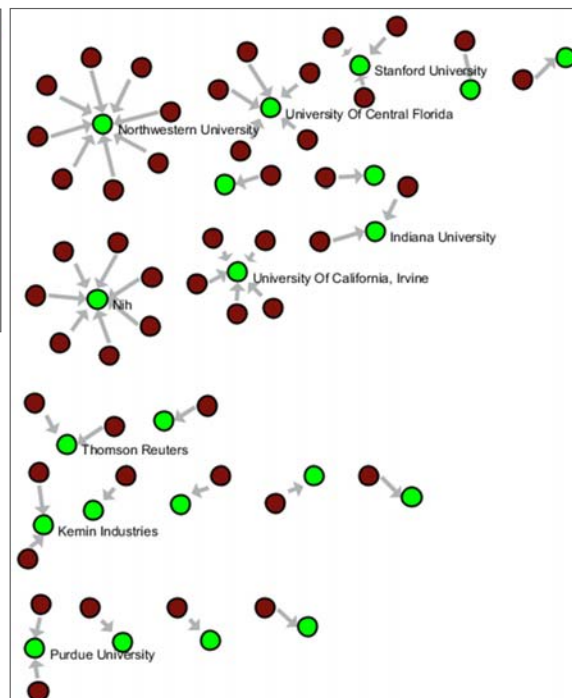
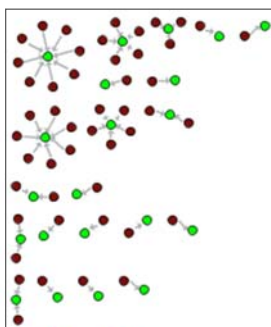
from *Last Name and Org* using *Visualization > Network > GUESS* and

Layout > GEM, *Layout > Bin Pack*

35



Sci2 Tool – Visualize Workshop Attendees



● Last Name
● Affiliation

36



Sci2 Tool – Visualize SciTS Co-Author Network Based on Holly’s EndNote File

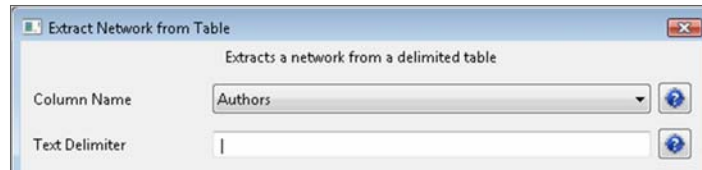
Open Holly’s ‘*SciTS-Library-03-04-2011.enl*’ in EndNote and save as ‘*SciTS-Library-03-04-2011.enw*’ following instructions on

<http://cishell.wiki.cns.iu.edu/Endnote+Export+Format>

Use ‘*File > Read*’ to load ‘*SciTS-Library-03-04-2011.enw*’

Run ‘*Data Preparation > Extract Co-Occurrence Network*’

With parameter values:



Optional: Calculate Node Degree

Visualize resulting ‘*Bipartite network*

from *Last Name and Org*’ using ‘*Visualization > Network > GUESS*’ and

‘*Layout > GEM*’, ‘*Layout > Bin Pack*’

37



Sci2 Tool – Visualize SciTS Co-Author Network Based on Holly’s EndNote File

.....
Network Analysis Toolkit (NAT) was selected.
Implementer(s): Timothy Kelley
Integrator(s): Timothy Kelley
Reference: Robert Sedgewick. Algorithms in Java, Third Edition, Part 5 - Graph Algorithms. Addison-Wesley, 2002. ISBN 0-201-31663-3. Section 19.8, pp.205
Documentation:
<http://wiki.cns.iu.edu/display/CISHELL/Network+Analysis+Toolkit+%28NAT%29>
This graph claims to be undirected.

Nodes: 706
Isolated nodes: 100
Node attributes present: label

Edges: 1687
No self loops were discovered.
No parallel edges were discovered.

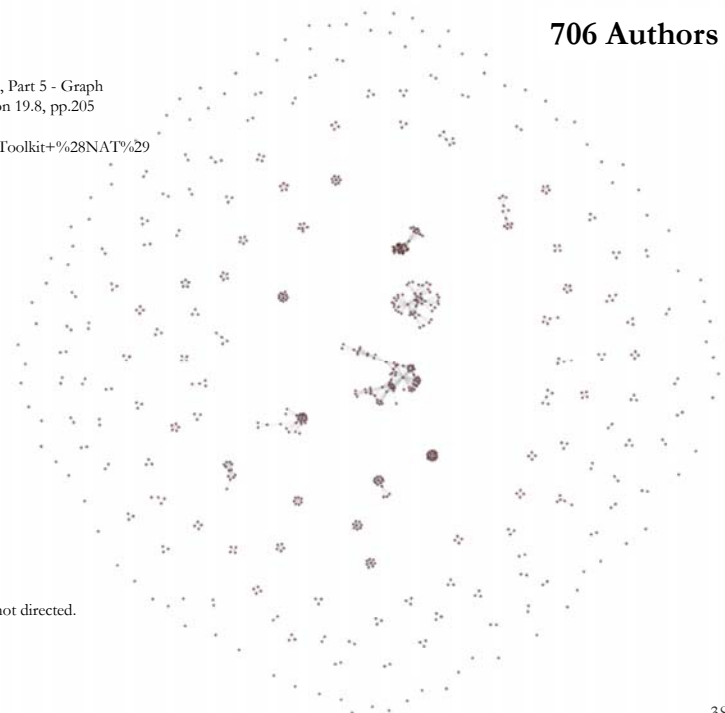
Edge attributes:
Did not detect any nonnumeric attributes.
Numeric attributes:
minmaxmean
weight 151.15412

This network seems to be valued.

Average degree: 4.779
This graph is not weakly connected.
There are 223 weakly connected components. (100 isolates)
The largest connected component consists of 73 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0.0068
Additional Densities by Numeric Attribute

706 Authors

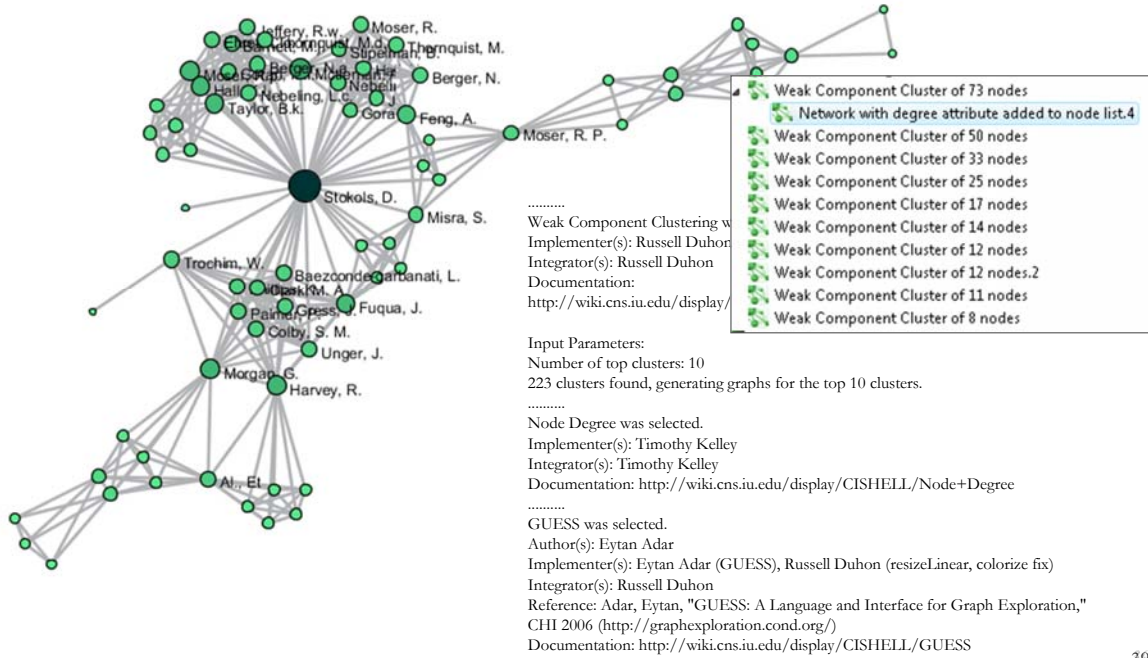


38

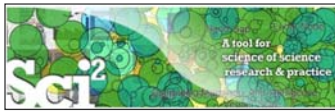


Sci2 Tool – Visualize SciTS Co-Author Network Based on Holly’s EndNote File

Largest “Giant” Component

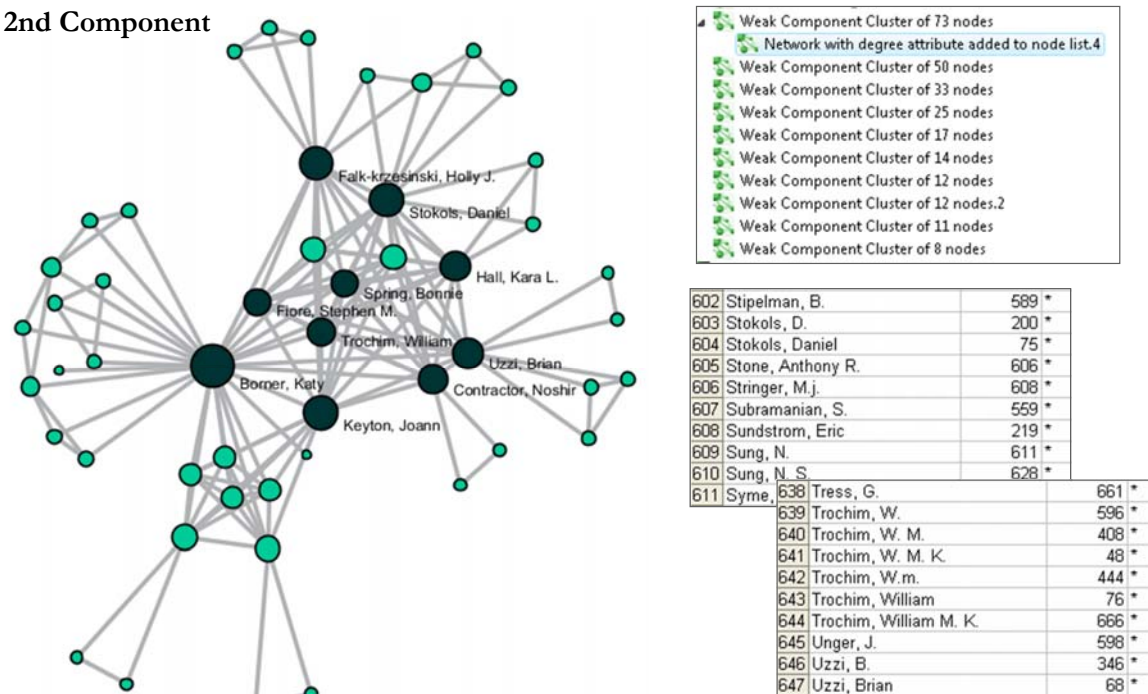


39



Sci2 Tool – Visualize SciTS Co-Author Network Based on Holly’s EndNote File

2nd Component



40



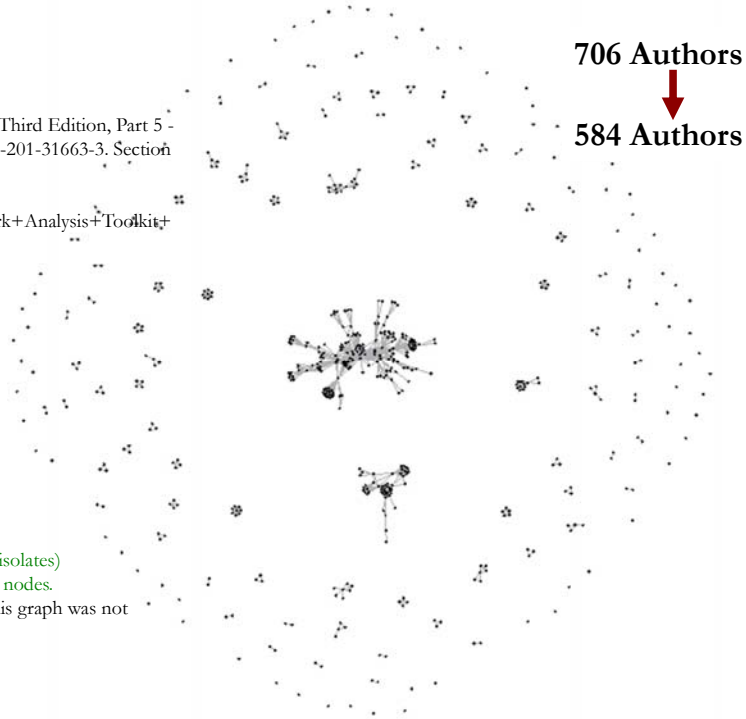
Sci2 Tool – Visualize SciTS Co-Author Network Based on Holly’s EndNote File – Last Names

.....
 Network Analysis Toolkit (NAT) was selected.
 Implementer(s): Timothy Kelley
 Integrator(s): Timothy Kelley
 Reference: Robert Sedgewick. Algorithms in Java, Third Edition, Part 5 - Graph Algorithms. Addison-Wesley, 2002. ISBN 0-201-31663-3. Section 19.8, pp.205
 Documentation:
<http://wiki.cns.iu.edu/display/CISHELL/Network+Analysis+Toolkit+%28NAT%29>
 This graph claims to be undirected.

Nodes: 584
 Isolated nodes: 79
 Node attributes present: label, totalDegree

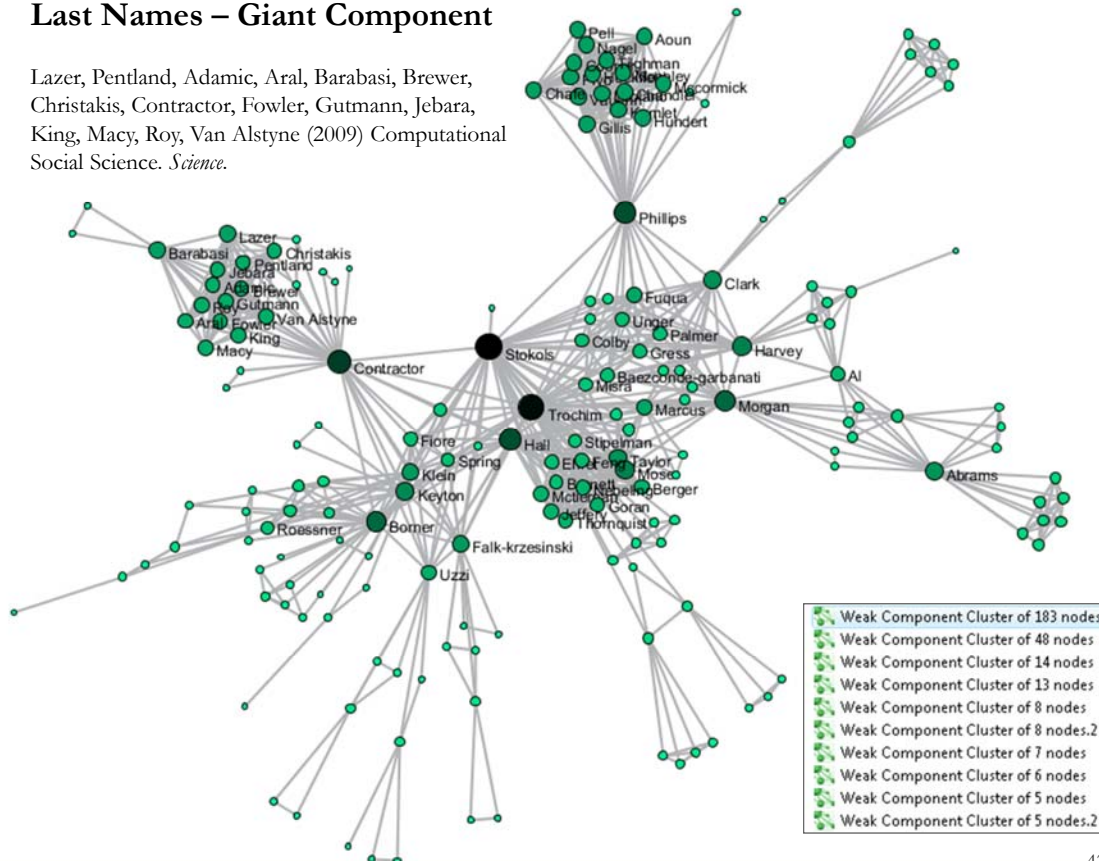
Edges: 1472
 No self loops were discovered.
 No parallel edges were discovered.

Average degree: 5.0411
 This graph is not weakly connected.
 There are 162 weakly connected components. (79 isolates)
 The largest connected component consists of 183 nodes.
 Did not calculate strong connectedness because this graph was not directed.
 Density (disregarding weights): 0.0086
 Additional Densities by Numeric Attribute



Last Names – Giant Component

Lazer, Pentland, Adamic, Aral, Barabasi, Brewer, Christakis, Contractor, Fowler, Gutmann, Jebara, King, Macy, Roy, Van Alstyne (2009) Computational Social Science. *Science*.



- Weak Component Cluster of 183 nodes
- Weak Component Cluster of 48 nodes
- Weak Component Cluster of 14 nodes
- Weak Component Cluster of 13 nodes
- Weak Component Cluster of 8 nodes
- Weak Component Cluster of 8 nodes.2
- Weak Component Cluster of 7 nodes
- Weak Component Cluster of 6 nodes
- Weak Component Cluster of 5 nodes
- Weak Component Cluster of 5 nodes.2



Workshop Overview

1:30 Macroscope Design and Usage & CShell Powered Tools: NWB & Sci2

1:45 Sci2 Tool Basics

- Download and run the tool.

2:00 Sci2 Sample Workflow: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

2:30 Sci2 Sample Workflow: Studying Four Major NetSci Researchers.

- Load and clean a dataset as text file; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize as either a circular hierarchy or network

3:30 Break

4:00 Sci2 Demo I: Geospatial maps with congressional districts

4:30 Sci2 Demo II: Evolving collaboration networks

4:45 Outlook and Discussion

5:00 Adjourn

43



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>

44



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

- Select network and run 'Visualization > GUESS' to open GUESS with file loaded.
- Apply 'Layout > GEM'.

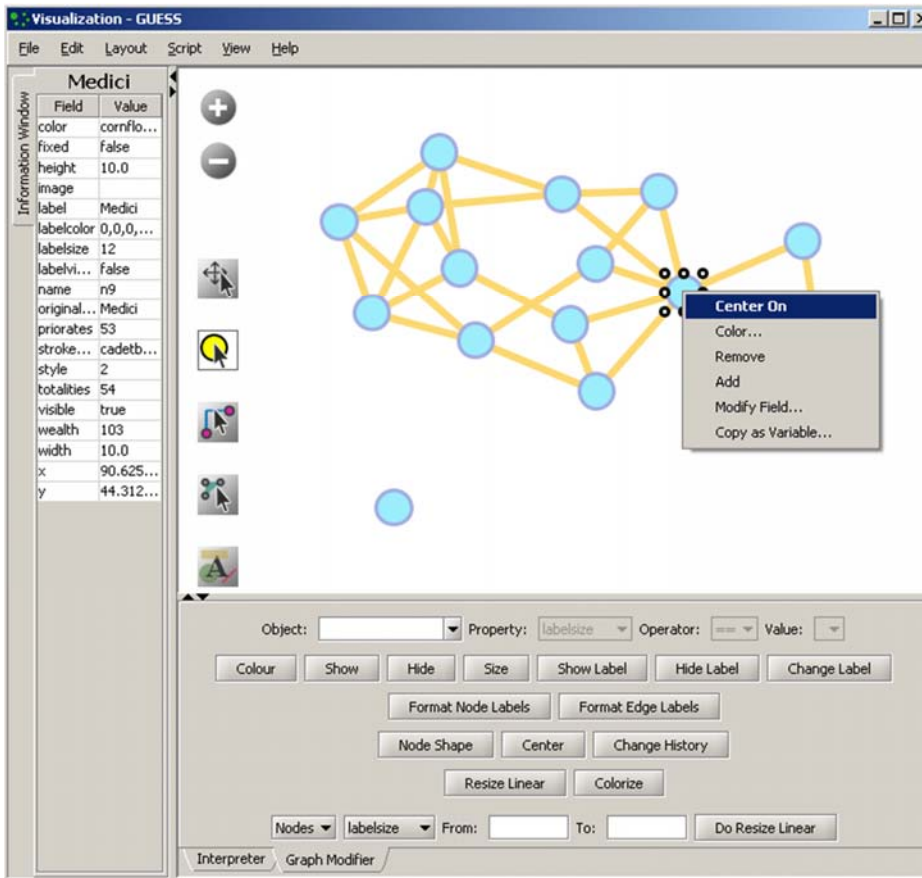
45

The screenshot shows the Network Workbench Tool interface. The main window displays a network visualization titled 'Medici-Acciaiuoli'. The network consists of 16 nodes and 27 edges, visualized as a cluster of blue nodes connected by yellow edges. The 'Information Window' on the left 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 on the left shows the GUESS execution log, including the command 'Visualization > GUESS' and the resulting network properties. The 'Data Manager' window on the right shows the loaded network file and its associated analysis results.

46

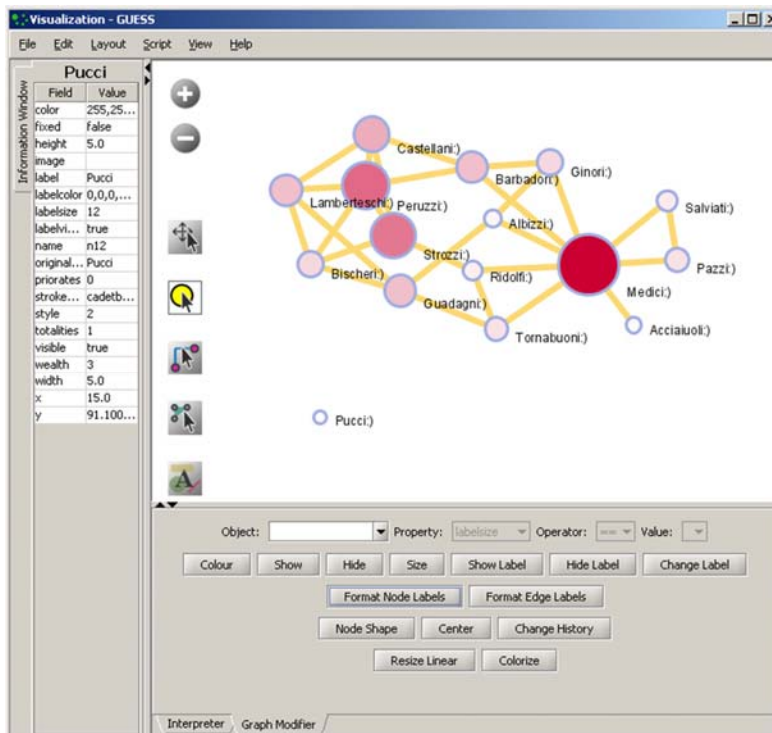


Pan:
 “grab” the background by holding left-click and moving your mouse.

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.’

The screenshot shows the 'Visualization - GUESS' application window. On the left is an 'Information Window' for the 'Acciaiuoli' node, displaying a table of fields and values. The main area shows a network graph with nodes of various sizes and colors (pink, white, red) connected by yellow edges. A tooltip for the 'Pucci' node is visible. At the bottom, an 'Interpreter' window shows the following code:

```
>>> resizeLinear(totalities,5,20)
>>> colorize(wealth,white,red)
>>>
```

Interpreter:
Uses Jython a combination of Java and Python.
Try
colorize(wealth, white, red)

49



Workshop Overview

1:30 Macroscopic Design and Usage & CShell Powered Tools: NWB & Sci2

1:45 Sci2 Tool Basics

- Download and run the tool.

2:00 Sci2 Sample Workflow: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

2:30 Sci2 Sample Workflow: Studying Four Major NetSci Researchers.

- **Load and clean a dataset; process raw data into networks.**
- **Find basic statistics and run various algorithms over the network.**
- **Visualize as either a circular hierarchy or network.**

3:30 Break

4:00 Sci2 Demo I: Geospatial maps with congressional districts

4:30 Sci2 Demo II: Evolving collaboration networks

4:45 Outlook and Discussion

5:00 Adjourn

50



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. 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.

[http://sci2.wiki.cns.iu.edu/5.1.4+Studying+Four+Major+NetSci+Researchers+\(ISI+Data\)](http://sci2.wiki.cns.iu.edu/5.1.4+Studying+Four+Major+NetSci+Researchers+(ISI+Data))

51



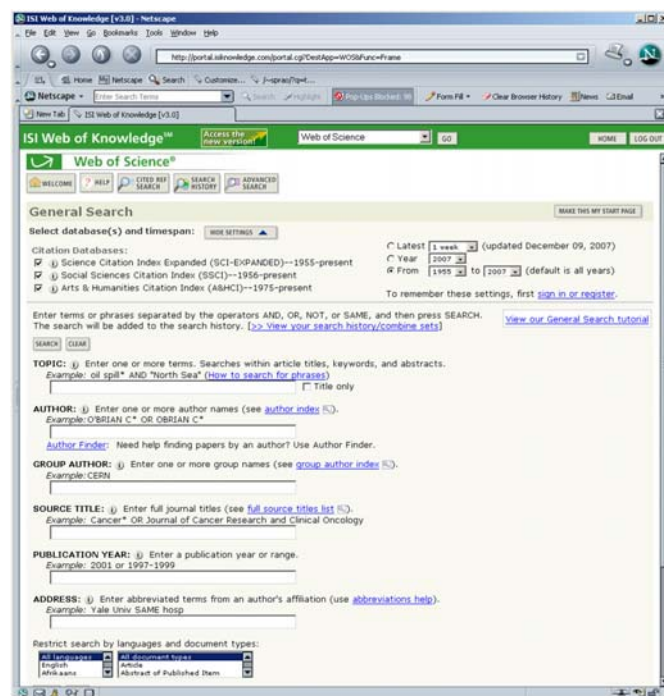
Data Acquisition from Web of Science

In Dec 2007, we downloaded 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



52

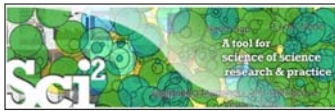


Comparison of Counts

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 <i>(Dec 2007)</i>
	41	16,920	159	52 <i>(Dec 2008)</i>
	44	30,102	201	68 <i>(April 11)</i>

53

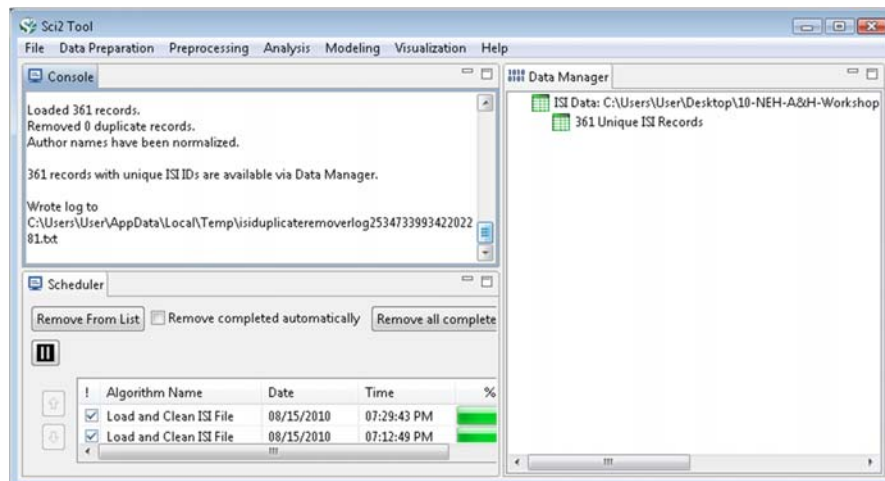


Extract Co-Author Network

Load **yoursci2directory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi'*
using *'File > Load ...'*

And file with 361 records appears in the Data Manager.

Duplicates were removed, author names normalized. Log file exists.



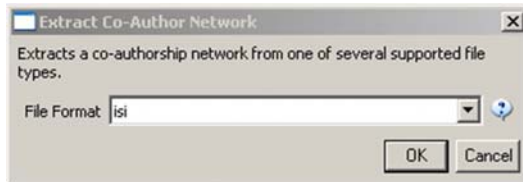
54



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 > 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 independent of co-authorship weight.

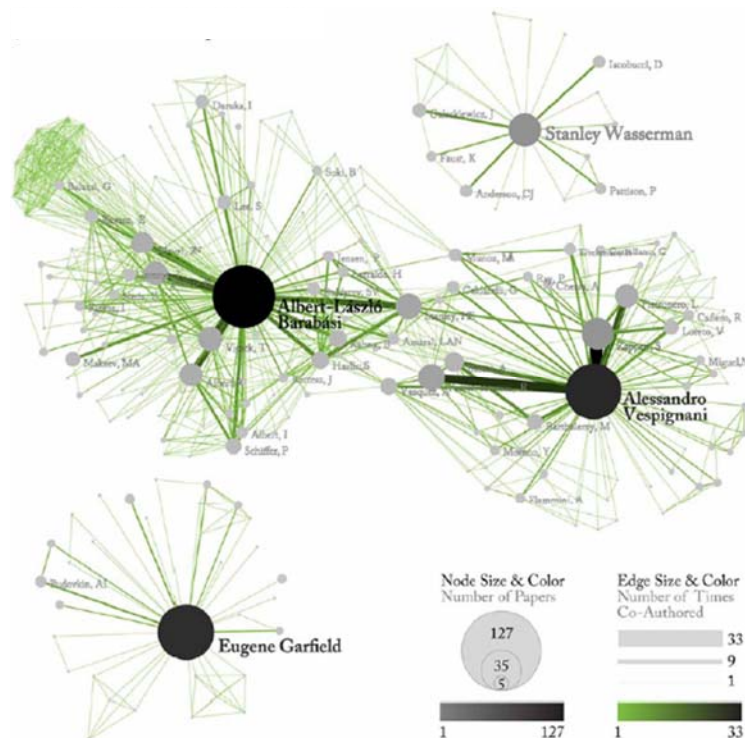
To view the complete network, select the 'Extracted Co-Authorship Network' and run 'Visualization > Networks > GUESS'.

Network is loaded with random layout. In GUESS, run 'Layout > GEM' and 'Layout > Bin Pack' to improve layout. Run 'Script > Run Script ...' and select 'yoursci2directory/scripts/GUESS/co-author-nw.py'.

55



Co-Author Network of all Four NetsSci Researchers



56

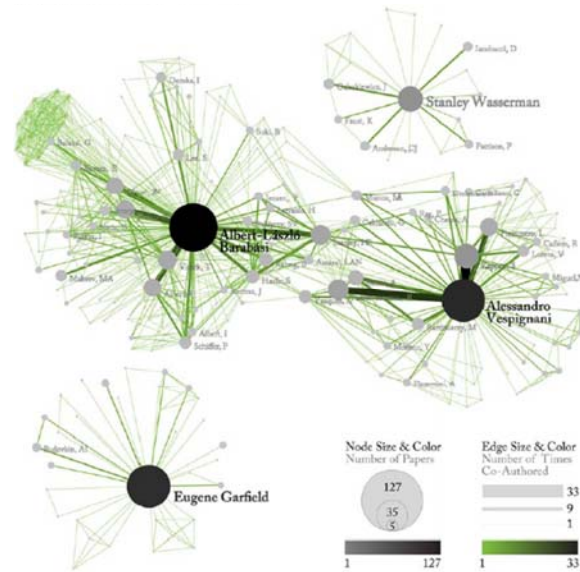


Co-Author Network of all Four NetsSci Researchers

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

Calculate node degrees in Sci2 Tool.

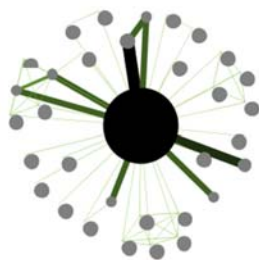
Use a graphic program to add legend.



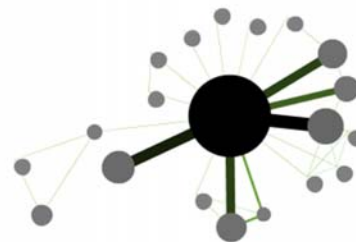
57



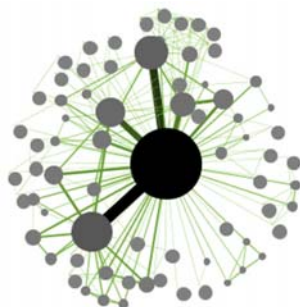
Individual Co-Author Networks (Read/map 4 files separately)



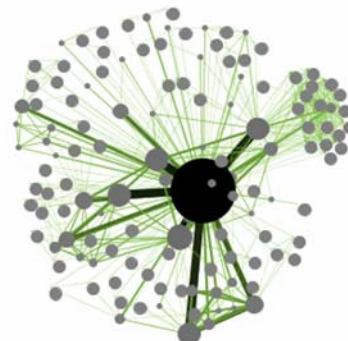
Eugene Garfield



Stanley Wasserman

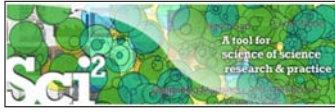


Alessandro Vespignani



Albert-László Barabási

58

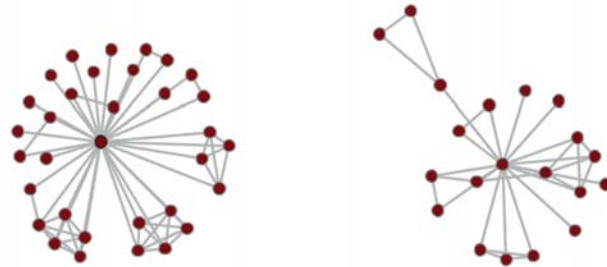
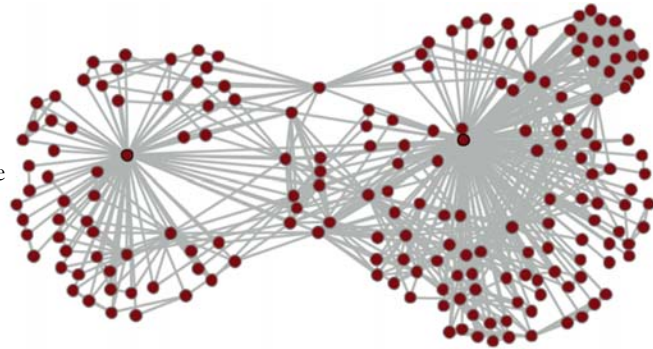


Network Visualization: Node Layout

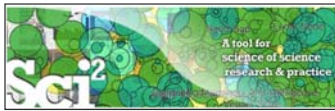
Load and Clean ISI File was selected.
Loaded 361 records.
Removed 0 duplicate records.
Author names have been normalized.
361 records with unique ISI IDs are available
via Data Manager.

.....
Extract Co-Author Network was selected.
Input Parameters:
File Format: isi

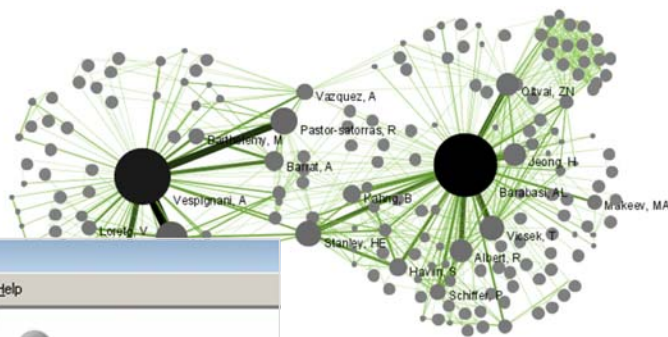
.....
Network Analysis Toolkit (NAT) was selected.
Nodes: 247
Edges: 891
.....
GUESS was selected.



59



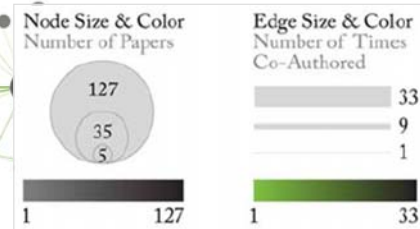
Network Visualization: Color/Size Coding by Data Attribute Values



Visualization - GUESS

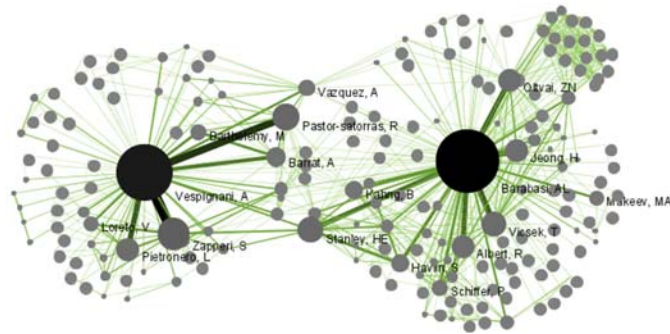
File Edit Layout Script View Help

Vespignani, A	
Field	Value
color	125,12,17,255
fixed	False
height	10.0
image	
label	Vespignani, A
labelcolor	0,0,0,255
labelsize	12
labelvisible	False
name	n161
numberofworks	101
originallabel	Vespignani, A
strokecolor	black
style	2
timescited	3811
visible	true
width	10.0
x	586.75
y	107.25





Network Visualization: Giant Component



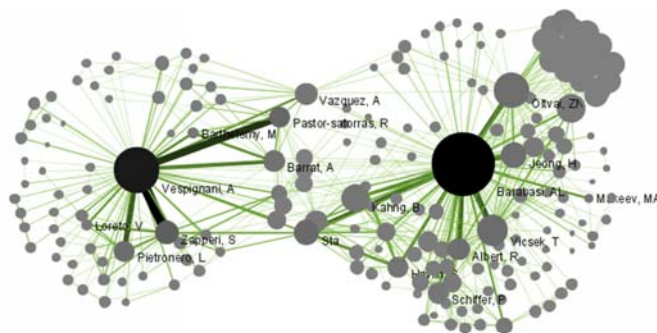
.....
Weak Component Clustering was selected.
Implementer(s): Russell Duhon
Integrator(s): Russell Duhon

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

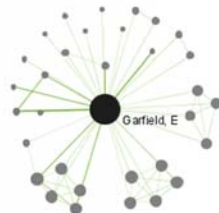
61



Network Visualization: Color/Size Coding by Degree



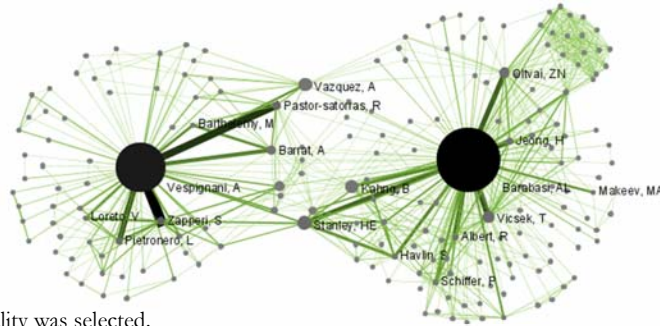
.....
Node Degree was selected.
Documentation:
<https://nwb.slis.indiana.edu/community/?n=AnalyzeData.No deDegree>
.....



62



Network Visualization: Color/Size Coding by Betweenness Centrality

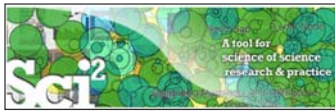
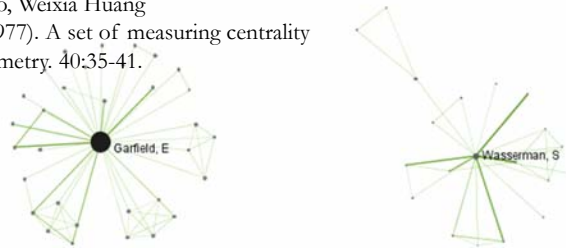


.....
Node Betweenness Centrality was selected.
Author(s): L. C. Freeman
Implementer(s): Santo Fortunato
Integrator(s): Santo Fortunato, Weixia Huang
Reference: Freeman, L. C. (1977). A set of measuring centrality based on betweenness. Sociometry. 40:35-41.

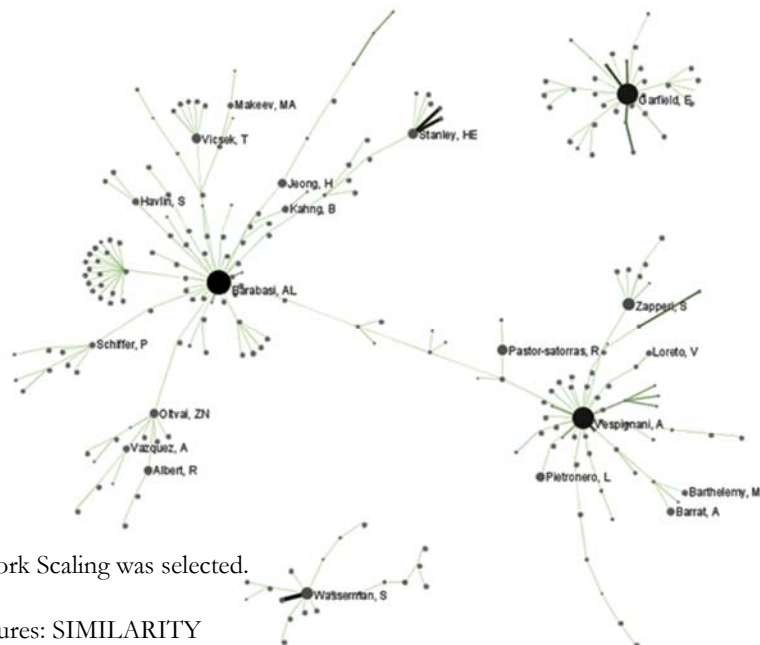
Input Parameters:
Number of bins: 10

umber of bins: 10

.....



Network Visualization: Reduced Network After Pathfinder Network Scaling



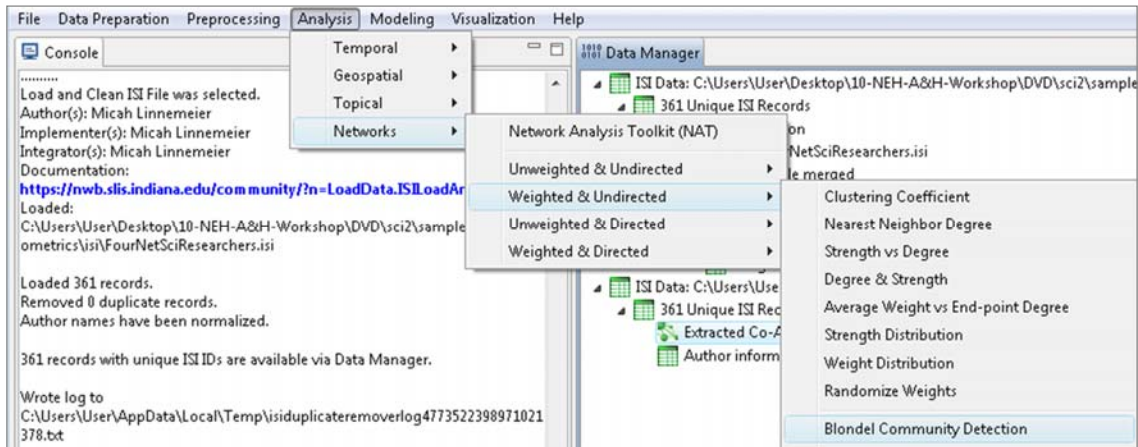
.....
MST-Pathfinder Network Scaling was selected.
Input Parameters:
Weight Attribute measures: SIMILARITY
Edge Weight Attribute: weight

.....

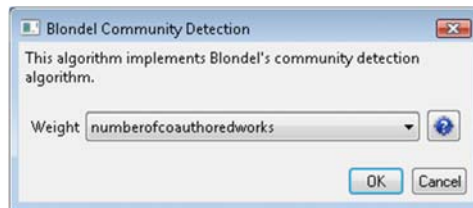


Network Visualization: Circular Hierarchy Visualization

Select Co-Author Network and run Blondel Community detection:



With parameter values

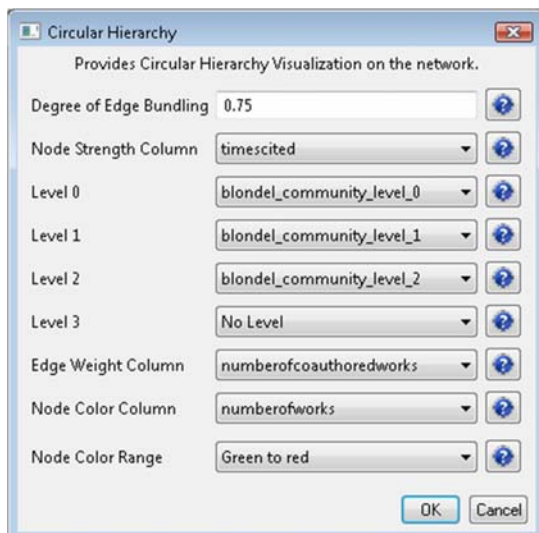


65



Network Visualization: Circular Hierarchy Visualization

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



66



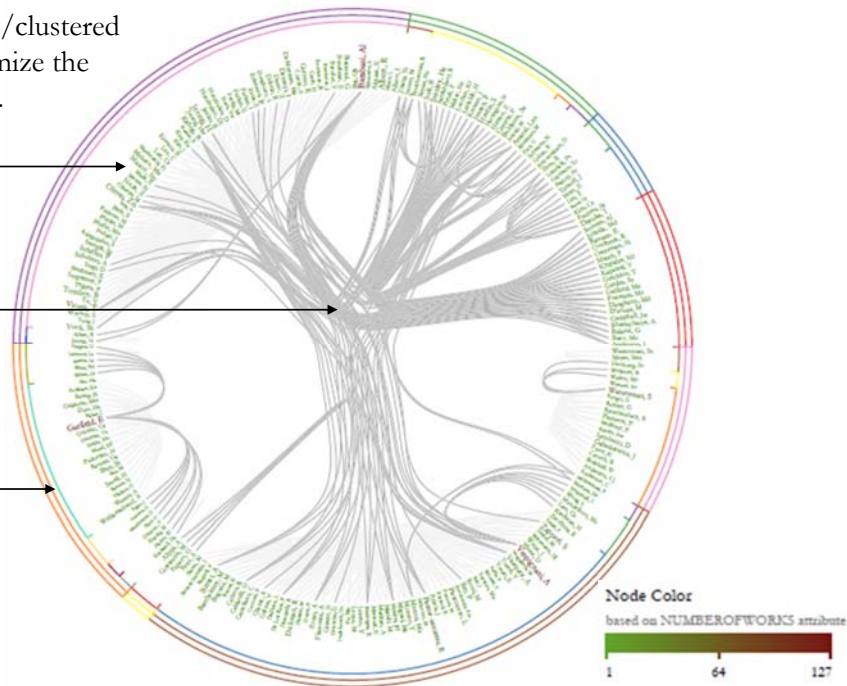
Network Visualization: Circular Hierarchy Visualization

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.

67



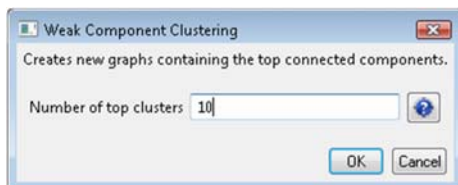
Paper-Citation Network Layout

To extract the paper-citation network, select the '361 Unique ISI Records' table and run *Data Preparation > 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)

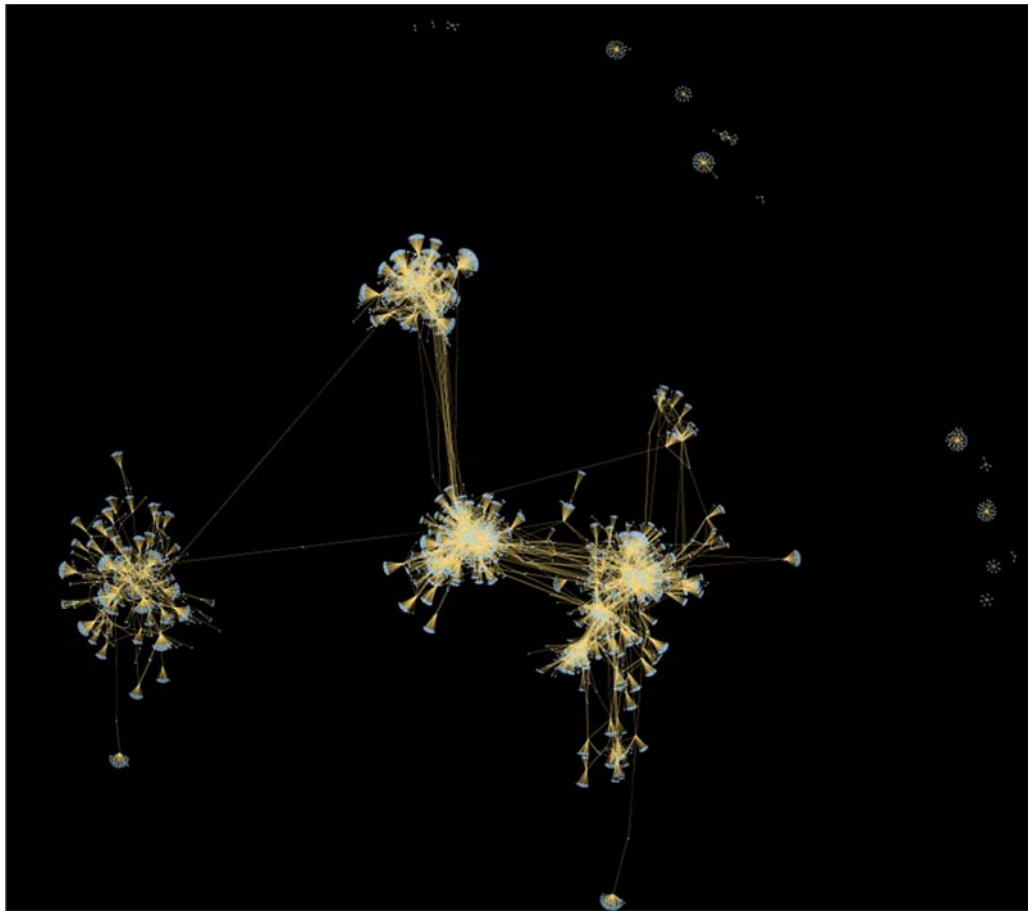
Run *'Analysis > Networks > Unweighted and Directed > Weak Component Clustering'* with parameters



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'*.

68

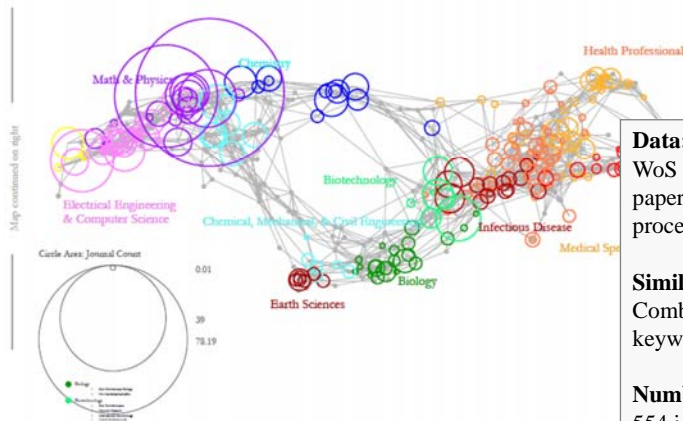


Topic Mapping: UCSD Science Map

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:

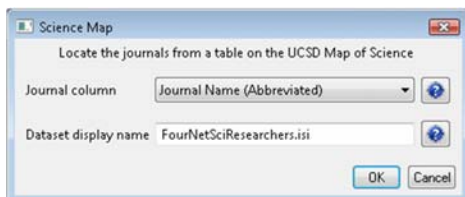
WoS and Scopus for 2001–2005, 7.2 million papers, more than 16,000 separate journals, proceedings, and series

Similarity Metric:

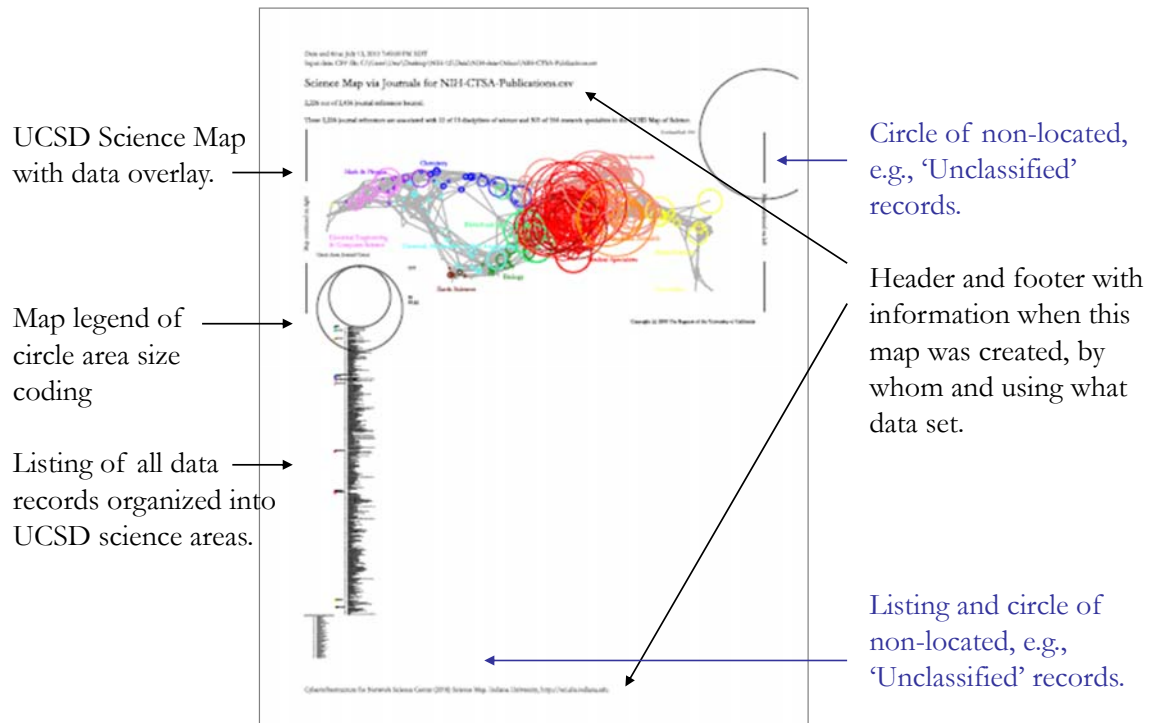
Combination of bibliographic coupling and keyword vectors

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.



How to Read the UCSD Map



71

Break

72



Workshop Overview

1:30 Macroscope Design and Usage & CShell Powered Tools: NWB & Sci2

1:45 Sci2 Tool Basics

➤ Download and run the tool.

2:00 Sci2 Sample Workflow: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

2:30 Sci2 Sample Workflow: Studying Four Major NetSci Researchers.

➤ Load and clean a dataset as text file; process raw data into networks.

➤ Find basic statistics and run various algorithms over the network.

➤ Visualize as either a circular hierarchy or network

3:30 Break

4:00 Sci2 Demo I: Geospatial maps with congressional districts

4:30 Sci2 Demo II: Evolving collaboration networks

4:45 Outlook and Discussion

5:00 Adjourn

73



Sci2 Demo I:

Geospatial maps with congressional districts

	A
1	Zip code
2	90095
3	4672
4	232980568
5	10032
6	10039242
7	46091500
8	191112434
9	27705
10	981959472
11	10065
12	10065



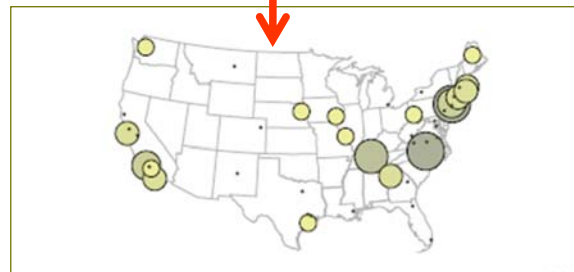
Identify Congressional District, Latitude, Longitude

	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



Aggregate/Count identical Congressional Districts

	A	B	C	D
1	Congressional District	Latitude	Longitude	Count
2	CA-30	34.0735035	-118.6645815	4
3	ME-02	45.818717	-69.0290345	2
4	VA-03	37.270472	-77.0699835	1
5	NY-15	40.8341475	-73.9342095	4



74



Relevant Sci2 Manual entry

- Home
- 1 Introduction
- 2 Getting Started
- 3 Algorithms, Tools, and Plugins
- 4 Workflow Design
- 5 Sample Workflows
 - 5.1 Individual Level Studies - Micro
 - 5.2 Institution Level Studies - Meso
 - 5.3 Global Level Studies - Macro
 - 5.3.1 Geo USPTO (SDB Data)
 - 5.3.2 Congressional District Geocoder
- 6 Sample Science Studies & Online Services
- 7 Extending the Sci2 Tool
- 8 Relevant Datasets and Tools
- 9 References
- Appendix 1 Glossary
- Appendix 2 CShell Algorithms
- Appendix 3 Sci2 Release Notes v0.5 alpha



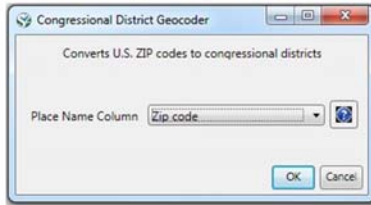
5.3.2 Congressional District Geocoder

14 Added by Scott Weingart, last edited by Ted Polley on Mar 28, 2011 (view change)

Tools ▾

zip code.csv	
Region(s):	United States
Analysis Type(s):	Geospatial Analysis

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: 'yoursci2directory/sampledata/geo/zipcode.csv'. Load the file in Standard csv format. Then select the file in the data manager and use 'Analysis > 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

<http://sci2.wiki.cns.in.edu/5.3.2+Congressional+District+Geocoder>

75



Relevant CShell plugin



Congressional District Geocoder

Tools ▾

1 Added by Ted Polley, last edited by Chin Hua Kong on Mar 29, 2011 (view change)

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.

<http://cishell.wiki.cns.in.edu/Congressional+District+Geocoder>

76



Console Messages

Load... was selected.
Documentation: <http://wiki.cns.iu.edu/display/CISHELL/Data+Formats>
Loaded: C:\Users\katy\Desktop\NWB-SCI2\sci2-2011.04.04-v0.5a\sampladata\geo\zip code.csv
.....
Congressional District Geocoder was selected.
Implementer(s): Chin Hua Kong
Integrator(s): Chin Hua Kong
Documentation: <https://nwb.cns.iu.edu/community/?n=SampleData.CongressionalDistrictGeocoder>

Input Parameters:

Place Name Column: Zip code

District values added to Congressional District, Latitude and Longitude respectively.

There are 2 rows with "33612" ZIP code, which could not been given a congressional district.

There are 1 rows with "2472" ZIP code, which could not been given a congressional district.

There are 3 rows with "10016" ZIP code, which could not been given a congressional district.

There are 1 rows with "11203" ZIP code, which could not been given a congressional district.

There are 1 rows with "60637" ZIP code, which could not been given a congressional district.

There are 1 rows with "70118" ZIP code, which could not been given a congressional district.

There are 1 rows with "60612" ZIP code, which could not been given a congressional district.

There are 3 rows with "21205" ZIP code, which could not been given a congressional district.

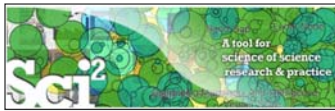
There are 1 rows with "2467" ZIP code, which could not been given a congressional district.

5-digit ZIP codes may often be insufficient, as many zip codes contain multiple congressional districts. 9-digit zip codes may be required. If a zip code was recently created, it may also not be contained in our database.

Successfully converted 86 out of 100 ZIP codes to congressional districts.

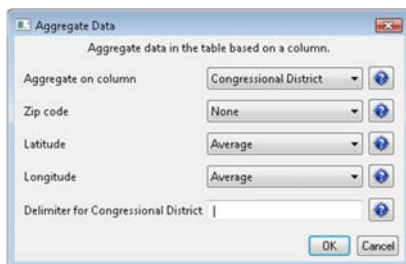
.....

77



Sci2 Demo I: Geospatial maps with congressional districts

Run *'Preprocessing > General > Aggregate Data'*
using parameter values



Note: Need lat/long for geomap.

Input Parameters:

Aggregate on column: Congressional District

Longitude: AVERAGE

Latitude: AVERAGE

Delimiter for Congressional District: |

Zip code: NONE

Aggregated by ": All rows of Latitude column were skipped due to no non-null, non-empty values.

Aggregated by ": All rows of Longitude column were skipped due to no non-null, non-empty values.

Frequency of unique "Congressional District" values added to "Count" column.

"Zip code" column has been deleted from the output. Since No aggregation was mentioned for it.

78



Create Geo Map (Circle Annotation)

.....

Geo Map (Circle Annotations) was selected.

Author(s): Joseph R. Biberstine

Implementer(s): Joseph R. Biberstine

Integrator(s): Joseph R. Biberstine

Documentation: <http://wiki.cns.iu.edu/display/CISHELL/Geo+Map>

Input Parameters:

Longitude: Longitude

Size Circles By: CircleSize

Color Circle Exteriors By: None (no outer color)

Color Circle Interiors By: CircleSize

Exterior Color Scaling: Linear

Exterior Color Range: Yellow to Blue

Interior Color Range: Blue to Red

Size Scaling: Linear

Map: US States

Author Name:

Interior Color Scaling: Linear

Latitude: Latitude

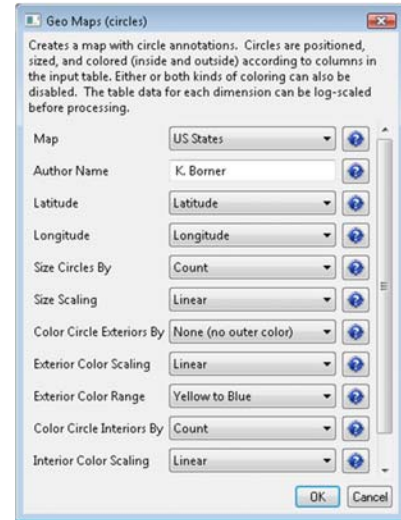
14 rows in the table did not specify all values needed to make a circle; those rows were skipped.

Printing PostScript..

Done.

Saved: C:\Users\katy\Desktop\geoMaps2903082942930990749.ps

Save ps file, convert to pdf, view.



79

How to Read the Geo Map

U.S. Map with data overlay.

Listing of map type, author, and parameters used.



Header and footer with information when this map was created, by whom and using what data set.

Map legend with color coding.



Workshop Overview

1:30 Macroscopic Design and Usage & CShell Powered Tools: NWB & Sci2

1:45 Sci2 Tool Basics

➤ Download and run the tool.

2:00 Sci2 Sample Workflow: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

2:30 Sci2 Sample Workflow: Studying Four Major NetSci Researchers.

➤ Load and clean a dataset as text file; process raw data into networks.

➤ Find basic statistics and run various algorithms over the network.

➤ Visualize as either a circular hierarchy or network

3:30 *Break*

4:00 Sci2 Demo I: Geospatial maps with congressional districts

4:30 Sci2 Demo II: Evolving collaboration networks

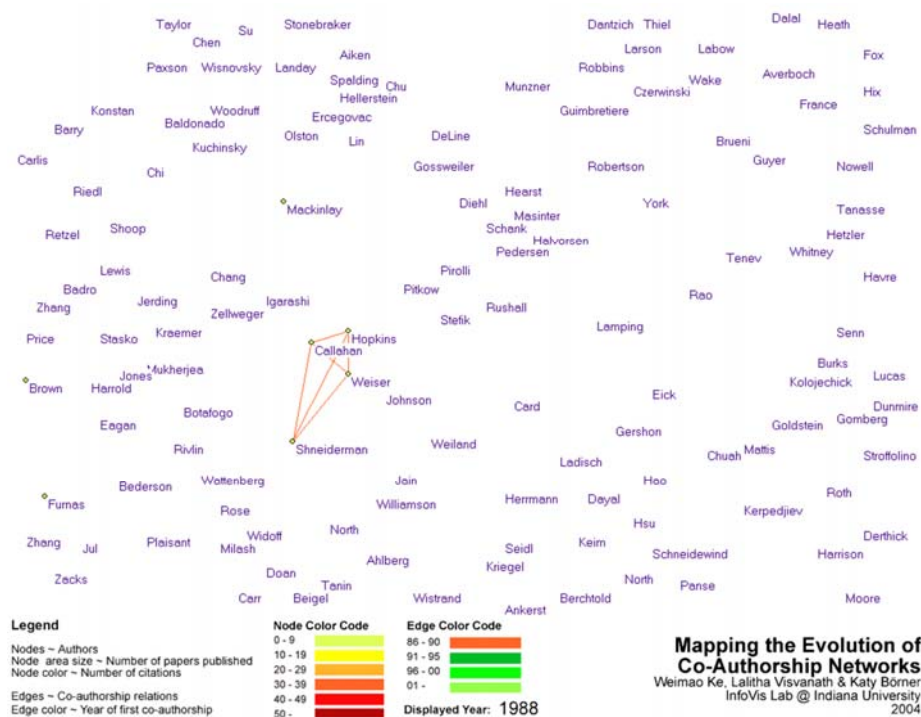
4:45 Outlook and Discussion

5:00 *Adjourn*

81



Sci2 Demo II: Evolving collaboration networks

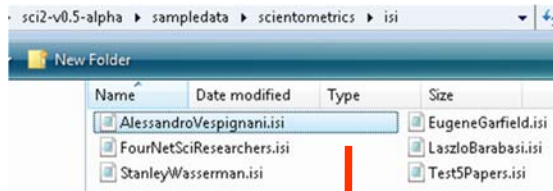


82



Sci2 Demo II: Evolving collaboration networks

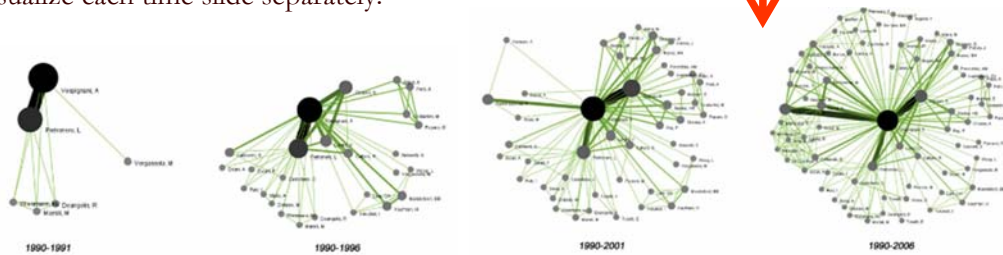
Load isi formatted file



As csv, file looks like:

	A	B	C	D	E	F	G
1	Abstract	Authors	Authors (Full Names)	Beginning	Book Serie	Book Serie	Cited Pate
2	The systematic study of	Colizza, V Barrat, A Barthelemy, M Vespignani, A		2015			
3	Uncovering the hidden r	Colizza, V Flammini, A Serrano, MA Vespignani, A		110			
4	Computer viruses can s	Vespignani, A		135			
5	Mapping the Internet ge	Dall'Asta, L Alvarez-Hamelin, I Barrat, A Vazquez, A Vespignani, A		140			LECTURE NOTES IN

Visualize each time slide separately:



83



Relevant Sci2 Manual entry

- Home
- 1 Introduction
- 2 Getting Started
- 3 Algorithms, Tools, and Plugins
- 4 Workflow Design
- 5 Sample Workflows
 - 5.1 Individual Level Studies - Micro
 - 5.1.1 Mapping Collaboration, Publication, and Funding Profiles of One Researcher (EndNote and NSF Data)
 - 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)
 - 5.2 Institution Level Studies - Meso
 - 5.3 Global Level Studies - Macro
- 6 Sample Science Studies & Online Services
- 7 Extending the Sci2 Tool
- 8 Relevant Datasets and Tools
- 9 References

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

Tools ▾

Added by Ted Polley, last edited by Scott Weingart on Mar 16, 2011 (view change)

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 Sci² 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: 'Slice the data into five year intervals from 1990-2006 using 'Preprocessing > Temporal' > 'Slice Table by Time' and the following parameters:

Slice Table by Time

Slice a table into groups of rows by time.

Date/Time Column: Publication Year

Date/Time Format: yyyy

Slice Into: Years

How Many?: 5

From Time: 1990

To Time: 2006

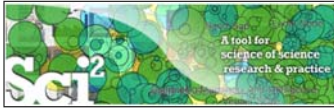
Cumulative?

Align With Calendar

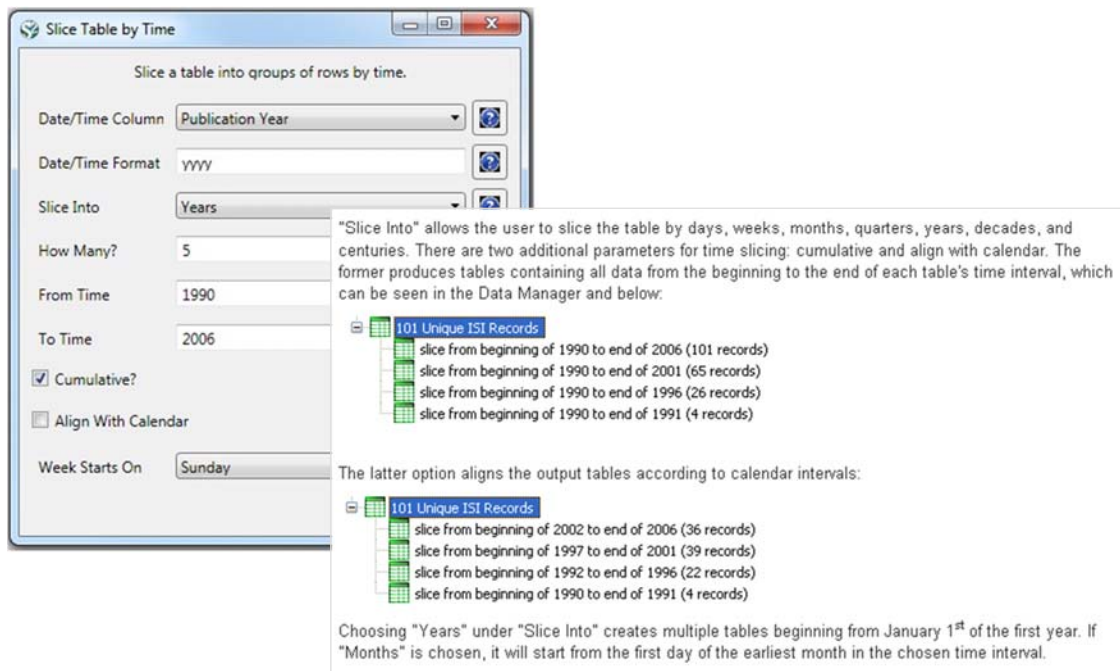
Week Starts On: Sunday

[http://sci2.wiki.cns.iu.edu/5.1.2+Time+Slicing+of+Co-Authorship+Networks+\(ISI+Data\)](http://sci2.wiki.cns.iu.edu/5.1.2+Time+Slicing+of+Co-Authorship+Networks+(ISI+Data))

84



Slice Table by Time



Slice a table into groups of rows by time.

Date/Time Column: Publication Year

Date/Time Format: YYYY

Slice Into: Years

How Many?: 5

From Time: 1990

To Time: 2006

Cumulative?

Align With Calendar

Week Starts On: Sunday

"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.

- 101 Unique ISI Records
 - slice from beginning of 1990 to end of 2006 (101 records)
 - slice from beginning of 1990 to end of 2001 (65 records)
 - slice from beginning of 1990 to end of 1996 (26 records)
 - slice from beginning of 1990 to end of 1991 (4 records)

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

- 101 Unique ISI Records
 - slice from beginning of 2002 to end of 2006 (36 records)
 - slice from beginning of 1997 to end of 2001 (39 records)
 - slice from beginning of 1992 to end of 1996 (22 records)
 - slice from beginning of 1990 to end of 1991 (4 records)

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.

[http://sci2.wiki.cns.iu.edu/5.1.2+Time+Slicing+of+Co-Authorship+Networks+\(ISI+Data\)](http://sci2.wiki.cns.iu.edu/5.1.2+Time+Slicing+of+Co-Authorship+Networks+(ISI+Data))

85



Visualize Each Network, Keep Node Positions

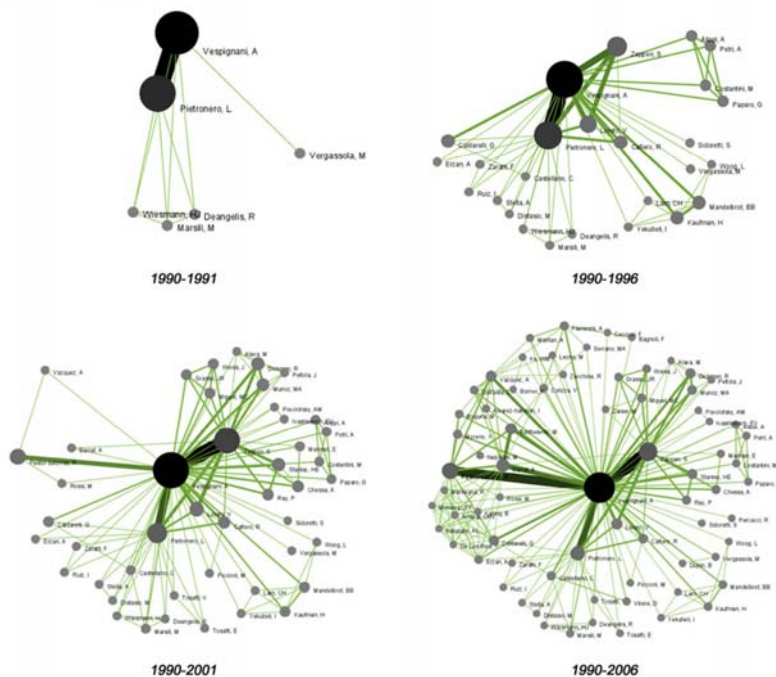
1. To see the evolution of Vespignani's co-authorship network over time, check 'cumulative'.
2. Extract co-authorship networks one at a time for each sliced time table using 'Data Preparation > Extract Co-Author Network', making sure to select "ISI" from the pop-up window during the extraction.
3. To 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'.
4. From here, run 'Layout > GEM' followed by 'Layout > Bin Pack'. Run 'Script > Run Script ...' and select 'yoursci2directory/scripts/GUESS/co-author-nw.py'.
5. In order to save the x, y coordinates of each node and to 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 'yoursci2directory/NodePositions.csv'.
6. 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.

[http://sci2.wiki.cns.iu.edu/5.1.2+Time+Slicing+of+Co-Authorship+Networks+\(ISI+Data\)](http://sci2.wiki.cns.iu.edu/5.1.2+Time+Slicing+of+Co-Authorship+Networks+(ISI+Data))

86



Visualize Each Network, Keep Node Positions



[http://sci2.wiki.cns.iu.edu/5.1.2+Time+Slicing+of+Co-Authorship+Networks+\(ISI+Data\)](http://sci2.wiki.cns.iu.edu/5.1.2+Time+Slicing+of+Co-Authorship+Networks+(ISI+Data))

87



Relevant CShell plugin



Tools ▾

Added by [Artha Alencar](#), last edited by [Ted Polley](#) on Jan 12, 2011 ([view change](#))

Description

Slice Table By Time is an algorithm to chop a table up into new tables, based on a date/time column. It takes the column with the date/time data, a string describing the format of that column, the intervals that the data should be sliced into, whether or not the slices are cumulative, whether or not the slices should be aligned with the calendar, and what day the week is considered to start on (which only matters if the slices are aligned with the calendar) as parameters.

The column to use for date/time values should have a single value for each row of data. It is used by the algorithm to choose which slice(s) the row should end up in. In order to determine what date/time is represented by that row, you must provide the algorithm with a descriptive format, in the second parameter. For instance, a four digit year would be represented by yyyy (the default value). See <http://joda-time.sourceforge.net/api-release/org/joda/time/format/DateTimeFormat.html> for details of all the various formatting options.

The next dropdown has the available intervals to slice the table into. These include milliseconds, seconds, minutes, hours, days, weeks, fortnights, months, quarters, years, decades, and centuries. A future version of the algorithm may include the ability to select how many of these intervals should be grouped together at once.

The checkbox that follows determines if the slices will be cumulative. If the slices are not cumulative, every row in the original table is in one and only one resulting slice. However, if the slices are cumulative, every row in the original table is in the slice it is for and every slice for a period after that.

The checkbox that follows determines if the slices will be aligned with the calendar. For instance, if the first row is for June 7th, 2006 and yearly slices are chosen, then the default behavior will be to have the first slice be from June 7th, 2006 to June 6th, 2007. However, if the slices are aligned with the calendar, the first slice will be from January 1st, 2006 to December 31st, 2006. Alignment does not affect the output for intervals of fortnights, quarters, decades, or milliseconds.

If the slices are aligned with the calendar and are weekly, then the day the week starts is used to determine how they are aligned.

Pros & Cons

The output of the slice algorithm is in separate tables, so a longitudinal analysis will require working with each slice separately, which can be awkward. There will likely be future versions of the time slice algorithm that annotate the original table with the slice the rows belong to.

Applications

When doing longitudinal analysis of data, it can be useful to consider it in chunks, such as to calculate how statistics have changed over time. Alternatively, only a particular time period might be of interest, and this algorithm can extract it from data for a larger time range.

Implementation Details

This algorithm uses the Joda Time library extensively, which provides significantly improved capabilities compared to the default Java algorithms for dates and times.

<http://cishell.wiki.cns.iu.edu/Slice+Table+by+Time>

88



Workshop Overview

1:30 Marcoscope Design and Usage & CShell Powered Tools: NWB & Sci2

1:45 Sci2 Tool Basics

- Download and run the tool.

2:00 Sci2 Sample Workflow: Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

2:30 Sci2 Sample Workflow: Studying Four Major NetSci Researchers.

- Load and clean a dataset as text file; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize as either a circular hierarchy or network

3:30 Break

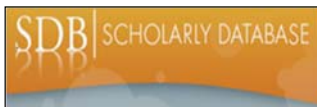
4:00 Sci2 Demo I: Geospatial maps with congressional districts

4:30 Sci2 Demo II: Evolving collaboration networks

4:45 Marcoscopes: Outlook and Discussion

5:00 Adjourn

89



Scholarly Database at Indiana University

<http://sdb.wiki.cns.iu.edu>

Supports federated search of 25 million publication, patent, grant records.

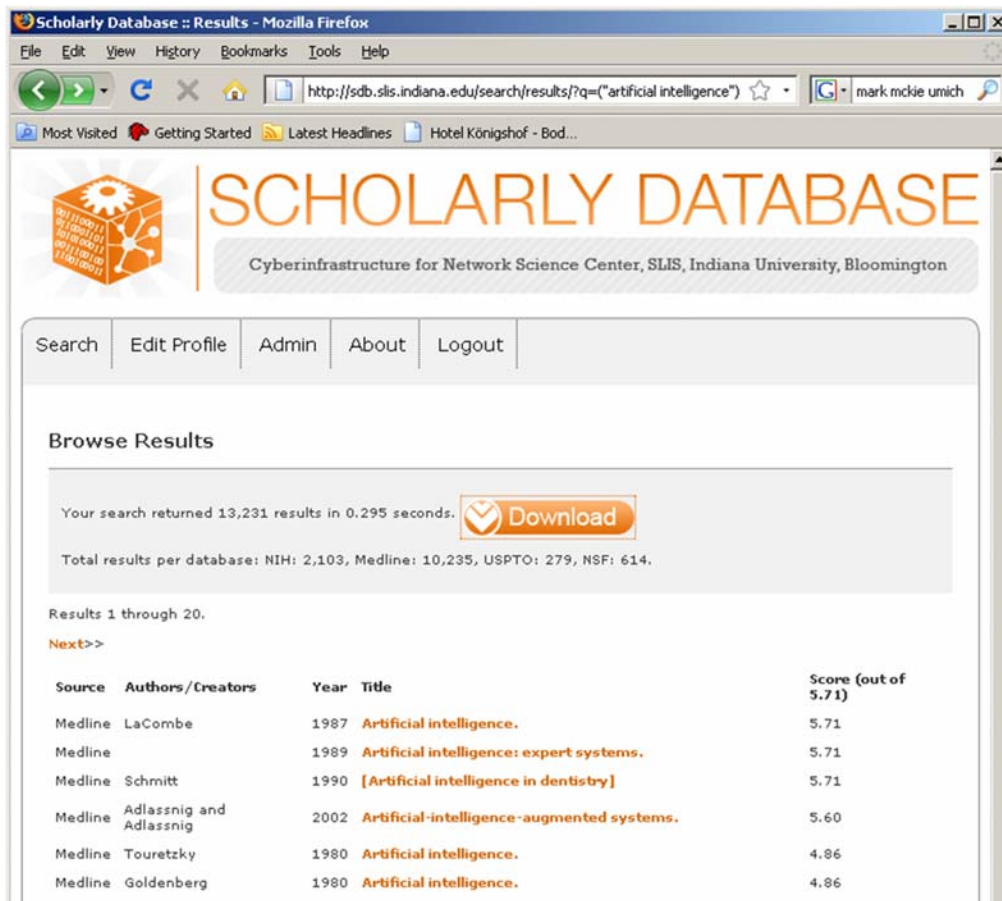
Results can be downloaded as data dump and (evolving) co-author, paper-citation networks.

The screenshot displays the Scholarly Database website. On the left is the login page with fields for 'IU User' and 'Non-IU User'. The 'IU User' section includes a 'Go to IU Login' button. The 'Non-IU User' section includes 'Email' and 'Password' fields with a 'Login' button. Below the login fields are links for 'Not Registered? Test!', 'Register as an IU User', and 'Register as a Non-IU User'. There is also a 'In the News' section with a citation for Wolfald, John (2006) and a 'Please Cite As' section with a citation for La Rosa, Sean, Ambler, Sumner, Burgess, John, Ye, Weisman and Böhm, Kath. (2007). At the bottom are logos for Indiana University School of Library and Information Science and the James S. McDonnell Foundation.

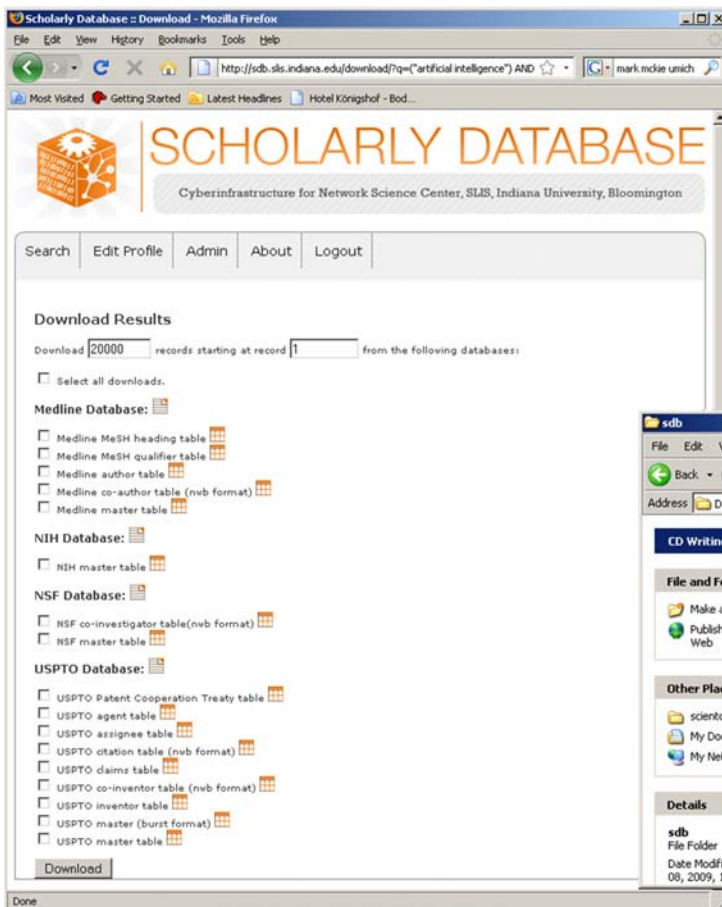
On the right is the search interface. It features a navigation menu with 'Search', 'Edit Profile', 'Admin', 'About', and 'Logout'. The search form includes fields for 'Creators', 'Title', 'Abstract' (with 'RNAi' entered), and 'Full Text'. There are dropdown menus for 'First Year' (1898) and 'Last Year' (2008). Below these are checkboxes for 'Medline (1898 - 2009)', 'NIH (1961 - 2002)', 'NSF (1985 - 2004)', and 'USPTO (1976 - 2007)'. A 'Search' button is at the bottom left. On the right side of the search interface, there is explanatory text about search operators: 'If multiple terms are entered in a field, they are automatically combined using "OR". So, "breast cancer" matches any record with "breast" or "cancer" in that field.', 'You can put AND between terms to combine with "AND". Thus "breast AND cancer" would only match records that contain both terms.', 'Double quotation can be used to match compound terms, e.g., "breast cancer" retrieves records with the phrase "breast cancer", and not records where "breast" and "cancer" are both present, but not the exact phrase.', and '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".'

Register for free access at <http://sdb.cns.iu.edu>

90



91

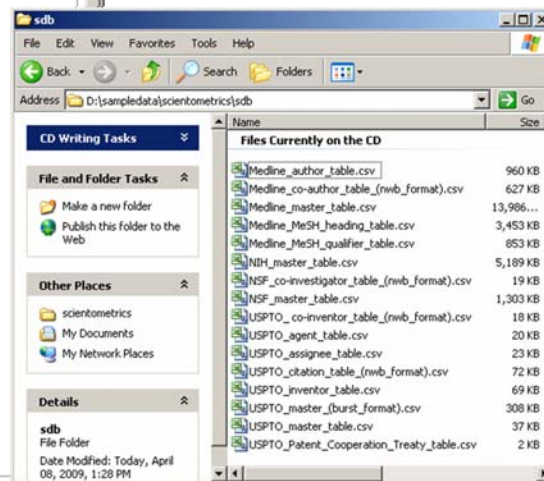


Since March 2009:

Users can download networks:

- Co-author
- Co-investigator
- Co-inventor
- Patent citation

and tables for burst analysis in NWB.



92

Borner, Katy
Person

This information is based solely on publications which have been loaded into the VIVO system. This may only be a small sample of the person's total work.

General Statistics

- 35 publication(s) from 2001 to 2010 [\(.CSV File\)](#)
- 87 co-author(s) from 2001 to 2010 [\(.CSV File\)](#)

Co-Author Network [\(GraphML File\)](#)

Legend

No. of publication(s) | No. of times(s) co-authored

Interact

Hover over any name to see the number of past publications and co-author with Borner, Katy. Click on a name to see details on the right.

Disambiguation

Only people that co-authored more than 1 paper(s) with Borner, Katy are shown.

Tables

Publications per Year [\(.CSV File\)](#)

Year	Count
2001	2
2002	4
2003	2
2004	7
2005	7
2006	3
2007	10
2010	1

Co-author(s) [\(.CSV File\)](#)

Author(s)	Count
Chen C.	5
Boyack K.W.	4
Mane K.K.	4
Ka W.	3
Penumarthy S.	3
Vespijnani, Alessandro	2
Hart B.	2
Hart E.	2
Holloway T.	2
Hart S.W.	2
Thakur S.	2
Feng Y.	2
Mane H.	2

Download Data

General Statistics

- 36 publication(s) from 2001 to 2010 [\(.CSV File\)](#)
- 80 co-author(s) from 2001 to 2010 [\(.CSV File\)](#)

Co-Author Network

[\(GraphML File\)](#)

Save as Image (.PNG file)

Tables

- Publications per year [\(.CSV File\)](#)
- Co-authors [\(.CSV File\)](#)

http://vivo-netsci.cns.iu.edu/vivo/visualization?uri=http%3A%2F%2Fvivo-trunk.indiana.edu%2Findividual%2FPerson74&vis=person_level&render_mode=standalone

93

36 publication(s) from 2001 to 2010 [\(.CSV File\)](#)

Year	Publications
2001	2
2002	4
2003	2
2004	7
2005	7
2006	3
2007	10
2010	1

80 co-author(s) from 2001 to 2010 [\(.CSV File\)](#)

Year	Count	Co-Author(s)
2001	1	Chen C.
2002	3	Chen C.; McMahon T.; Feng Y.
2003	2	Chen C.; Boyack K.W.
2004	17	Sengupta A.; Penumarthy S.; Thakur S.; Sooriamurthi R.; Maru J.T.; Shiffrin R.M.; Mane K.; Moor K.A.;

Co-author network [\(GraphML File\)](#)

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <graphml xmlns="http://graphml.graphdrawing.org/xmlns"
3 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4 xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
5 http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
6 <key id="label" for="node" attr.name="label" attr.type="string" />
7 <key id="number_of_authored_works" for="node" attr.name="number_of_authored_works" attr.type="int" />
8 <key id="num_unknown_publication" for="node" attr.name="num_unknown_publication" attr.type="int" />
9 <key id="num_latest_publication" for="node" attr.name="num_latest_publication" attr.type="int" />
10 <key id="latest_publication" for="node" attr.name="latest_publication" attr.type="int" />
11 <key id="profile_url" for="node" attr.name="profile_url" attr.type="string" />

```

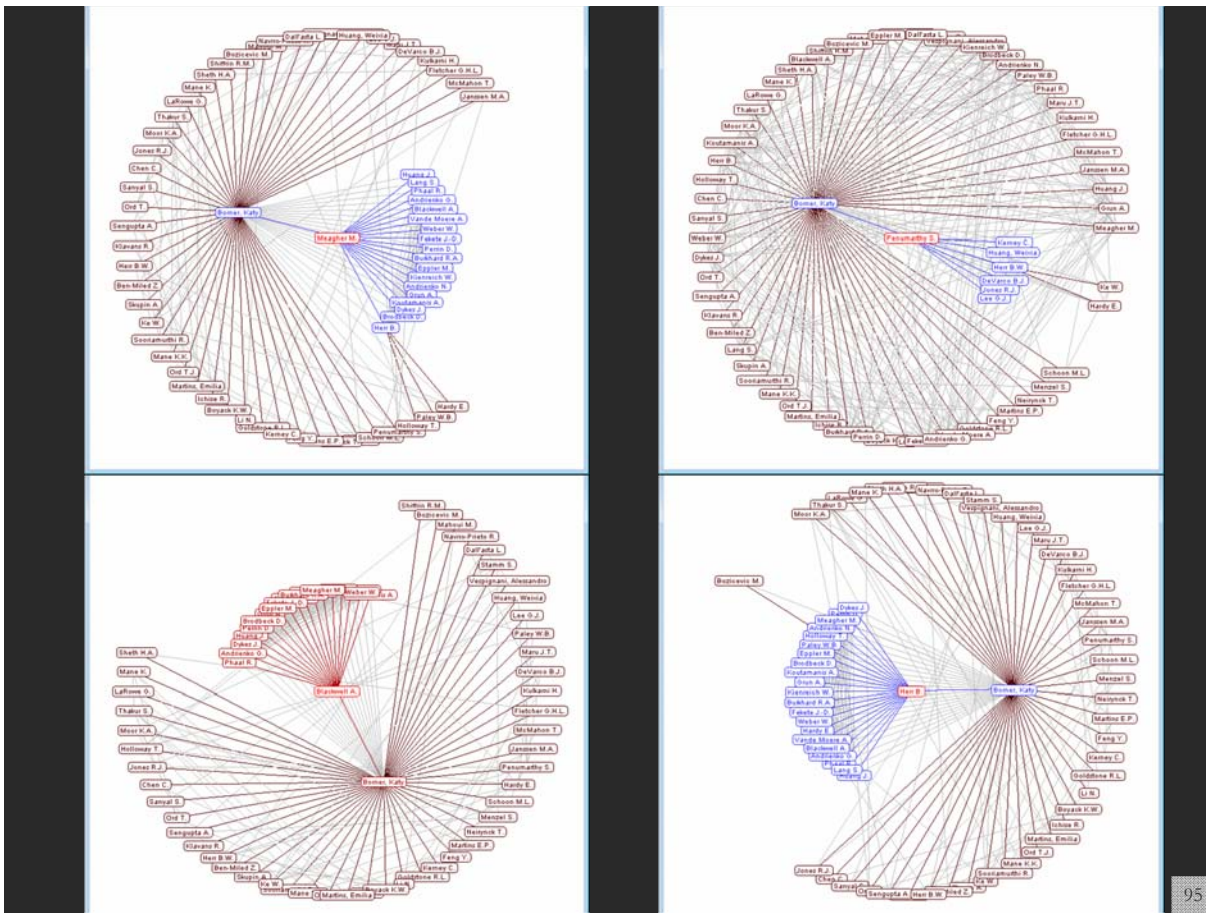
Save as Image (.PNG file)

Publications per year [\(.CSV File\)](#), see top file.

Co-authors [\(.CSV File\)](#)

Co-Author	Count
Andrienko G.	1
Andrienko N.	1
Ben-Miled Z.	1
Blackwell A.	1
Boyack K.W.	4
Bozicevic M.	1
Brodbeck D.	1
Burkhard R.A.	1
Chen C.	5

94



CIShell – Integrate New Algorithms

About the Cyberinfrastructure Shell

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.

CIShell is the basis of [Network Workbench](#), [TexTrend](#), [Sci2](#) and the upcoming [EpiC](#) tool.

CIShell supports remote execution of algorithms. A standard web service definition is in development that will allow pools of algorithms to transparently be used in a peer-to-peer, client-server, or web front-end fashion.

CIShell Features

A framework for easy integration of new and existing algorithms written in any programming language

Using CIShell, an algorithm writer can fully concentrate on creating their own algorithm in whatever language they are comfortable with. Simple tools are provided to then take their algorithm and

Learn More...

- [CIShell Papers](#)
- [CIShell Powered Tools](#)
- [Algorithms](#)
- [Plugins \(coming soon\)](#)
- [Misc. Tool Documentation](#)
- [CIShell Web Services \(coming soon\)](#)
- [Screenshots](#)

Getting Started...

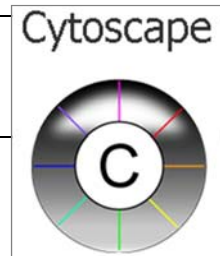
- [Documentation & Developer Resources](#)
- [Download](#)

Getting Involved...

- [Contact Us](#)

CIShell Developer Guide is at <http://cishell.wiki.cns.iu.edu>

Additional Sci2 Plugins are at <http://sci2.wiki.cns.iu.edu/3.2+Additional+Plugins>

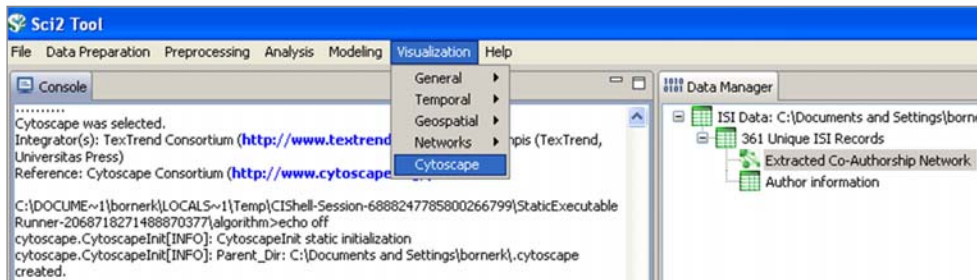


Adding more layout algorithms and network visualization interactivity via Cytoscape <http://www.cytoscape.org>.

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

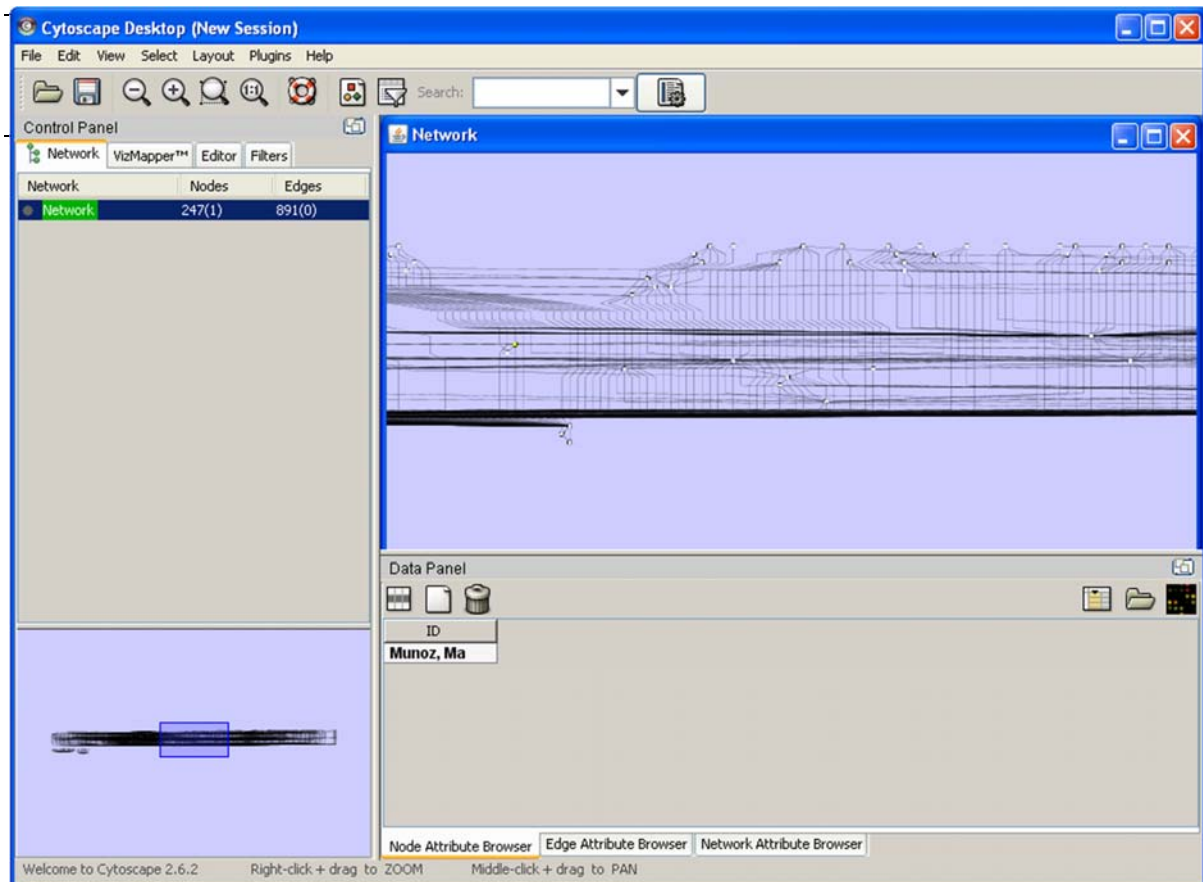
Restart Sci2 Tool.

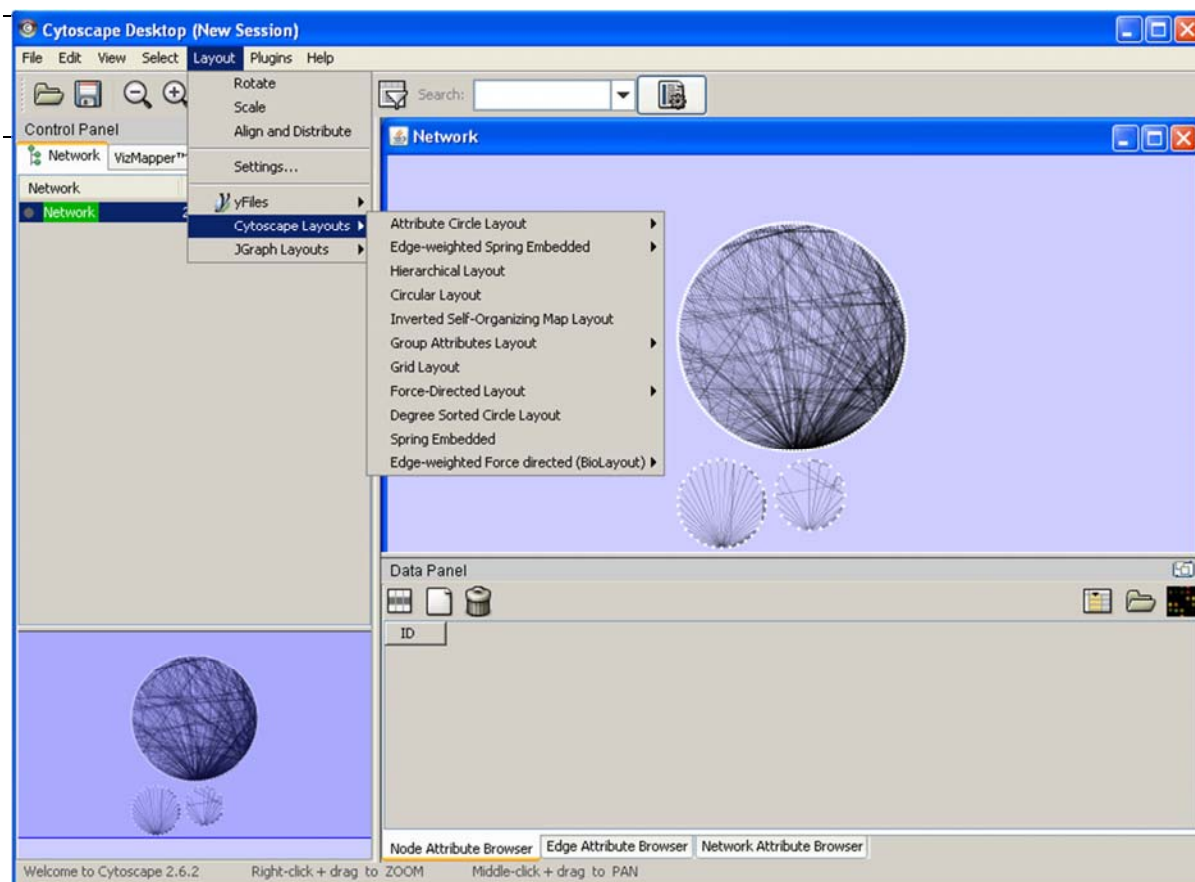
Cytoscape now shows in the Visualization Menu.



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

97





OSGi/CISHell Adoption

A number of other projects recently adopted OSGi and/or CISHell:

Cytoscape (<http://cytoscape.org>) Led by Trey Ideker at the University of California, San Diego 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.org.uk>) Developed by the myGrid team (<http://mygrid.org.uk>) led by Carol Goble at the University of Manchester, U.K. 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 Jong Lee at NCSA is an open-source, extensible software platform which supports seismic risk assessment based on the Mid-America Earthquake (MAE) Center research.
- *TEXTrend* (<http://textrend.org>) Led by George Kamps at Eötvös Loránd University, Budapest, Hungary supports natural language processing (NLP), classification/mining, and graph algorithms for the analysis of business and governmental text corpora with an inherently temporal component.
- *DynaNets* (<http://www.dynanets.org>) Coordinated by Peter M.A. Sloot at the University of Amsterdam, The Netherlands develops algorithms to study evolving networks.

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.

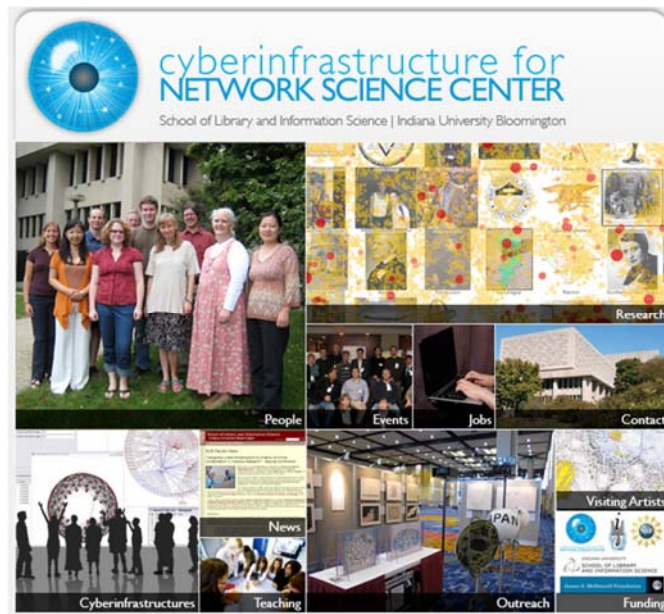
TEXTrend adds R bridge, WEKA, Wordij, CFinder, and more.

See the latest versions of TEXTrend Toolkit modules at

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

101

102



All papers, maps, tools, talks, press are linked from <http://cns.iu.edu>

CNS Facebook: <http://www.facebook.com/cnscenter>

Mapping Science Exhibit Facebook: <http://www.facebook.com/mappingscience>