



## Plug-and-Play Macroscopes Tutorial

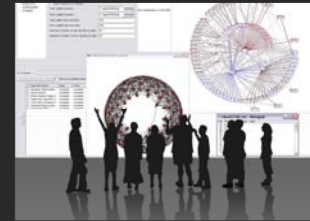
**Dr. Katy Börner**

Cyberinfrastructure for Network Science Center, Director  
 Information Visualization Laboratory, Director  
 School of Library and Information Science  
 Indiana University, Bloomington, IN  
[katy@indiana.edu](mailto:katy@indiana.edu)

Assisted by **Angela M. Zoss**

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,  
 Scott Weingart

*Social Computing, Behavioral Modeling, and Prediction, Natcher Conference Center, NIH  
 March 29, 2010 8:30-12:30am and 1-5pm*



### Overview

- Macroscopes and the Changing Scientific Landscape 15 mins
  - Introduction to network science with sample maps and insights 15 mins
  - Introduction to the Network Workbench Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins

---

  - Introduction to science studies with sample maps and insights 15 mins
  - Introduction to the Sci<sup>2</sup> Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - Overview of validation approaches for science studies
  - Plug-and-play tool design using OSGi/CIshell 30 mins
  - Scholarly Marketplaces 15 mins
- 240 mins**



These slides are available at

<http://info.slis.indiana.edu/~katy/outgoing/10-NIH-Tutorial.pdf>



## 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).

**Users -> Contributors:** Web 2.0 technologies empower anybody to contribute to Wikipedia and 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.

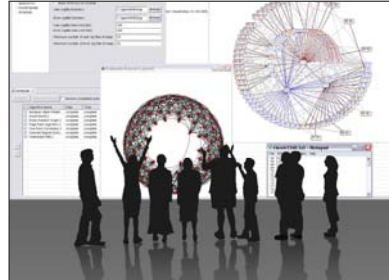
**Modularity:** The design of software modules with well defined functionality that can be flexibly combined helps reduce costs, makes it possible to have many contribute, and increases flexibility in tool development, augmentation, and customization.

**Standardization:** Adoption of 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.



## 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.

5



## Desirable Features of Plug-and-Play Macroscopes

**Division of Labor:** Ideally, labor is divided in a way that the expertise and skills of computer scientists are utilized for the design of standardized, modular, easy to maintain and extend “core architecture”. Dataset and algorithm plugins, i.e., the “filling”, are initially 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. Wizard-driven integration of new algorithms and data sets by domain experts, sharing via email or online sites, deploying plugins by adding them to the ‘plugin’ directory, and running them via a Menu driven user interfaces (as used in Word processing systems or Web browsers) seems to work well.

**Plugin Content and Interfaces:** Should a plugin represent one algorithm or an entire tool? What about data converters needed to make the output of one algorithm compatible with the input of the next? Should those be part of the algorithm plugin or should they be packaged separately?

**Supported (Central) Data Models:** Some tools use a central data model to which all algorithms conform, e.g., Cytoscape, see Related Work section. Other tools support many internal data models and provide an extensive set of data converters, e.g., Network Workbench, see below. The former often speeds up execution and visual rendering while the latter eases the integration of new algorithms. In addition, most tools support an extensive set of input and output formats.

**Core vs. Plugins:** As will be shown, the “core architecture” and the “plugin filling” can be implemented as sets of plugin bundles. Answers to questions such as: “Should the graphical user interface (GUI), interface menu, scheduler, or data manager be part of the core or its filling?” will depend on the type of tools and services to be delivered.

**Supported Platforms:** If the software is to be used via Web interfaces then Web services need to be implemented. If a majority of domain experts prefers a stand-alone tool running on a specific operating system then a different deployment is necessary.

6



## CIs Developed and Served by CNS



Scholarly Database: 23 million scholarly records

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



Information Visualization Cyberinfrastructure

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



Network Workbench Tool + Community Wiki

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



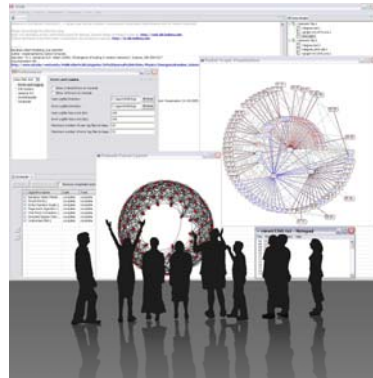
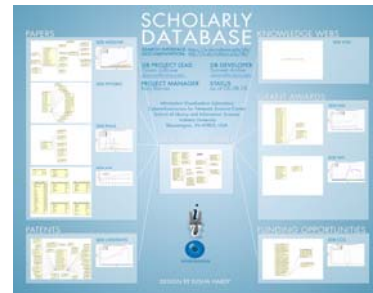
Sci<sup>2</sup> Tool and Science of Science CI Portal

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



Epidemics Cyberinfrastructure

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



7



## Overview

- Macroscopes and the Changing Scientific Landscape 15 mins
  - **Introduction to network science with sample maps and insights 15 mins**
  - Introduction to the Network Workbench Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - Introduction to science studies with sample maps and insights 15 mins
  - Introduction to the Sci<sup>2</sup> Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - **Overview of validation approaches for science studies**
  - Plug-and-play tool design using OSGi/CIshell 30 mins
  - Scholarly Marketplaces 15 mins
- 240 mins**

8

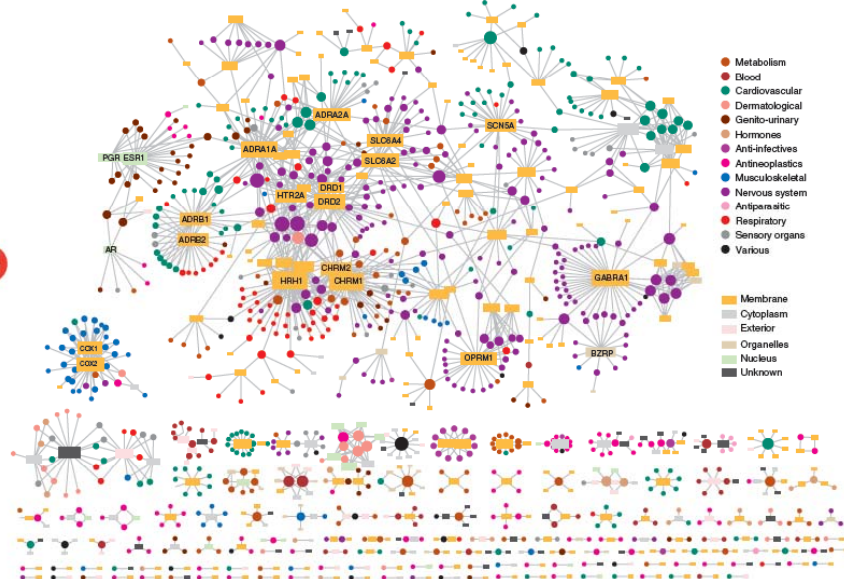


## Computational Proteomics

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

© 2007 Nature Publishing Group



**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.



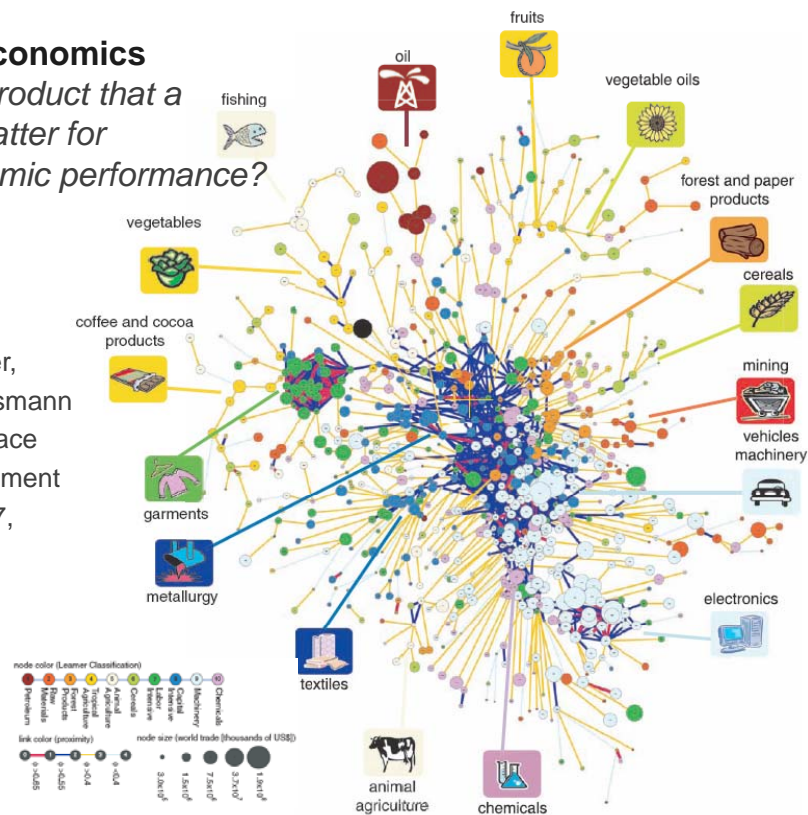
Network Workbench (<http://nwb.slis.indiana.edu>).

9

## 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 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.



10

# Computational Social Science

Studying large scale social networks such as Wikipedia

## Second sight

Image: Bruce W. Herr and Todd M. Holloway

### Vizzards 2007 Entry

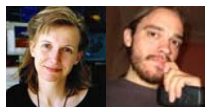
Second Sight: An Emergent Mosaic of Wikipedian Activity, The NewScientist, May 19, 2007

#### 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



locked until the mood cools (locked 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.



www.newscientist.com

19 May 2007 | NewScientist | 55

# Computational Epidemics

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

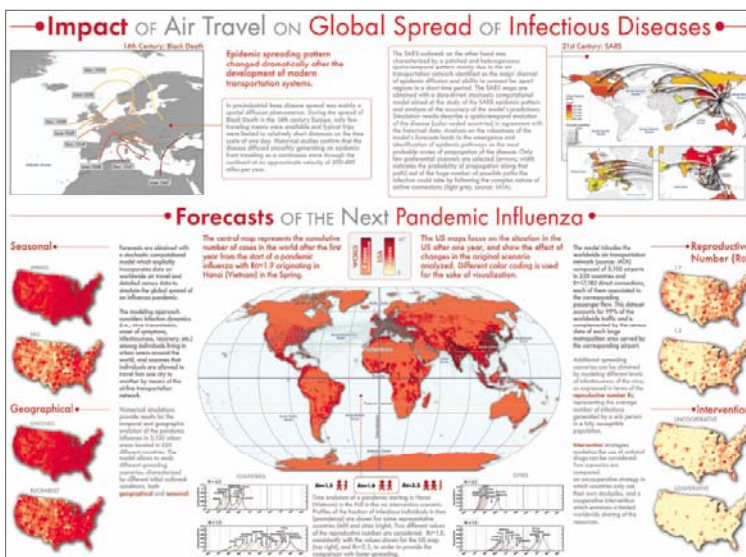
Epidemic Modeling in Complex realities, V. Colizza, A. Barrat, M. Barthelemy, A. Vespignani, Comptes Rendus Biologie, 330, 364-374 (2007).

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).



Network Workbench (<http://nwb.slis.indiana.edu>).





## Overview

- Macroscopes and the Changing Scientific Landscape 15 mins
  - Introduction to network science with sample maps and insights 15 mins
  - **Introduction to the Network Workbench Tool** 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  
  - Introduction to science studies with sample maps and insights 15 mins
  - Introduction to the Sci<sup>2</sup> Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - **Overview of validation approaches for science studies**
  
  - Plug-and-play tool design using OSGi/CIshell 30 mins
  - Scholarly Marketplaces 15 mins
- 240 mins**

13



## 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 Aug. 2009, the tool provides more 160 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

It has been downloaded more than 59,000 times since October 2006.

**Network Workbench**  
A Workbench for Network Scientists

Home People Research Publications Community Download Documentation Dev Zone About

**Summary**  
Network Workbench: A Large-Scale Network Analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research. This project will design, evaluate, and operate a unique distributed, shared resources environment for large-scale network analysis, modeling, and visualization, named Network Workbench (NWB). The envisioned data-code-computing resources environment will provide ...  
[more](#)  
[How to cite this project](#)

**News & Updates**

- 5.1.09 Kaelble, Steve. 2009. [Mapping the Future of Knowledge, Research & Creative Activity](#), 31, 2:12-15. [\(website\)](#) accessed 5/1/09
- 3.23.09 [1.0.0 beta 5](#) Released
- 1.23.09 Ann Mcranie's [tutorial abstract](#) for Sunbelt 2009
- 11.4.08 Two NWB PIs featured in "[Connected—The Power of Six Degrees](#)." 2008. Anna Maria Talas, Director. Australian Broadcasting Corporation, Ltd. [\(YouTube\)](#) [\(Full Video\)](#) (300MB)

**Download 1.0.0 beta 5 Release**  
Note: save the download as jar

Select Your Operating System  
Windows (XP & Vista) **DOWNLOAD**

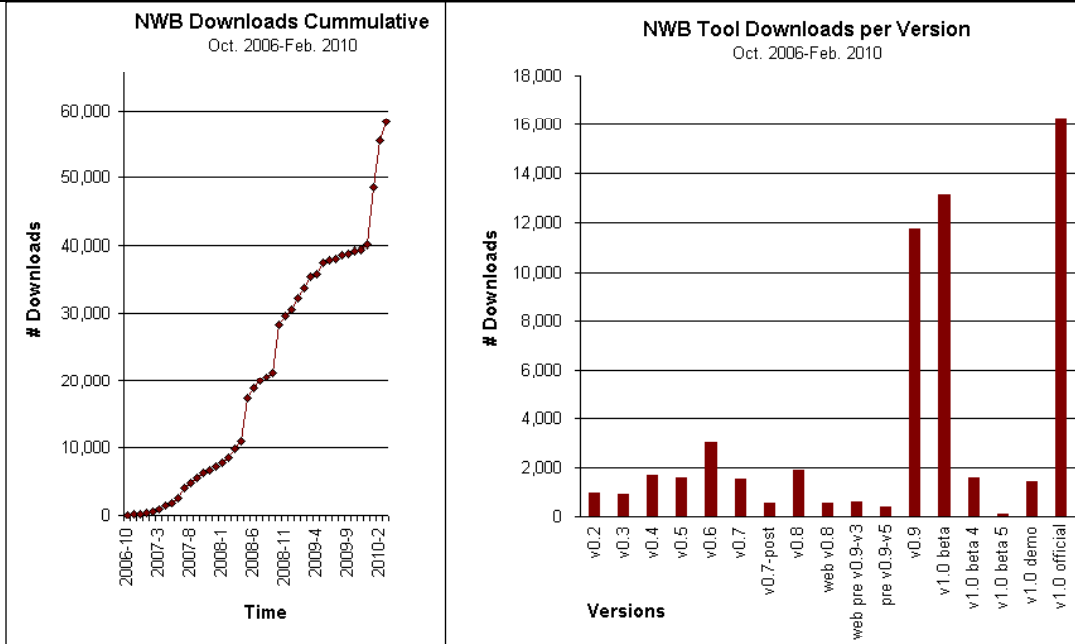
[Getting Started](#)  
See more [documentation](#)

**Get Involved**

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

14





Herr II, Bruce W., Huang, Weixia (Bonnie), Penumarthy, 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.

**NWB Community Wiki - Home Page**

NetworkWorkbench  
A Workbench for Network Scientists

**Main**  
[People](#)  
[NWB Tool](#)  
[Update Sites](#)  
[Tutorials](#)  
[Algorithms](#)  
[Datasets](#)  
[Data Formats](#)  
[Glossary](#)  
[FAQ](#)

**About the Network Workbench Community Wiki**  
 The Network Workbench Community Wiki is the part of Network Workbench (NWB) project. It provides descriptions for algorithms and datasets that have been integrated in the NWB Tool. It is also a place for users of the NWB Tool, the Cyberinfrastructure Shell, or any other CShell based program to get, upload, and request algorithms & datasets to be used in the tool. This site is a sounding board to be used by the community to work together and create a tool which will meet their needs and the needs of the scientific community at large.

Check out the lists of available [algorithms](#) and [datasets](#). Download the [NWB Tool](#) and play with it.

You are invited to add or edit your own dataset and algorithm descriptions (sign up [here](#)), or post wanted algorithms and datasets.

If you are interested in joining the NWB community, please sign up the [NWB mailing list](#), post your question there, or contact Weixia (Bonnie) Huang [huangb@indiana.edu](mailto:huangb@indiana.edu) for more information.

Recent Changes (All) | [Edit SideBar](#) | Page last modified on June 23, 2008, at 05:40 PM | [Upload files](#) | [Edit Page](#) | [Page History](#)  
 Powered by [PmWiki](#)

**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>

17

### 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/>



18

### Personal Bibliographies

- Bibtext (.bib)
- Endnote Export Format (.enw)

### Data Providers

- Web of Science by Thomson Scientific/Reuters (.isi)
- Scopus by Elsevier (.scopus)
- Google Scholar (access via *Publish or Perish* save as CSV, Bibtext, EndNote)
- Awards Search by National Science Foundation (.nsf)

### Scholarly Database (all text files are saved as .csv)

- Medline publications by National Library of Medicine
- NIH funding awards by the National Institutes of Health (NIH)
- NSF funding awards by the National Science Foundation (NSF)
- U.S. patents by the United States Patent and Trademark Office (USPTO)
- Medline papers – NIH Funding

### Network Formats

- NWB (.nwb)
- Pajek (.net)
- GraphML (.xml or .graphml)
- XGMML (.xml)

### Burst Analysis Format

- Burst (.burst)

### Other Formats

- CSV (.csv)
- Edgelist (.edge)
- Pajek (.mat)
- TreeML (.xml)

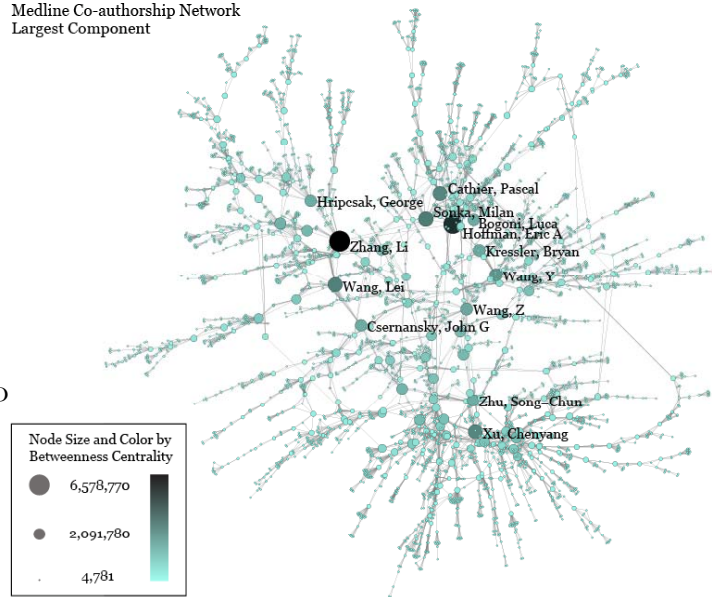
<p><b>Preprocessing</b> <small>Edit</small></p> <p><b>Remove Nodes</b></p> <ul style="list-style-type: none"> <li><a href="#">Extract Top Nodes</a></li> <li><a href="#">Extract Nodes Above or Below Val</a></li> <li><a href="#">Delete High Degree Nodes</a></li> <li><a href="#">Delete Random Nodes</a></li> <li><a href="#">Delete Isolates</a></li> </ul> <p><b>Remove Edges</b></p> <ul style="list-style-type: none"> <li><a href="#">Extract Top Edges</a></li> <li><a href="#">Extract Edges Above or Below Val</a></li> <li><a href="#">Remove Self Loops</a></li> <li><a href="#">Trim By Degree<sup>2</sup></a></li> <li><a href="#">Pathfinder Network Scaling</a></li> </ul> <p><b>Sampling</b></p> <ul style="list-style-type: none"> <li><a href="#">Snowball Sampling (n nodes)</a></li> <li><a href="#">Node Sampling</a></li> <li><a href="#">Edge Sampling</a></li> </ul> <p><b>Transformations</b></p> <ul style="list-style-type: none"> <li><a href="#">Symmetrize</a></li> <li><a href="#">Dichotomize</a></li> <li><a href="#">Multipartite Joining</a></li> </ul> <p><b>Modeling</b> <small>Edit</small></p> <p><b>General</b></p> <ul style="list-style-type: none"> <li><a href="#">Random Graph</a></li> <li><a href="#">Watts-Strogatz Small World</a></li> <li><a href="#">Barabási-Albert Scale-Free</a></li> </ul> <p><b>Structured</b></p> <ul style="list-style-type: none"> <li><a href="#">CAN</a></li> <li><a href="#">Chord</a></li> </ul> <p><b>Unstructured</b></p> <ul style="list-style-type: none"> <li><a href="#">Hypergrid</a></li> <li><a href="#">PRU</a></li> </ul> <p><b>Other</b></p> <ul style="list-style-type: none"> <li><a href="#">TARL</a></li> <li><a href="#">Discrete Network Dynamics</a></li> </ul>	<p><b>Analysis</b> <small>Edit</small></p> <p><b>General Purpose</b></p> <ul style="list-style-type: none"> <li><a href="#">Network Analysis Toolkit<sup>2</sup></a></li> </ul> <p><b>Unweighted &amp; Undirected</b></p> <ul style="list-style-type: none"> <li>Based on degree/</li> <li><a href="#">Node Degree</a></li> <li><a href="#">Node Distribution</a></li> </ul> <ul style="list-style-type: none"> <li>Based on clustering</li> <li><a href="#">k-Nearest Neighbor</a></li> <li><a href="#">Watts Strogatz Clustering Coefficient</a></li> <li><a href="#">Watts Strogatz Clustering Coefficient Over k</a></li> </ul> <ul style="list-style-type: none"> <li>Based on path</li> <li><a href="#">Diameter</a></li> <li><a href="#">Average Shortest Path</a></li> <li><a href="#">Shortest Path Distribution</a></li> <li><a href="#">Node Betweenness Centrality</a></li> </ul> <ul style="list-style-type: none"> <li>Based on components</li> <li><a href="#">Connected Components</a></li> <li><a href="#">Weak Component Clustering</a></li> </ul> <ul style="list-style-type: none"> <li><b>K-Core</b></li> <li><a href="#">Extract K-Core<sup>2</sup></a></li> <li><a href="#">Annotate K-Core<sup>2</sup></a></li> </ul> <p><b>Unweighted &amp; Directed</b></p> <ul style="list-style-type: none"> <li>Based on degree</li> <li><a href="#">Node Indegree</a></li> <li><a href="#">Node Outdegree</a></li> <li><a href="#">Indegree Distribution</a></li> <li><a href="#">Outdegree Distribution</a></li> </ul> <ul style="list-style-type: none"> <li>Based on local graph structure</li> <li><a href="#">k-Nearest Neighbor</a></li> <li><a href="#">Single Node In-Out Degree Correlations<sup>2</sup></a></li> </ul> <ul style="list-style-type: none"> <li>Unnamed Category?</li> <li><a href="#">Page Rank</a></li> </ul> <ul style="list-style-type: none"> <li>Based on local graph structure #2</li> <li><a href="#">Dyad Reciprocity<sup>2</sup></a></li> <li><a href="#">Arc Reciprocity<sup>2</sup></a></li> </ul>	<p><b>tion</b> <small>Edit</small></p> <p><b>Tools</b></p> <ul style="list-style-type: none"> <li><a href="#">GUESS</a></li> <li><a href="#">GnuPlot<sup>2</sup></a></li> </ul> <p><b>Predefined Positions Layout</b></p> <ul style="list-style-type: none"> <li><a href="#">DrL (VxOrd)</a></li> <li><a href="#">Pre-defined Positions (prefuse beta)<sup>2</sup></a></li> </ul> <p><b>Move</b></p> <ul style="list-style-type: none"> <li><a href="#">Circular</a></li> </ul> <p><b>Tree Layouts</b></p> <ul style="list-style-type: none"> <li><a href="#">Radial Tree (prefuse alpha)</a></li> <li><a href="#">Radial Tree with Annotations (prefuse beta)<sup>2</sup></a></li> <li><a href="#">Tree Map</a></li> <li><a href="#">Tree View</a></li> <li><a href="#">Balloon Graph (prefuse alpha)<sup>2</sup></a></li> </ul> <p><b>Network Layouts</b></p> <ul style="list-style-type: none"> <li><a href="#">Force Directed with Annotation (prefuse beta)</a></li> <li><a href="#">Kamada-Kawai (JUNG)</a></li> <li><a href="#">Fruchterman-Reingold (JUNG)</a></li> <li><a href="#">Fruchterman-Reingold with Annotation (prefuse beta)</a></li> <li><a href="#">Spring (JUNG)</a></li> <li><a href="#">Small World (prefuse alpha)</a></li> </ul> <p><b>Other Layouts</b></p> <ul style="list-style-type: none"> <li><a href="#">Parallel Coordinates (demo)<sup>2</sup></a></li> <li><a href="#">LaNet (k-Core Decomposition)</a></li> </ul> <p><b>etrics</b> <small>Edit</small></p> <p><b>Extract Network From Table</b></p> <ul style="list-style-type: none"> <li><a href="#">Extract Co-Authorship Network</a></li> <li><a href="#">Extract Co-Occurrence Network From Table<sup>2</sup></a></li> <li><a href="#">Extract Directed Network From Table<sup>2</sup></a></li> </ul> <p><b>Extract Network From Another Network</b></p> <ul style="list-style-type: none"> <li><a href="#">Extract Bibliographic Coupling Similarity Network</a></li> <li><a href="#">Extract Co-Citation Similarity Network<sup>2</sup></a></li> </ul> <p><b>Cleaning</b></p> <ul style="list-style-type: none"> <li><a href="#">Remove ISI Duplicate Records</a></li> </ul>
---	--	--



- NWB tool can be used for data conversion. Supported output formats comprise:
  - CSV (.csv)
  - NWB (.nwb)
  - Pajek (.net)
  - Pajek (.mat)
  - GraphML (.xml or .graphml)
  - XGMML (.xml)
- GUESS
 

Supports export of images into common image file formats.
- Horizontal Bar Graphs
- saves out raster and ps files.

Medline Co-authorship Network  
Largest Component



21



### Overview

- Macroscopes and the Changing Scientific Landscape 15 mins
  - Introduction to network science with sample maps and insights 15 mins
  - Introduction to the Network Workbench Tool 15 mins
  - **Demo and hands-on data analysis and visualization by participants** 60 mins
  - Introduction to science studies with sample maps and insights 15 mins
  - Introduction to the Sci<sup>2</sup> Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - **Overview of validation approaches for science studies**
  - Plug-and-play tool design using OSGi/CIshell 30 mins
  - Scholarly Marketplaces 15 mins
- 240 mins**

22

## NWB Tool Demo

- A. Download, install, and run.
- B. Load, view, convert, save data.
- C. Read and visualize a directory hierarchy.
- D. Load a network, compute its basic properties, and explore it in GUESS.
- E. Advanced community detection and scalable visualizations.

23



### Software on DVD

Name	Size	Type
<b>I - P (6)</b>		
NWB-Linux32bit		File Folder
NWB-Linux64bit		File Folder
NWB-MacG3G4G5		File Folder
NWB-MacIntel		File Folder
NWB-Windows		File Folder
NWB_Tutorial.pdf	4,046 KB	Adobe Acrobat Document
<b>Q - Z (7)</b>		
Sci2-Linux32bit		File Folder
Sci2-Linux64bit		File Folder
Sci2-MacG3G4G5		File Folder
Sci2-MacIntel		File Folder
Sci2-Windows		File Folder
README.txt	1 KB	Text Document
Sci2_Tutorial.pdf	10,947 KB	Adobe Acrobat Document

24



## Download, Install, and Run (Demo DVD has all installers)

### 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.

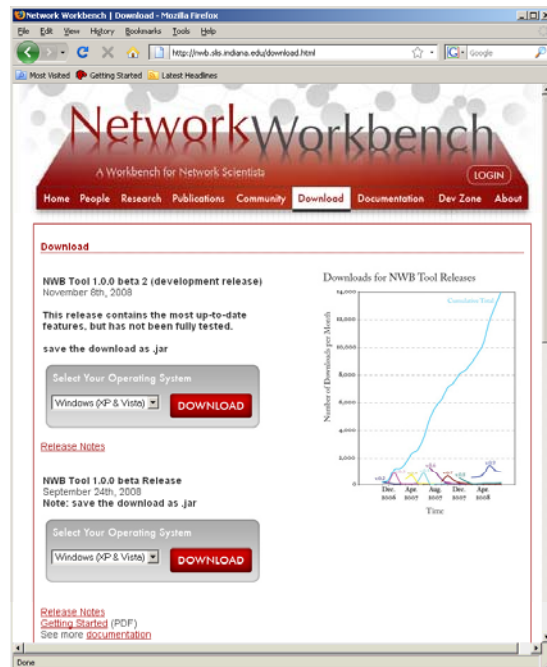
Save as \*.jar file.

Install and run.

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

NWB Demo DVD has all installers.

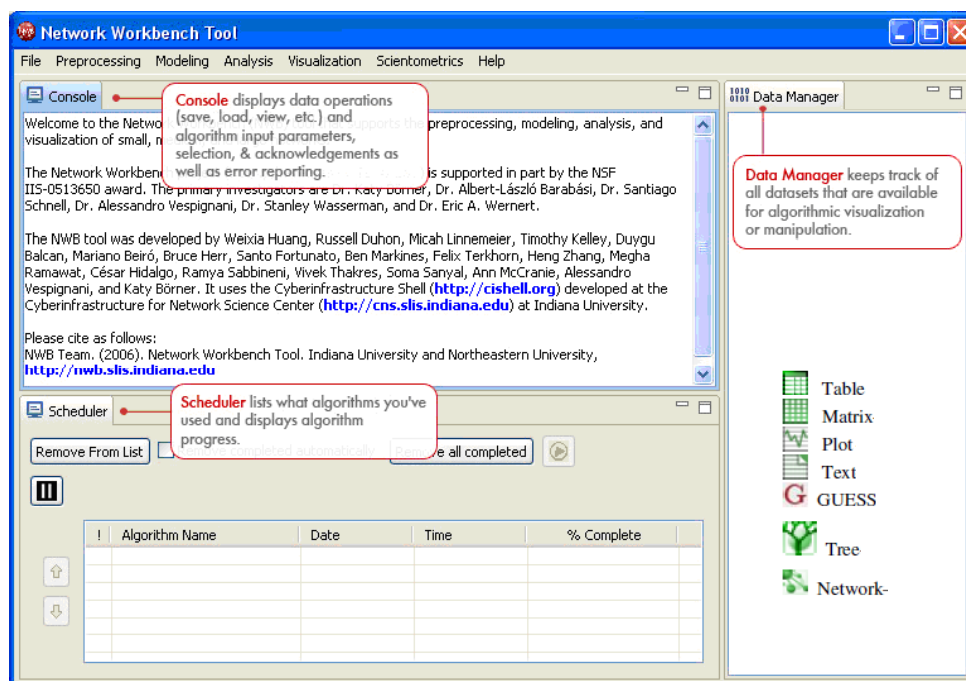
- NWB-Linux32bit
- NWB-Linux64bit
- NWB-MacG3G4G5
- NWB-MacIntel
- NWB-Windows



25



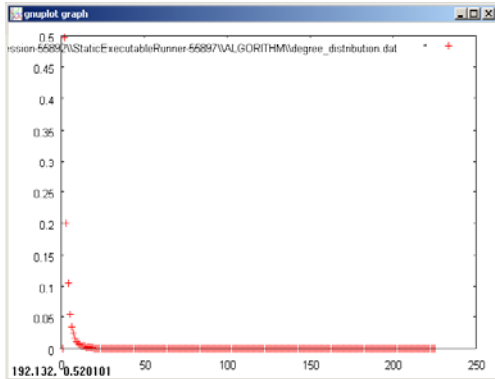
## NWB Tool Interface Components



26

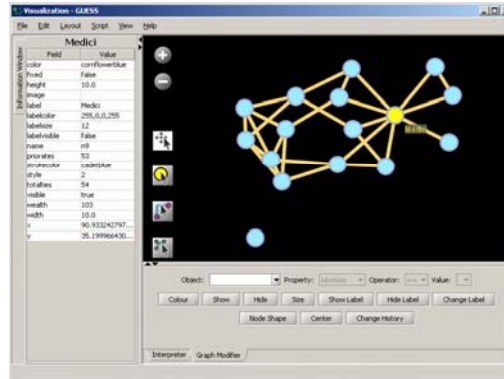
File	Preprocessing	Modeling	Visualization
Load...	Extract Top Nodes	Random Graph	GUESS
Load and Clean ISI File	Extract Nodes Above or Below Value	Watts-Strogatz Small World	Graphlet
Read Directory Hierarchy Datasets	Remove Node Attributes	Barrabási-Albert Scale-Free	DrL (VxOrd)
Save...	Delete High Degree Nodes	Can	Specified (prefuse beta)
View...	Delete Random Nodes	Chord	Circular (JUNG)
View with...	Delete Isolates	Hypergrid	Radial Tree/Graph (prefuse alpha)
Merge Node and Edge Files	Extract Top Edges	PRU	Radial Tree/Graph with Annotation (prefuse beta)
Split Graph to Node and Edge Files	Extract Edges Above or Below Value	TARL	Tree Map (prefuse beta)
Tests	Remove Edge Attributes	Discrete Network Dynamics (DND)	TreeView (prefuse beta)
Preferences	Remove Self Loops	Evolving Network (Weighted)	Balloon Graph (prefuse alpha)
Exit	Trim by Degree		Force Directed with Annotation (prefuse beta)
	Snowball Sampling (n nodes)		Kamada-Kawai (JUNG)
	Node Sampling		Fruchterman-Reingold (JUNG)
	Edge Sampling		Fruchterman-Reingold with Annotation (prefuse beta)
	Symmetrize		Spring (JUNG)
	Dichotomize		Small World (prefuse alpha)
	Multipartite Joining		Parallel Coordinates (demo)
	Normalize Text		LaNet
	Slice Table by Time		Circular Hierarchy

Analysis	Unweighted and Undirected	Unweighted and Directed	Weighted and Undirected	Weighted and Directed	Search	Textual	Discrete Network Dynamics
Network Analysis Toolkit (NAT)	Node Degree	Node Indegree	Clustering Coefficient	HITS	Can	Burst Detection	Extract and Annotate Attractors
Unweighted and Undirected	Degree Distribution	Node Outdegree	Nearest Neighbor Degree	Weak Component Clustering	Chord		
Weighted and Undirected	Watts-Strogatz Clustering Coefficient	Indegree Distribution	Strength vs Degree	Blondel Community Detection	K Random-Walk		
Unweighted and Directed	Watts Strogatz Clustering Coefficient over K	Outdegree Distribution	Degree & Strength	MST-Pathfinder Network Scaling	Random Breadth First		
Weighted and Directed	Diameter	K-Nearest Neighbor	Average Weight vs End-point Degree	Extract K-Core			
Search	Average Shortest Path	Single Node In-Out Degree Correlations	K-Nearest Neighbor (Java)	Annotate K-Core			
Discrete Network Dynamics	Shortest Path Distribution	PageRank	Strength Distribution				
Textual	Node Betweenness Centrality	HITS	Weight Distribution				
	Global Connected Components	Dyad Reciprocity	Randomize Weights				
	HITS	Arc Reciprocity	MST-Pathfinder Network Scaling				
	Weak Component Clustering	Adjacency Transitivity	Fast Pathfinder Network Scaling				
	Blondel Community Detection	Agency Transitivity	Blondel Community Detection				
	MST-Pathfinder Network Scaling	Weak Component Clustering	Extract K-Core				
	Extract K-Core	Strong Component Clustering	Annotate K-Core				
	Annotate K-Core	Blondel Community Detection					
		Extract K-Core					
		Annotate K-Core					



## 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>.

In November 2008, the NWB tool supports loading the following input file formats:

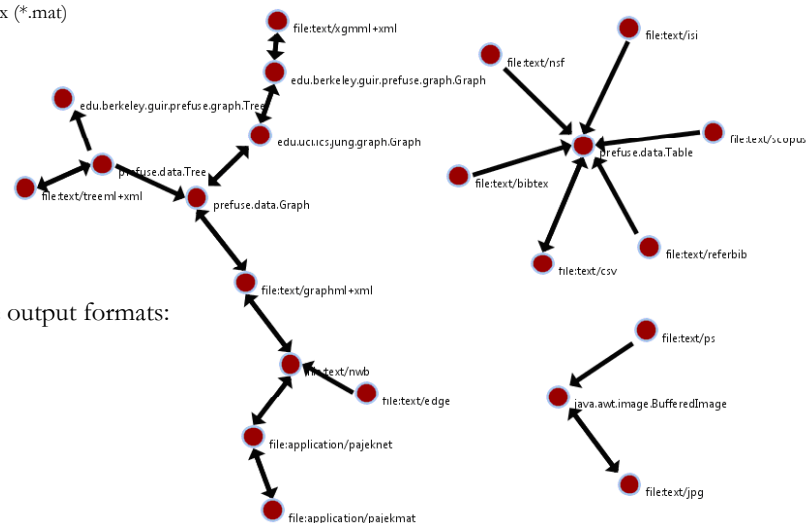
- GraphML (\*.xml or \*.graphml)
- XGMML (\*.xml)
- Pajek .NET (\*.net) & Pajek .Matrix (\*.mat)
- NWB (\*.nwb)
- TreeML (\*.xml)
- Edge list (\*.edge)
- CSV (\*.csv)
- ISI (\*.isi)
- Scopus (\*.scopus)
- NSF (\*.nsf)
- Bibtext (\*.bib)
- Endnote (\*.enw)

and the following network file output formats:

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

These formats are documented at

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



The ‘*\*yournwbdirectory\*/sampledata*’ directory provides sample datasets from the biology, network, scientometrics, and social science research domains:

**/biology**

**/network**

**/scientometrics**

/scientometrics/bibtex

/scientometrics/csv

/scientometrics/endnote

/scientometrics/isi

➤ FourNetSciResearchers.isi

/scientometrics/nsf

➤ Cornell.nsf

➤ Indiana.nsf

➤ Michigan.nsf

/scientometrics/scopus

**/socialscience**

➤ florentine.nwb

The ‘*\*yournwbdirectory\*/*’ directory also contains

*/sampledata/scientometrics/properties* // Used to extract networks and merge data

- bibtexCoAuthorship.properties
- endnoteCoAuthorship.properties
- isiCoAuthorship.properties
- isiCoCitation.properties
- isiPaperCitation.properties
- mergeBibtexAuthors.properties
- mergeEndnoteAuthors.properties
- mergeIsiAuthors.properties
- mergeNsfPIs.properties
- mergeScopusAuthors.properties
- nsfCoPI.properties
- scopusCoAuthorship.properties

*/sampledata/scripts/GUESS* // Used to do color/size/shape code networks

- co-author-nw.py
- co-PI-nw.py
- paper-citation-nw.py
- reference-co-occurrence-nw.py



## NWB Tool Overview

- A. Download, install, and run.
- B. Load, view, convert, save data.
- C. Read and visualize a directory hierarchy.
- D. Load a network, compute its basic properties, and explore it in GUESS.
- E. Advanced community detection and scalable visualizations.

33

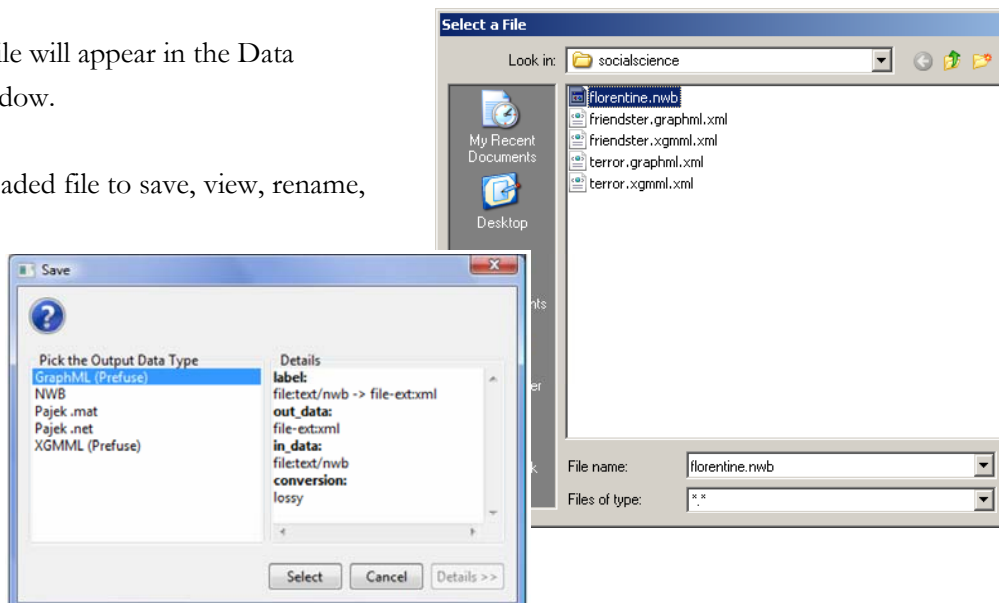


### Load, View and Save (Convert) Data

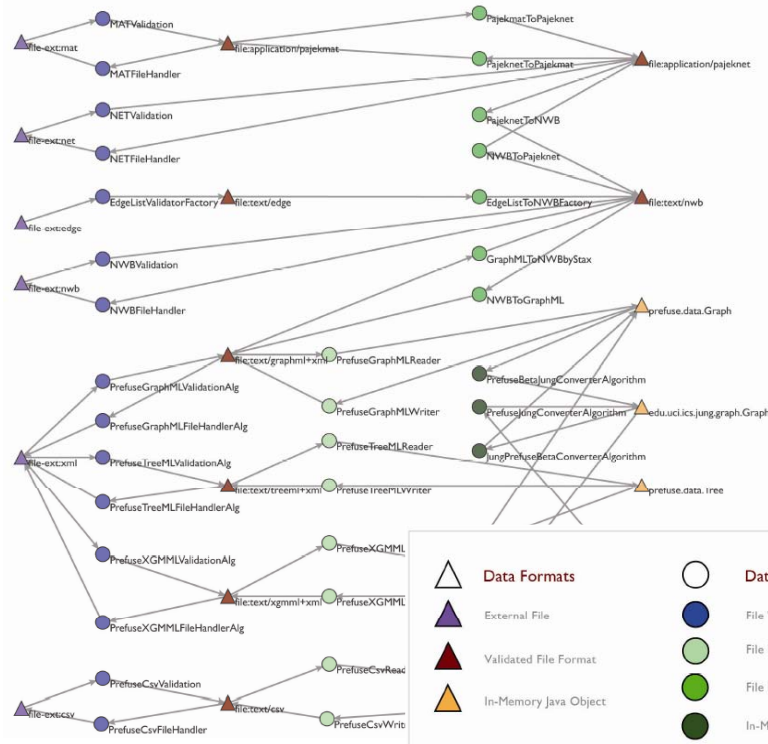
Use 'File > Load File' to load *florentine.nwb* in sample datasets in *'\*yournwbdirectory\*/sampledata/socialscience'*.

The loaded file will appear in the Data Manager window.

Right click loaded file to save, view, rename, or discard.



34



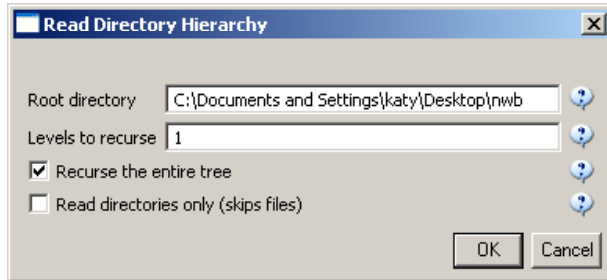
There is no central data format.

Instead, data formats used in different communities and required by the different algorithms are supported.

## NWB Tool Overview

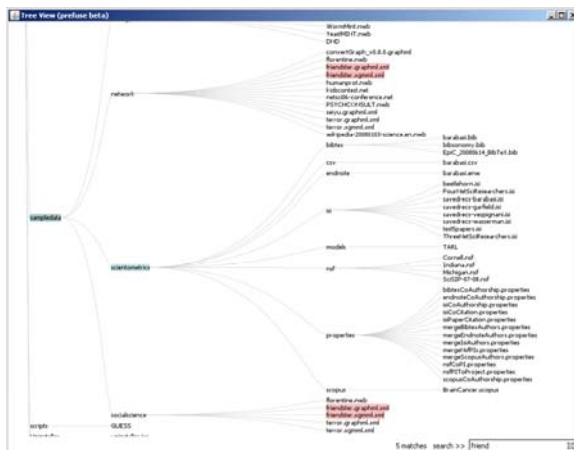
- A. Download, install, and run.
- B. Load, view, convert, save data.
- C. **Read and visualize a directory hierarchy.**
- D. Load a network, compute its basic properties, and explore it in GUESS.
- E. Advanced community detection and scalable visualizations.

Use *File > Read Directory Hierarchy* with parameters



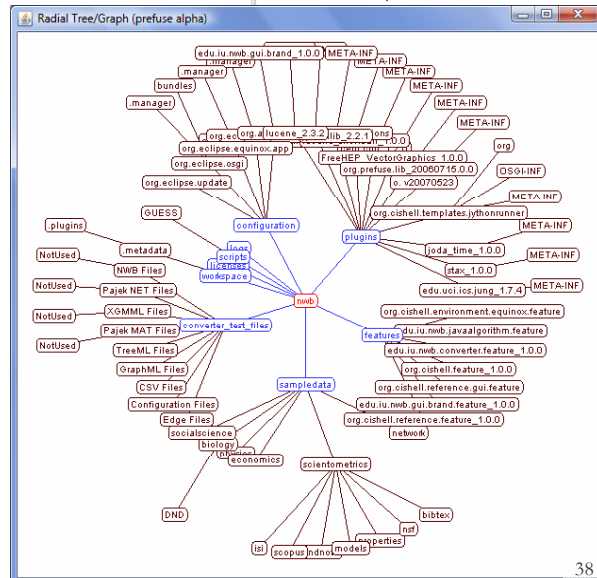
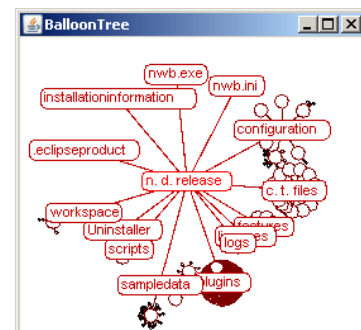
Visualize resulting *Directory Tree - Prefuse (Beta) Graph* using

- *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 /nwb directory hierarchy.

Note the size of the /plugin directory.



## NWB Tool Overview

- A. Download, install, and run.
- B. Load, view, convert, save data.
- C. Read and visualize a directory hierarchy.
- D. Load a network, compute its basic properties, and explore it in GUESS.
- E. Advanced community detection and scalable visualizations.

39



### Compute Basic NW Properties & View in GUESS

Select *florentine.mwb* in Data Manager.

- 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, totalties, priorates

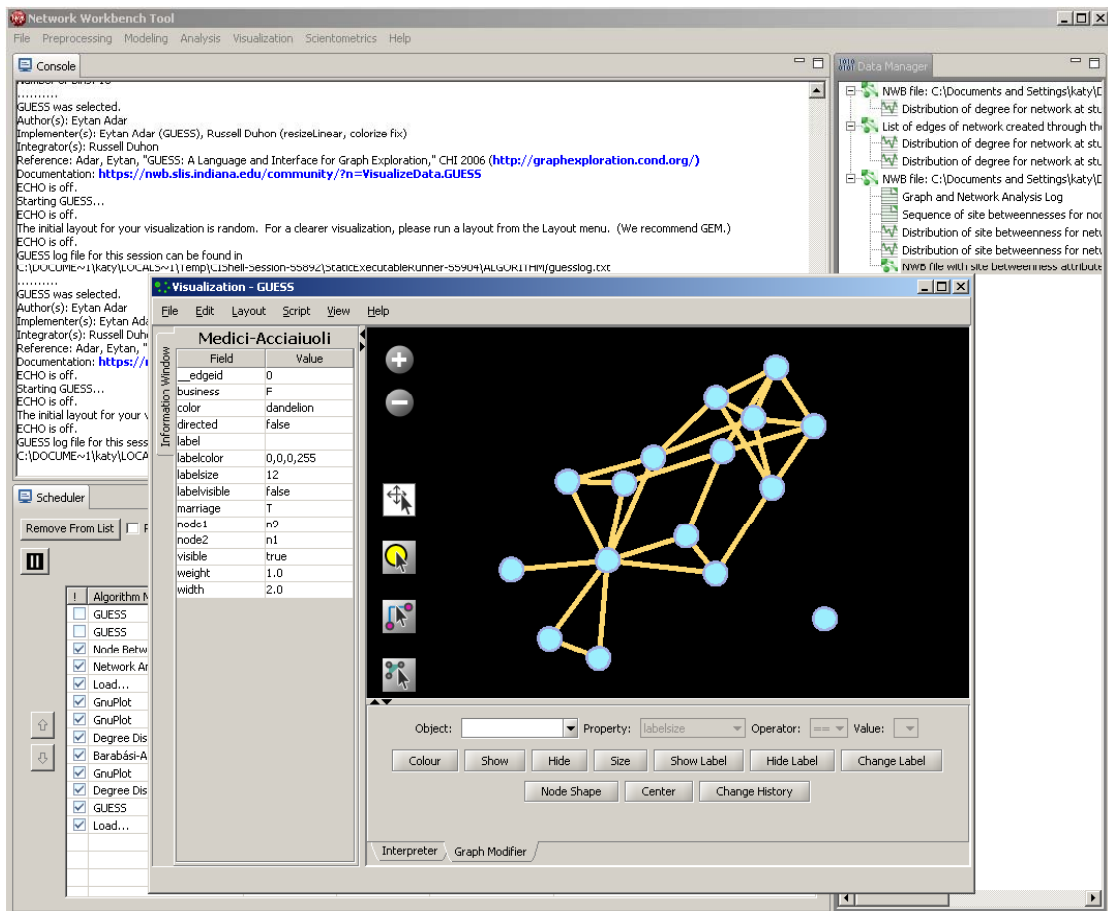
Edges: 27
No self loops were discovered.
No parallel edges were discovered.
Edge attributes:
  Nonnumeric attributes:
    marriag...   T   Example value
    busines...   F
  Did not detect any numeric attributes
  This network does not seem to be a valued network.

Average degree: 3.375
This graph is not weakly connected.
There are 2 weakly connected components. (1 isolates)
The largest connected component consists of 15 nodes.
Did not calculate strong connectedness because this graph was not directed.

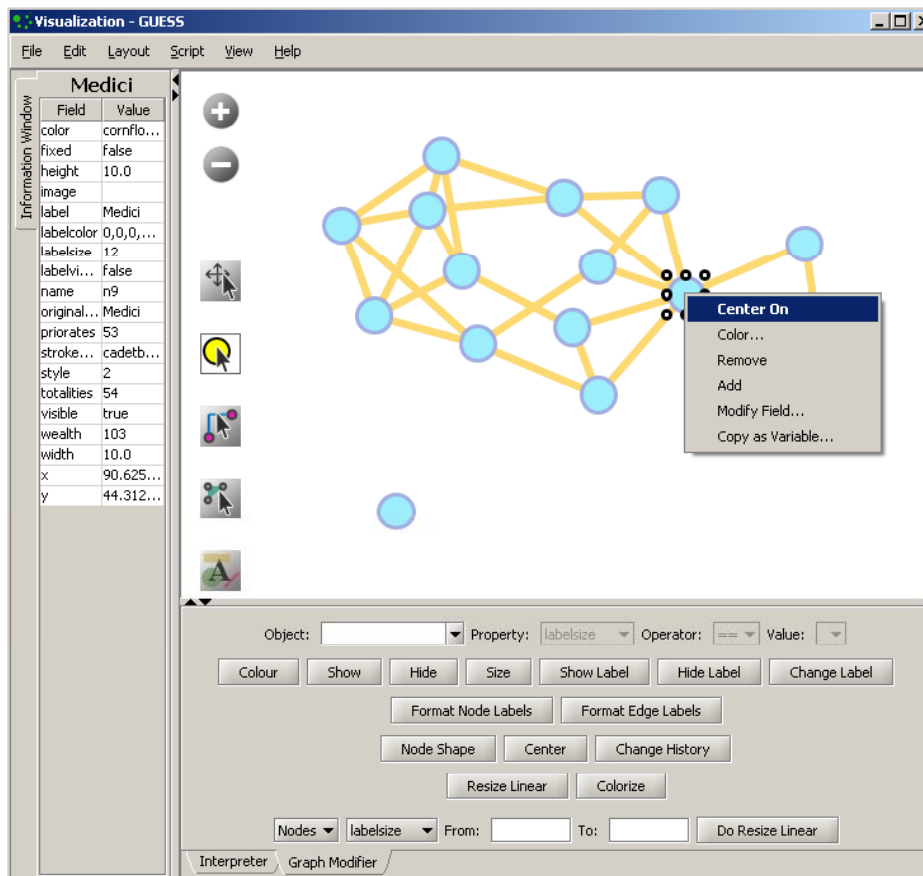
Density (disregarding weights): 0.225
```

- 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'.

40




41



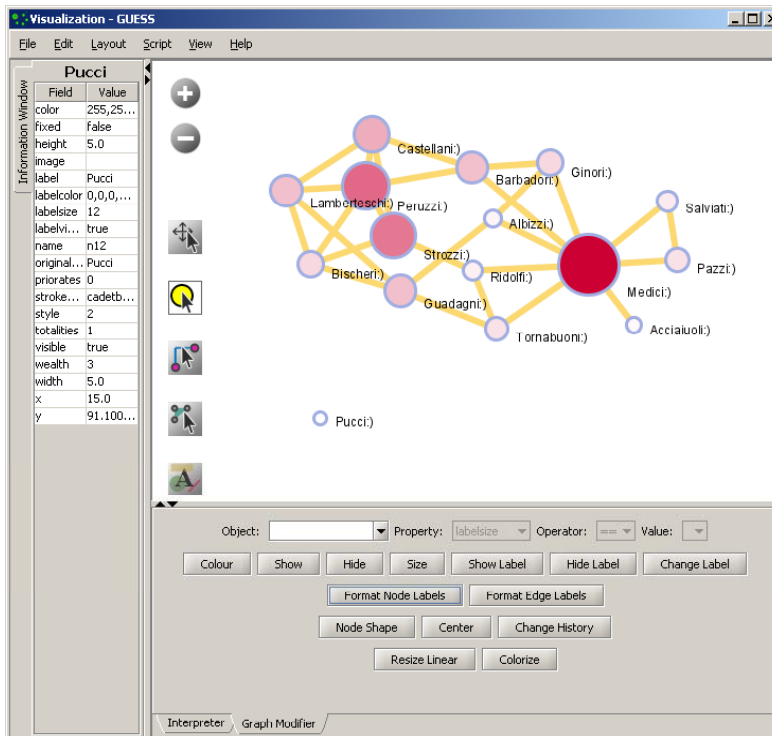
Pan:  
"grab" the background by holding left-click and moving your mouse.

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.

42



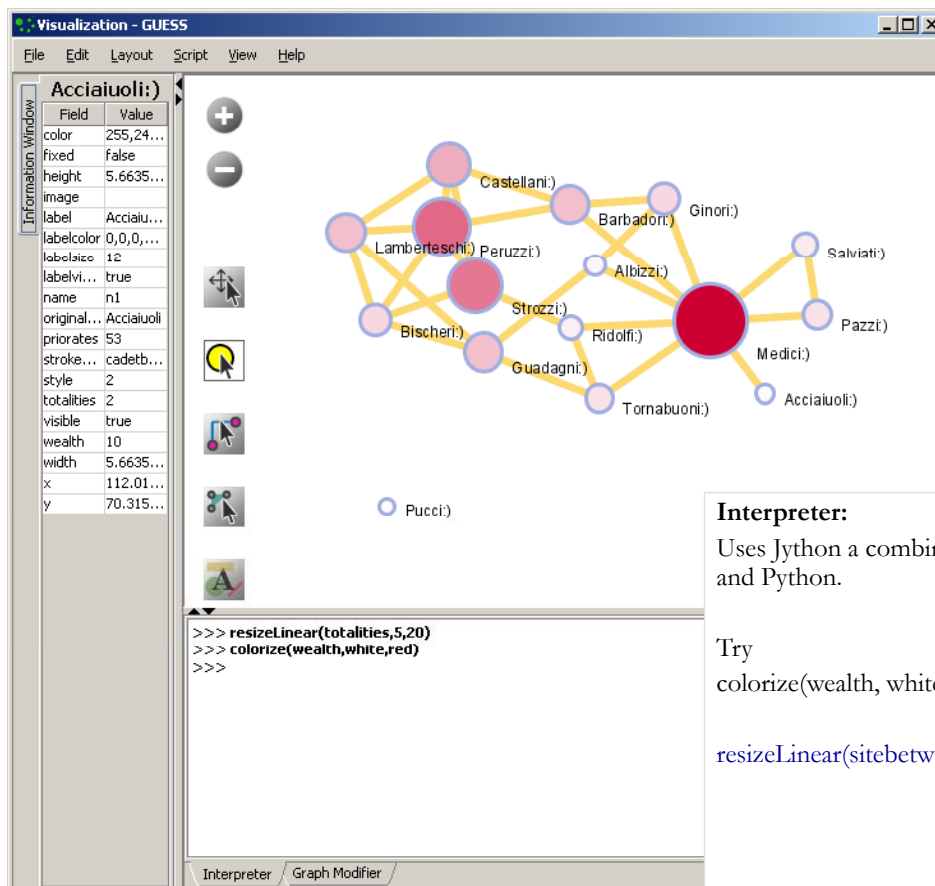
### Graph Modifier:

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.’



### Interpreter:

Uses Jython a combination of Java and Python.

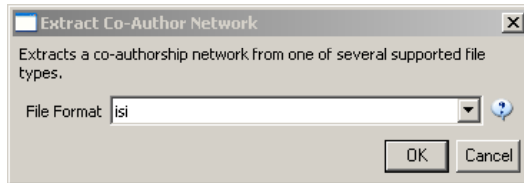
Try  
colorize(wealth, white, red)

resizeLinear(sitebetweenness, 5, 25)



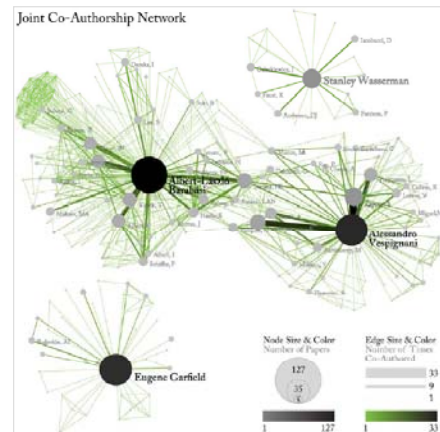
Load *\*yournwbdirectory\*/sampledata/scientometrics/isi/FourNetSciResearchers.isi*  
 using 'File > Load and Clean ISI File'.

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



The result is an undirected network of co-authors  
 in the Data Manager. It has 247 nodes and 891 edges.

To view the complete network, select the network  
 and run 'Visualization >  
 GUESS > GEM'. Run *Script > Run Script...*  
 And select *Script folder > GUESS > co-author-nw.py*.



45

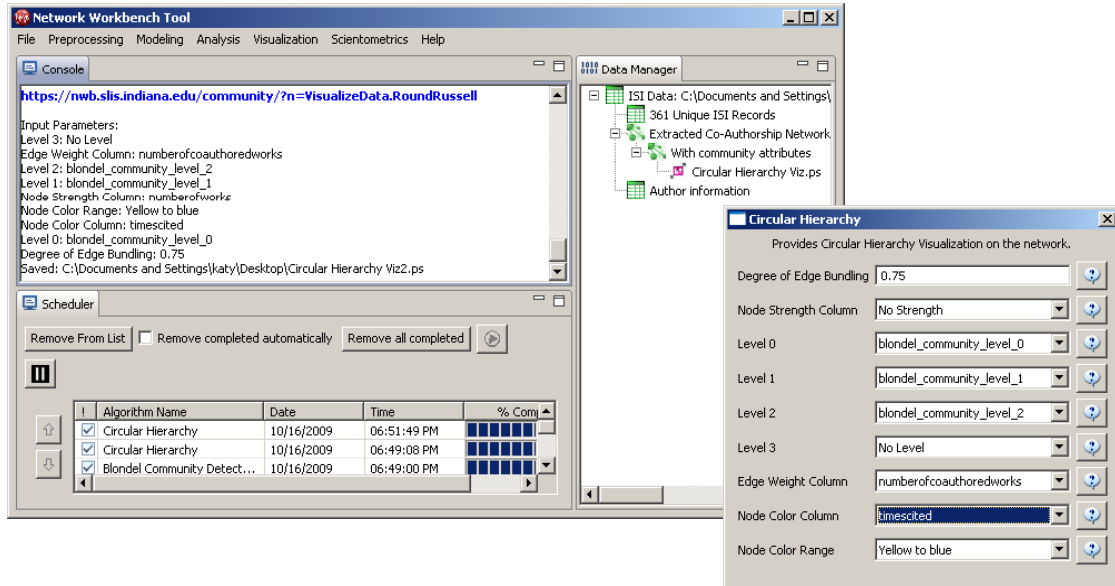
## NWB Tool Overview

- A. Download, install, and run.
- B. Load, view, convert, save data.
- C. Read and visualize a directory hierarchy.
- D. Load a network, compute its basic properties, and explore it in GUESS.
- E. **Advanced community detection and scalable visualizations.**

46

To cluster a network into subnetworks hierarchically use the Blondel community detection algorithms running *'Analysis > Weighted and Undirected > Blondel Community Detection'* using *numberofcoauthoredworks*.

Visualize result with *'Visualization > Circular Hierarchy'*.

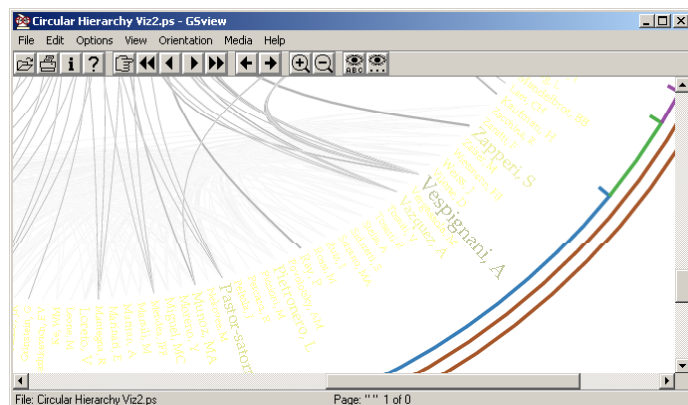


47

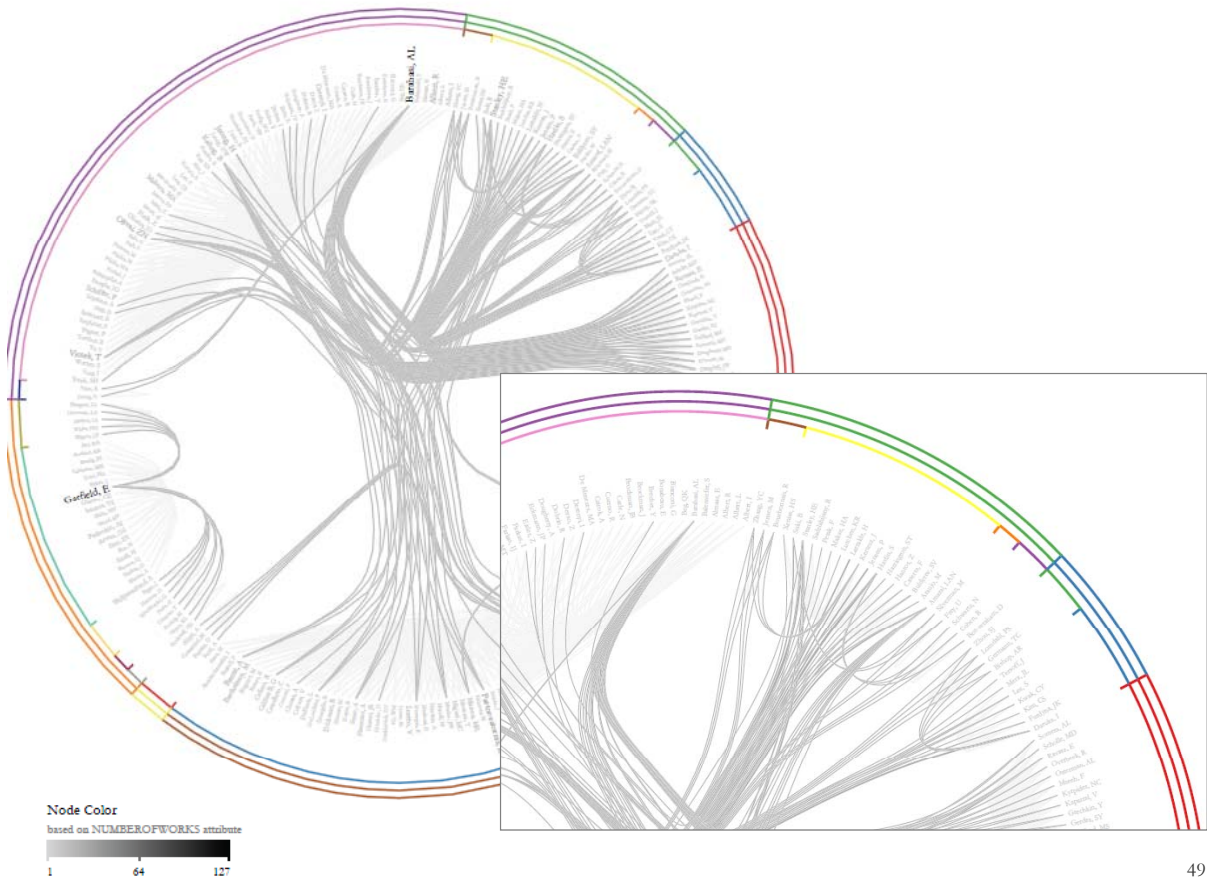
GSview is a graphical interface for Ghostscript under MS-Windows, OS/2 and GNU/Linux. Ghostscript is an interpreter for the PostScript page description language used by laser printers. For documents following the Adobe PostScript Document Structuring Conventions, GSview allows selected pages to be viewed or printed. GSview 4.9 requires Ghostscript 7.04 - 9.99.

On Sci2 Tool DVD run  
/PSView/gsview/gsview32.exe

File > Open



48



## Overview

- Macroscopes and the Changing Scientific Landscape 15 mins
  - Introduction to network science with sample maps and insights 15 mins
  - Introduction to the Network Workbench Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - **Introduction to science studies with sample maps and insights** 15 mins
  - Introduction to the Sci<sup>2</sup> Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - **Overview of validation approaches for science studies**
  - Plug-and-play tool design using OSGi/CIshell 30 mins
  - Scholarly Marketplaces 15 mins
- 240 mins**



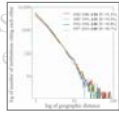

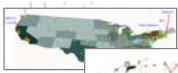


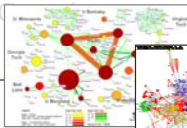
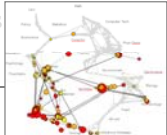
## Type of Analysis vs. Scale of Level of Analysis

	<i>Micro/Individual</i> (1-100 records)	<i>Meso/Local</i> (101–10,000 records)	<i>Macro/Global</i> (10,000 < records)
<i>Statistical Analysis/Profiling</i>	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of USA, all of science.
<i>Temporal Analysis (When)</i>	Funding portfolio of one individual	Mapping topic bursts in 20-years of PNAS	113 Years of physics Research
<i>Geospatial Analysis (Where)</i>	Career trajectory of one individual	Mapping a states intellectual landscape	PNAS publications
<i>Topical Analysis (What)</i>	Base knowledge from which one grant draws.	Knowledge flows in Chemistry research	VxOrd/Topic maps of NIH funding
<i>Network Analysis (With Whom?)</i>	NSF Co-PI network of one individual	Co-author network	NSF's core competency

51



## Type of Analysis vs. Scale of Level of Analysis

	<i>Micro/Individual</i> (1-100 records)	<i>Meso/Local</i> (101–10,000 records)	<i>Macro/Global</i> (10,000 < records)
<i>Statistical Analysis/Profiling</i>	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of USA, all of science. 
<i>Temporal Analysis (When)</i>	Funding portfolio of one individual	Mapping topic bursts in 20-years of PNAS	113 Years of physics Research 
<i>Geospatial Analysis (Where)</i>	Career trajectory of one individual	Mapping a states intellectual landscape	PNAS publications 
<i>Topical Analysis (What)</i>	Base knowledge from which one grant draws.	Knowledge flows in Chemistry research	VxOrd/Topic maps of NIH funding 
<i>Network Analysis (With Whom?)</i>	NSF Co-PI network of one individual 	Co-author network 	NSF's core competency 

Common analysis types are

- Temporal
- Geospatial
- Topical
- Network

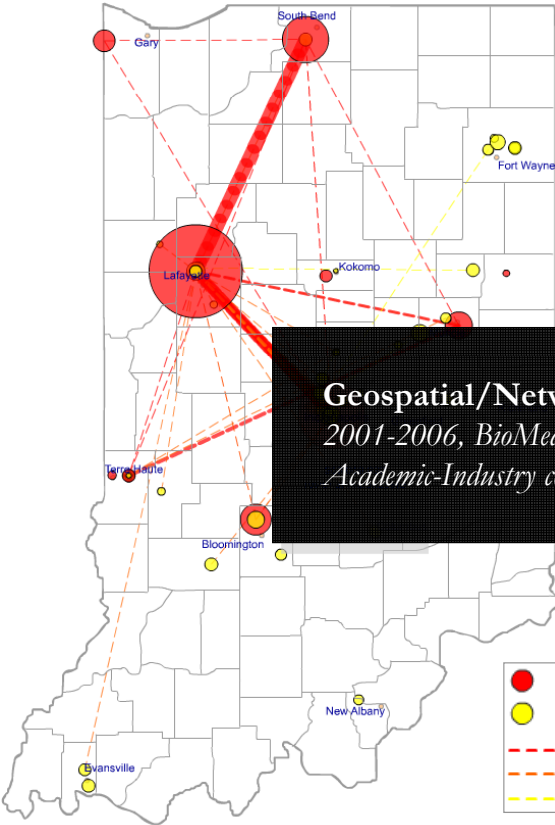
or combinations thereof.

*The data used determines the scope of the analysis.*

*We also list the main analysis goal.*

52

Mapping Indiana's Intellectual Space



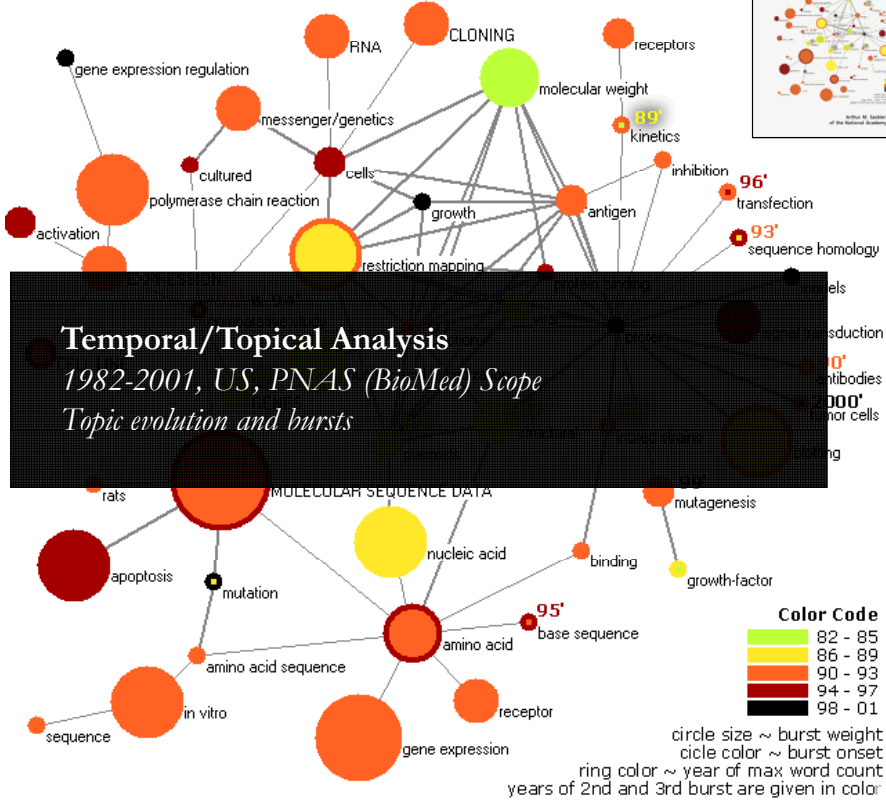
**Geospatial/Network Analysis**  
 2001-2006, BioMed, IN Scope  
*Academic-Industry collaborations and knowledge diffusion*

● (Red)	Academic
● (Yellow)	Industry
--- (Red)	Academic vs. Academic
--- (Orange)	Academic vs. Industry
--- (Yellow)	Industry vs. Industry

Mapping Topic Bursts

Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.

Mane & Börner. (2004) PNAS, 101(Suppl. 1): 5287-5290.





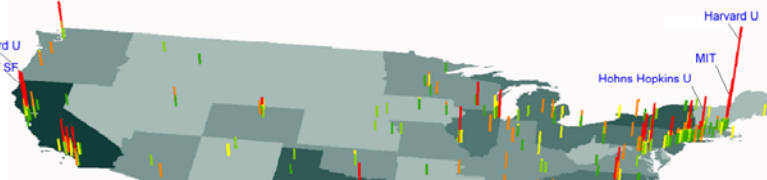
## Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions

Börner, Katy, Penumarthu, Shashikant, Meiss, Mark and Ke, Weimao. (2006)  
*Mapping the Diffusion of Scholarly Knowledge Among Major U.S. Research Institutions. Scientometrics. 68(3), pp. 415-426*

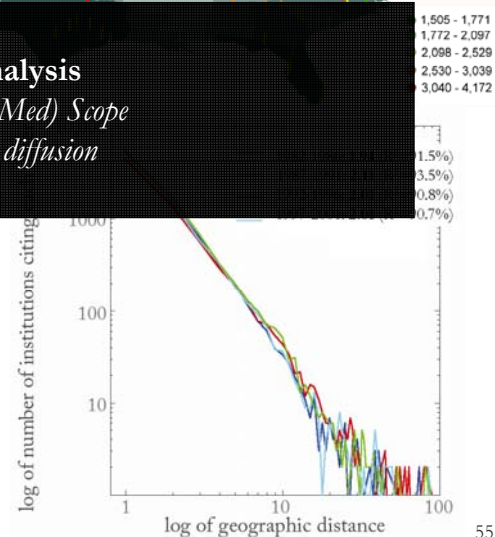


### Research questions:

1. Does space still matter in the Internet age?
2. Does one still have to study and work at institutions in order to do high quality data quality research?
3. Does the Internet change patterns, i.e., more produced at geographically distant research institutions?



### Temporal/Geospatial Analysis 1982-2001, US, PNAS (BioMed) Scope Citation impact and knowledge diffusion



### Contributions:

- Answer to Qs 1 + 2 is YES.
- Answer to Qs 3 is NO.
- Novel approach to analyzing the dual role of institutions as information producers and consumers and to study and visualize the diffusion of information among them.

## Research Collaborations by the Chinese Academy of Sciences

By Weixia (Bonnie) Huang, Russell J. Dubon, Elisha F. Hardy, Katy Börner, Indiana University, USA



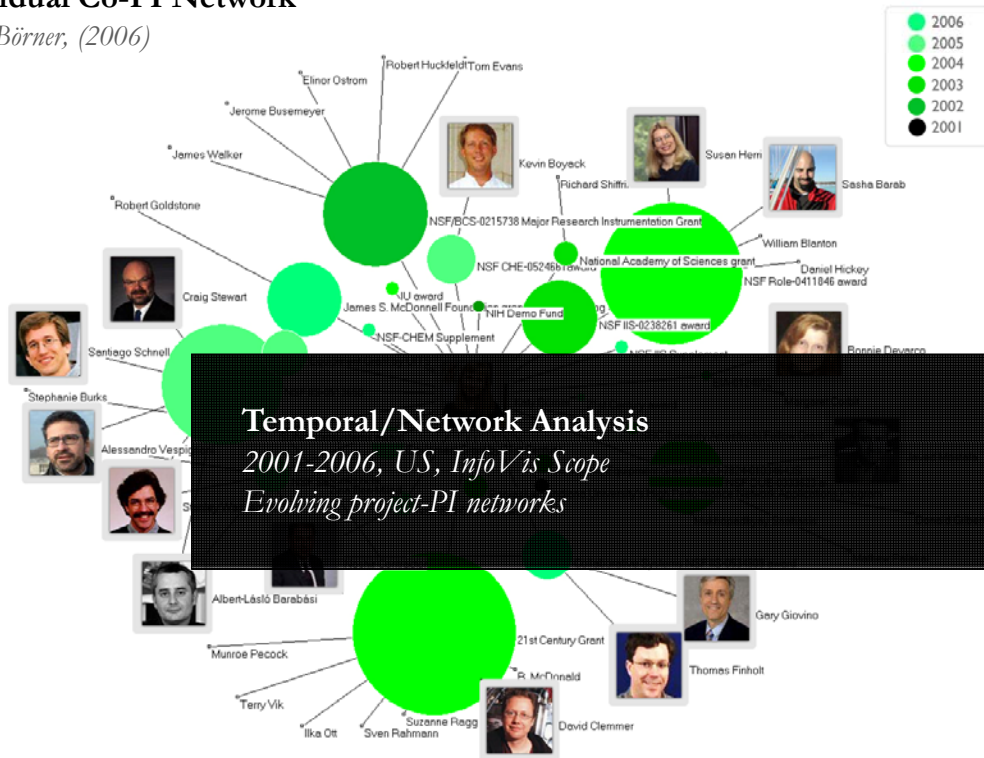
### Geospatial Analysis World, Chinese Academy of Science Collaboration and knowledge diffusion via co-author networks

This map highlights the research collaborations of the Chinese Academy of Sciences with locations in China and countries around the world. The large geographic map shows the research collaborations of all CAS institutes. Each smaller geographic map shows the research collaborations by the CAS researchers in one province-level administrative division. Collaborations between CAS researchers are not included in the data. On each map, locations are colored on a logarithmic scale by the number of collaborations from red to yellow. The darkest red is 3,395 collaborations by all of CAS with researchers in Beijing. Also, flow lines are drawn from the location of focus to all locations collaborated with. The width of the flow line is linearly proportional to the number of collaborations with the locations it goes to, with the smallest flow lines representing one collaboration and the largest representing differing amounts on each geographic map.



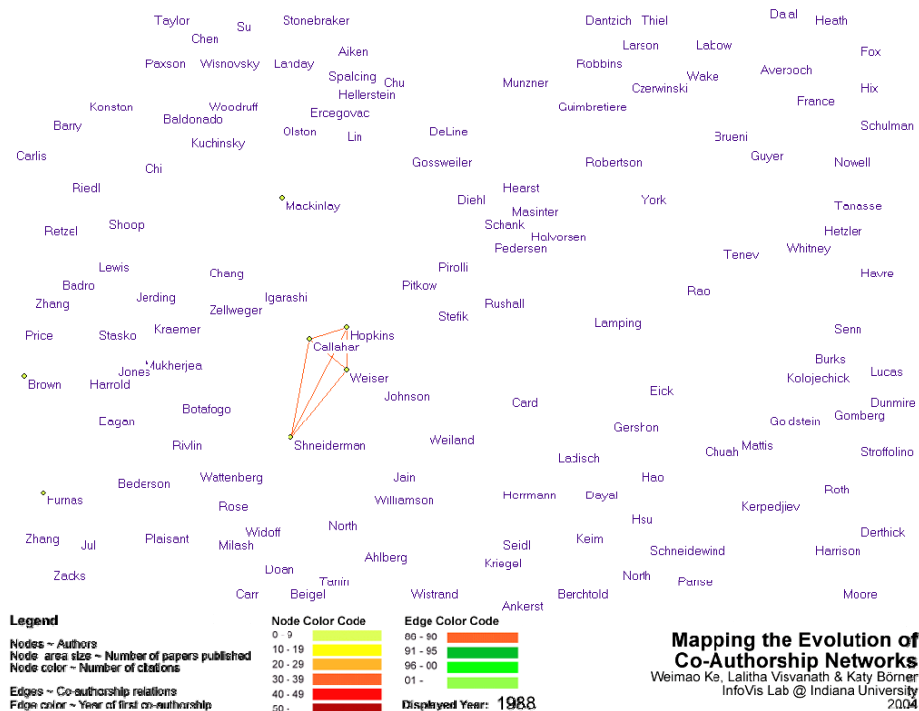
# Individual Co-PI Network

Ke & Börner, (2006)



# Mapping the Evolution of Co-Authorship Networks

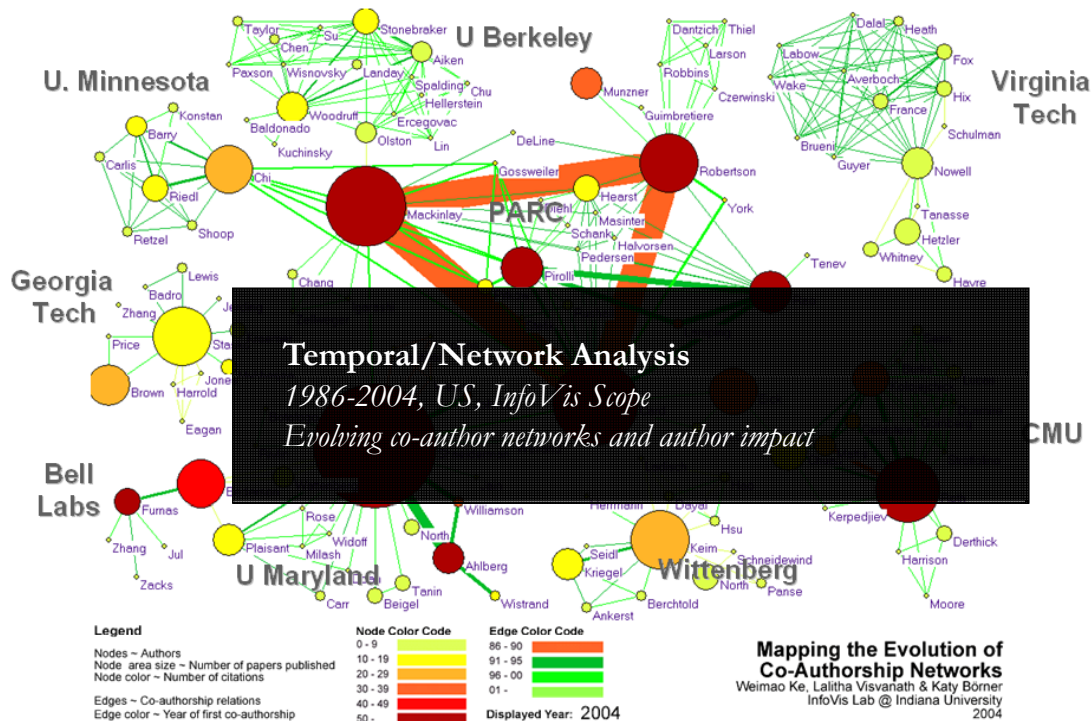
Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest.



**Mapping the Evolution of Co-Authorship Networks**  
Weimao Ke, Lalitha Visvanath & Katy Börner  
InfoVis Lab @ Indiana University  
2004

# Mapping the Evolution of Co-Authorship Networks

Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest



## Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams

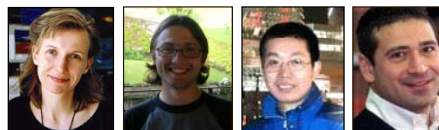
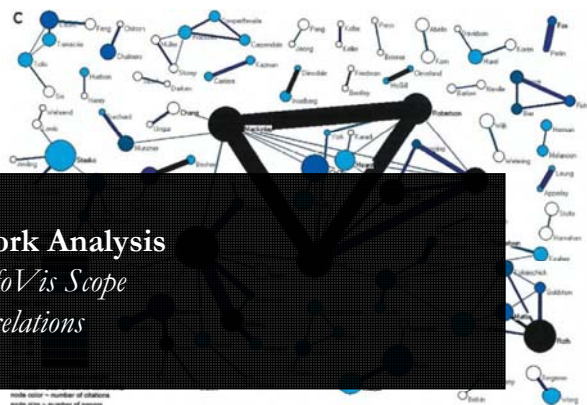
Börner, Dall'Asta, Ke & Vespignani (2005) *Complexity*, 10(4):58-67.

### Research question:

- Is science driven by prolific single experts or by high-impact co-authorship teams?

### Contributions:

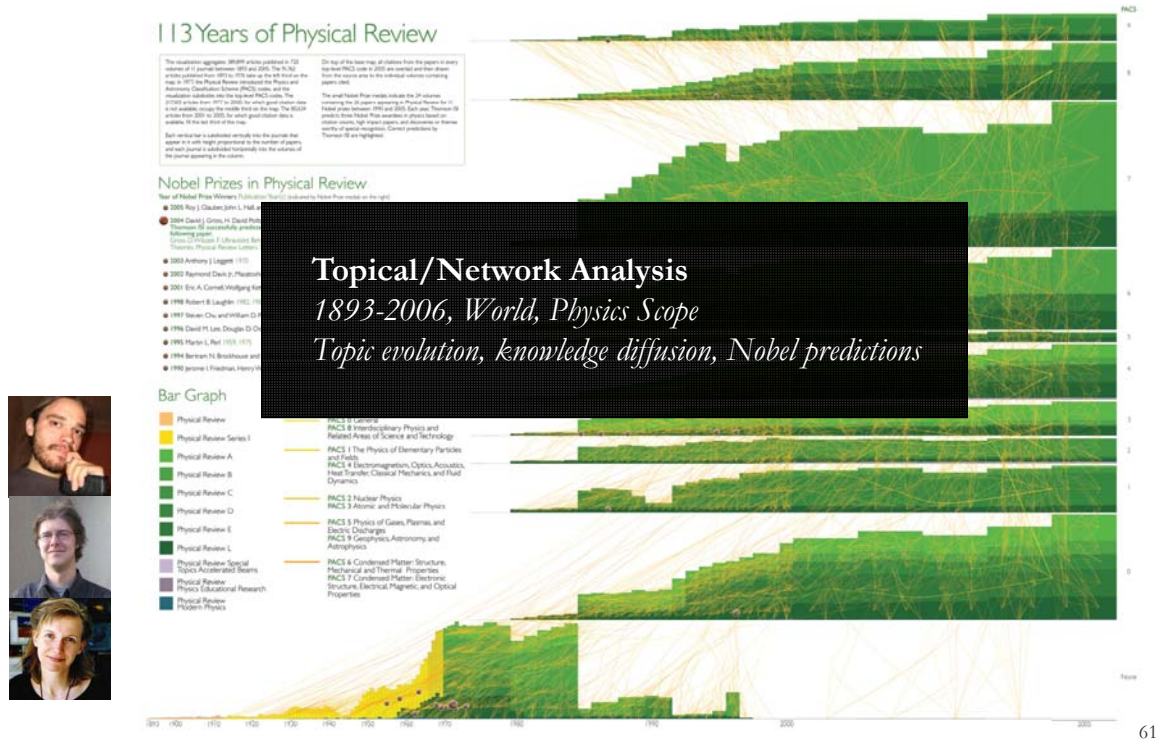
- New approach to allocate citational credit.
- Novel weighted graph
- Visualization of the co-author network
- Centrality measure impact.
- Global statistical analysis of paper production and citations in correlation with co-authorship team size over time.
- Local, author-centered entropy measure.



# 113 Years of Physical Review

[http://scimaps.org/dev/map\\_detail.php?map\\_id=171](http://scimaps.org/dev/map_detail.php?map_id=171)

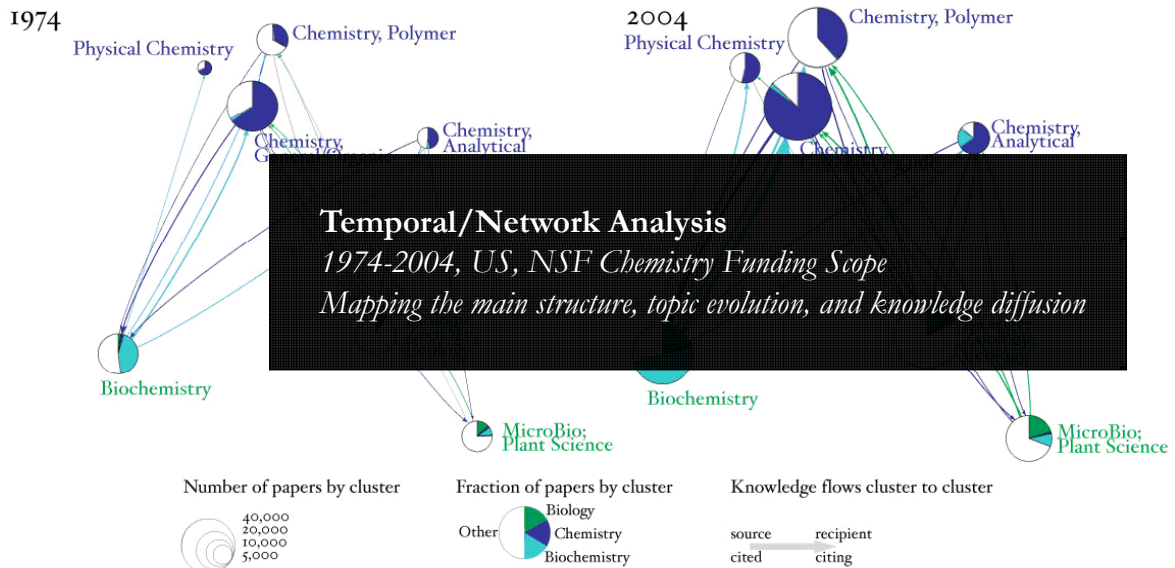
Bruce W. Herr II and Russell Dubon (Data Mining & Visualization), Elisha F. Hardy (Graphic Design), Shashikant Penumarthy (Data Preparation) and Katy Börner (Concept)



# Topical Composition and Knowledge Flow Patterns in Chemistry Research for 1974 and 2004

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007)

## Chemistry - Biology Interface



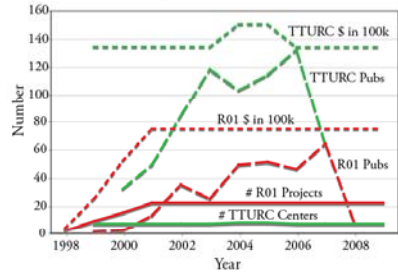


# Mapping Transdisciplinary Tobacco Use Research Centers Publications

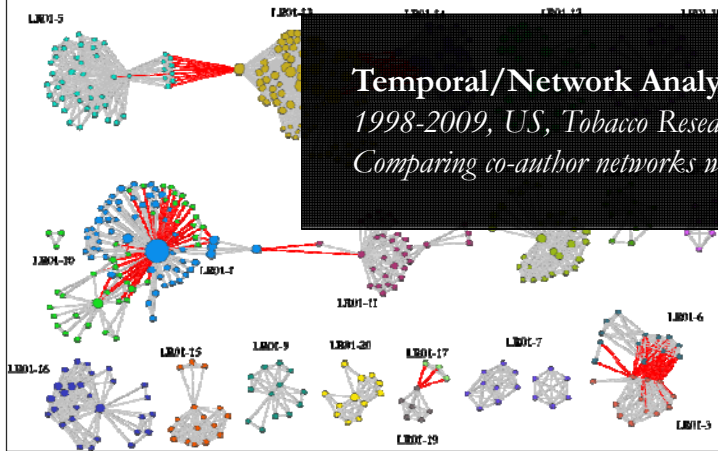
Compare R01 investigator based funding with TTURC Center awards in terms of number of publications and evolving co-author networks.

Zoss & Börner, *forthcoming*.

R01 & TTURC Project Information



Longitudinal R01 Co-Authorship Network



TTURC Co-Authorship Network

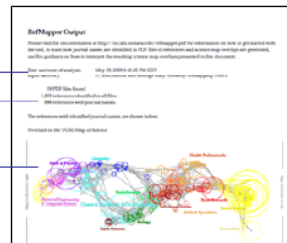


**Temporal/Network Analysis**  
 1998-2009, US, Tobacco Research Scope  
 Comparing co-author networks with different funding

# Reference Mapper

Dubon & Börner, *forthcoming*.

(a) Overview



Date and input directory  
 Basic counts  
 Overlay of all matched journal references from all PDF files on 554 scientific disciplines (nodes) in UCSD Map of Science  
 Circle size denotes # references  
 Listing of all references grouped by 13 science areas

(b) Visual Index



For each PDF file:  
 Basic counts and thumbnail science map  
 Max 18 per page

**Topical/Network Analysis**  
 2009, US, NSF Funding  
 Grouping interdisciplinary funding proposals for review

For each PDF file:  
 Overlay of all matched journal references on 554 scientific fields (nodes) in UCSD Map of Science  
 Circle size denotes # references  
 Colors and names of science areas that are cited  
 Alphabetic listing of cited journals and # of times cited

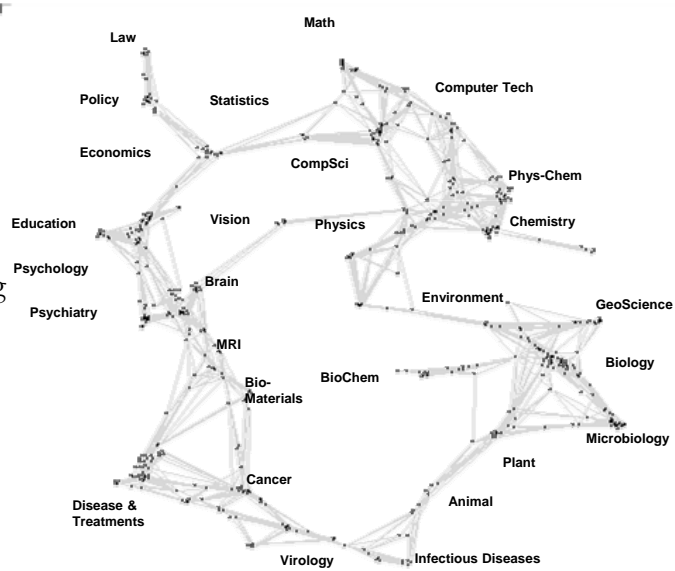


Top-n most similar PDF files identified based on journal name co-occurrences  
 The similarity of each PDF file to itself is 1  
 Overlay of matched journal references from all above listed PDF files on UCSD Map of Science and grouping by 13 science areas

## Latest 'Base Map' of Science

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007). *Mapping the Structure and Evolution of Chemistry Research*. 11th International Conference on Scientometrics and Informetrics. pp. 112-123.

- Uses combined SCI/SSCI from 2002
  - 1.07M papers, 24.5M references, 7,300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts were reaggregated at the journal cluster level to calculate the
  - (x,y) positions for each journal cluster
  - by association, (x,y) positions for each journal

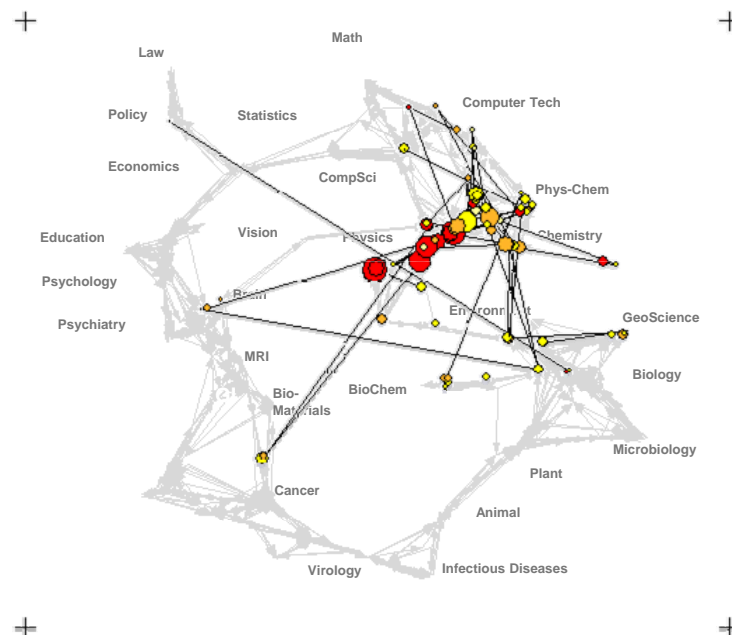


65

## Science map applications: Identifying core competency

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

Funding patterns of the US Department of Energy (DOE)

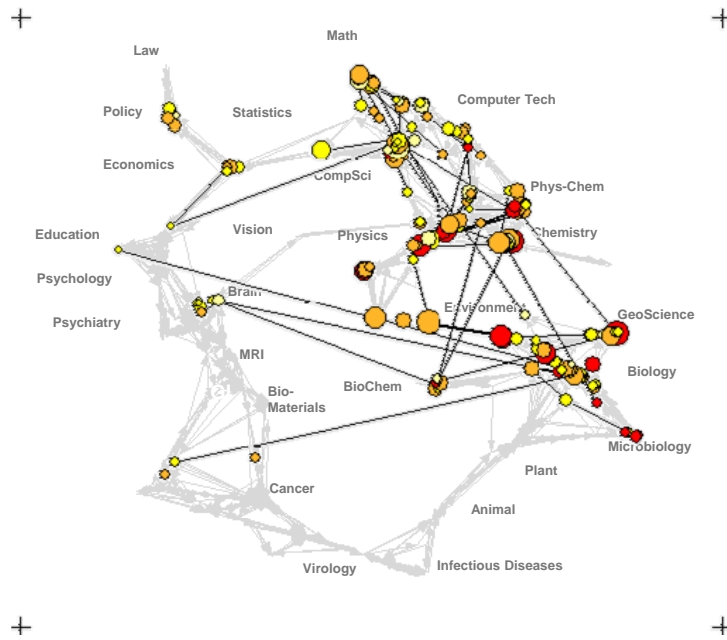


66

## Science map applications: Identifying core competency

*Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).*

### Funding Patterns of the National Science Foundation (NSF)

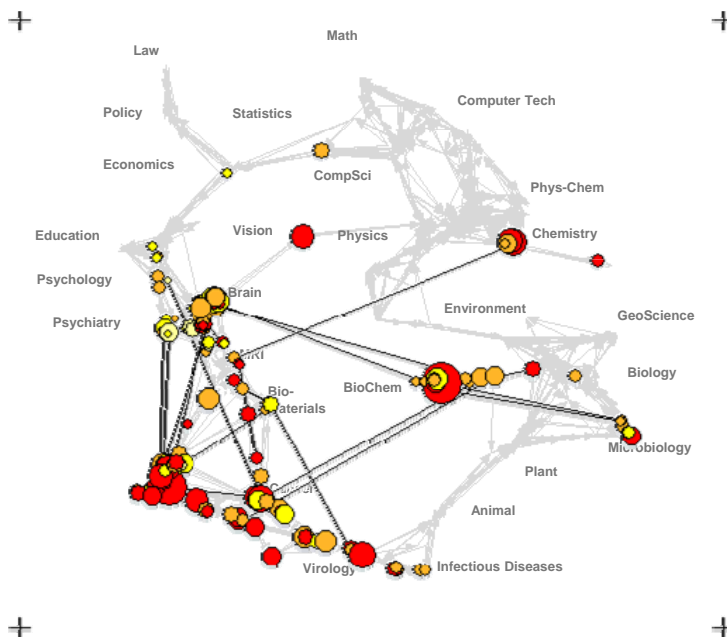


67

## Science map applications: Identifying core competency

*Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).*

### Funding Patterns of the National Institutes of Health (NIH)

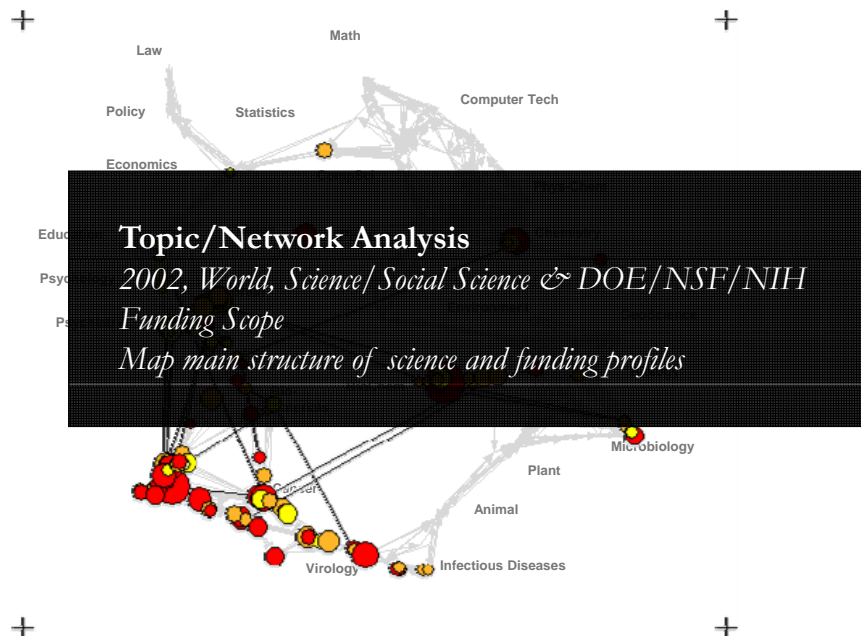


68

## Science map applications: Identifying core competency

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

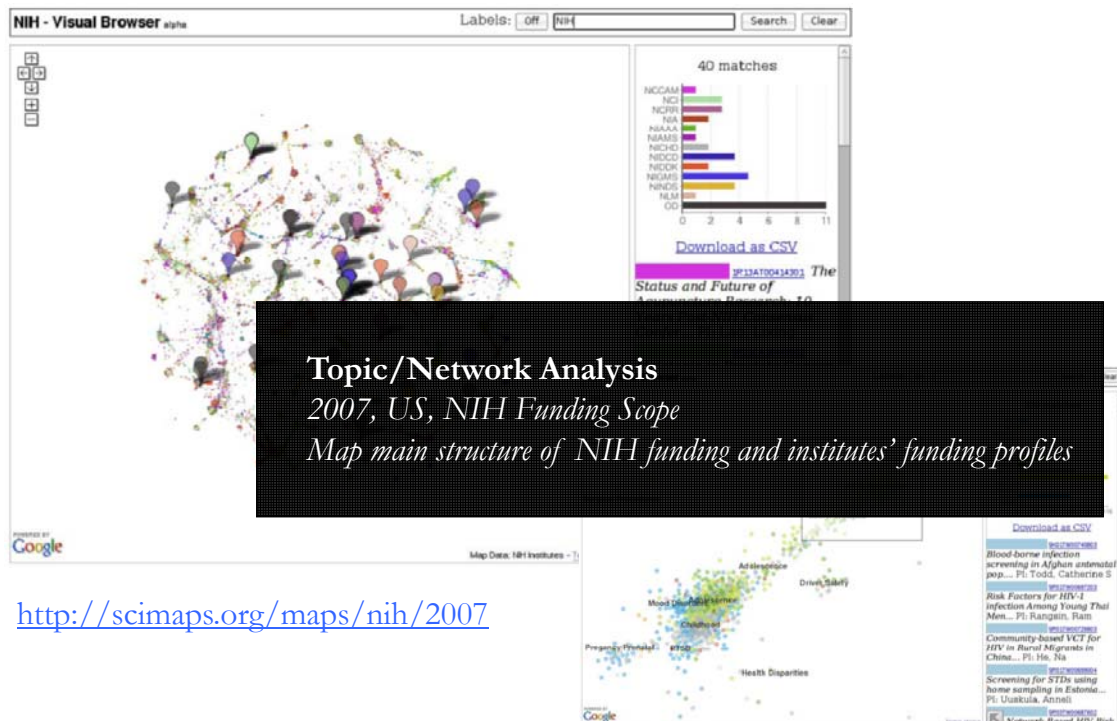
### Funding Patterns of the National Institutes of Health (NIH)



69

## Interactive Science Map of NIH Funding

Herr II, Bruce W., Talley, Edmund M, Burns, Gully APC, Newman, David & La Rowe, Gavin. (2009).



70



# Interactive World and Science Map of S&T Jobs

Angela Zoss, Michael Connover, Katy Börner (2010).

**Visualization of Job Postings**

Map of Science | Geographic

**Visualization of Job Postings**

Map of Science | Geographic

Postdoc at Harvard Medical School  
[Link to Post](#)

**Visualization of Job Postings**

Map of Science | Geographic

**Map of Science**  
Scientific domains are highly

**Geospatial/Topic Analysis**  
2008/2009, World, 100 Job RSS feeds  
Map evolving job market in real time

Earth Sciences | Biology | Humanities

POWERED BY Google

Copyright © 2008 The Regents of the University of California - [Terms of Use](#)

Search for Jobs

Between the papers and journals to cluster journals into small groups of highly related journals.

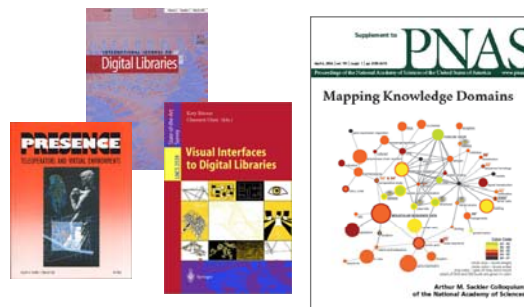
Those clusters are represented by 554 individual nodes in the network. The links between the clusters show that some clusters are related to other clusters but are not as tightly connected as the journals that make up each cluster. Then the clusters are labeled both by the content area shared by the journals in the cluster and by the overarching scientific domain for that cluster (represented by one of 13 colors).

Maps of science like this one can be used to understand many different data sets and how they can be represented by topic. Here we are looking at the topics that appear in job postings from large inh

71

## Computational Scientometrics: Studying Science by Scientific Means

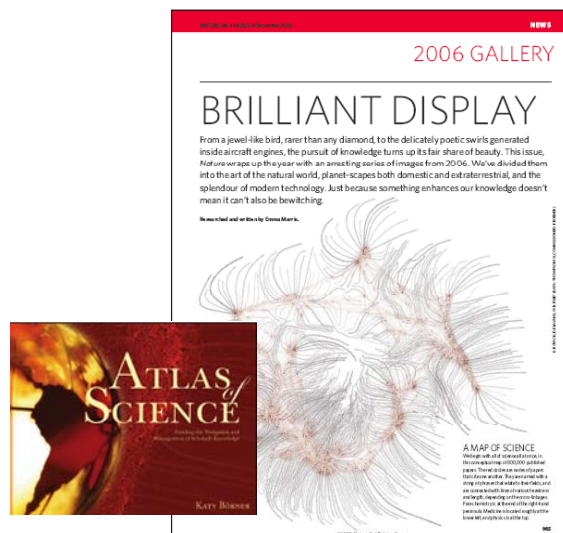
Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). **Visualizing Knowledge Domains**. In Blaise Cronin (Ed.), *ARIST*, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, Volume 37, Chapter 5, pp. 179-255.  
<http://ivl.slis.indiana.edu/km/pub/2003-borner-arist.pdf>



Shiffrin, Richard M. and Börner, Katy (Eds.) (2004). **Mapping Knowledge Domains**. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl\_1).  
[http://www.pnas.org/content/vol101/suppl\\_1/](http://www.pnas.org/content/vol101/suppl_1/)

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>

Börner, Katy (2010) *Atlas of Science*. MIT Press.  
<http://scimaps.org/atlas>



72



## There are many more questions than answers: First results from a questionnaire study on insights needed by science policy makers

*Priority scale of 1-5, with 1=urgent to 5=nice to know*

### Priority Questions

#### Temporal Analysis

- 1 funding trends in individual institutes, all NIH, all funding / Topical – to examine NIH scientific topic area broadly and in detail
- 1 Topical/temporal – how are the current structures of scientific/translational/clinical research changing, what are the emerging areas, and how are the submitted applications different from awarded grants in these areas.
- 2 What new biomedical fields of research are emerging, and 1) is NIH currently funding such research, 2) are there enough trained scientists to address these new research fields, and 3) where is the emerging fields research being conducted (are there geographic clusters)?
- 2 Temporal patterns of distribution / Temporal – examine scientific trends
- 3 What are the prevailing trends in topics receiving funding across NIH? By specific institute?
- 3 Meso vs global (topical/temporal) – how does NIH funding relate to funding from other agencies/countries

#### Geospatial Analysis

- 1 Diffusion of knowledge globally
- 5 Have there been any changes in degree of international collaboration in the biomedical sciences?

#### Topic Analysis

- 1 What NIH Funds / How do we identify emerging concept / Are there emerging areas of opportunity to which NIH should direct more support?
- 1 How are NIH research findings being used by partners, health providers and the public?
- 2 How do we identify gaps in knowledge?
- 2 How can we characterize (or categorize) the research that NIH supports? AND How do these areas of investment compare to public health needs?

#### Network Analysis

- 2 How can we quickly understand the current network of nodule and collaboration? What information will we need to do so?
- 4 Have our efforts to encourage interdisciplinary research been effective? And which strategies have been the most effective?
- ? Identify instances of knowledge transfer within and across research networks
- ? Network approaches to measuring or detecting innovation? E.g. publication or concept that disturbs the stability of a network.

73



## The Science of Science (Sci2) Tool

- 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**.

74

# BREAK



## Overview

- Macroscopes and the Changing Scientific Landscape 15 mins
  - Introduction to network science with sample maps and insights 15 mins
  - Introduction to the Network Workbench Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  
  - Introduction to science studies with sample maps and insights 15 mins
  - **Introduction to the Sci<sup>2</sup> Tool** 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - **Overview of validation approaches for science studies**
  
  - Plug-and-play tool design using OSGi/CIshell 30 mins
  - Scholarly Marketplaces 15 mins
- 240 mins**



## NWB Tool vs. Science of Science (Sci2) Tool

### Similarities:

- Both use OSGi/CIShell and are easy to extend/customize.
- Same general interface, look and feel.
- Sci<sup>2</sup> uses many NWB plugins.

### Differences:

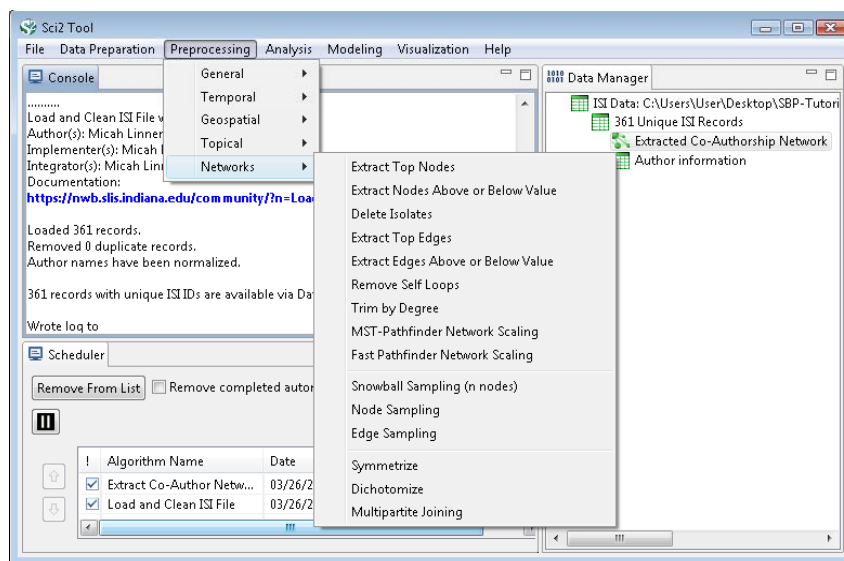
- Different target communities, branding, tutorials.
- Both come with different OSGi/CIShell plugin sets, sample datasets, and menu structures.
- Sci<sup>2</sup> has database support which makes it more scalable.
- Sci<sup>2</sup> has improved GUESS functionality.
- Sci<sup>2</sup> has more standard workflows and visualizations.
- Sci<sup>2</sup> (alpha 0.3) is less mature than NWB (v 1.0).

77



## Sci<sup>2</sup> Tool for Science of Science Research and Practice (See Demo DVD)

- Sci2-Linux32bit
- Sci2-Linux64bit
- Sci2-MacG3G4G5
- Sci2-MacIntel
- Sci2-Windows



### Acknowledgments

This work is supported in part by the Cyberinfrastructure for Network Science center and the School of Library and Information Science at Indiana University, the National Science Foundation under Grant No. SBE-0738111 and IIS-0513650, and the James S. McDonnell Foundation.



78



## Process of Computational Scientometrics

DATA EXTRACTION	UNIT OF ANALYSIS	MEASURES	LAYOUT (often one code does both similarity and ordination steps)		DISPLAY
			SIMILARITY	ORDINATION	
SEARCHES ISI INSPEC Eng Index Medline ResearchIndex Patents etc.	COMMON CHOICES Journal Document Author Term	COUNTS/FREQUENCIES Attributes (e.g. terms) Author citations Co-citations By year  THRESHOLDS By counts	SCALAR (unit by unit matrix) Direct citation Co-citation Combined linkage Co-word / co-term Co-classification  VECTOR (unit by attribute matrix) Vector space model (words/terms) Latent Semantic Analysis (words/terms) incl. Singular Value Decomposition (SVD)  CORRELATION (if desired) Pearson's R on any of above	DIMENSIONALITY REDUCTION Eigenvector/ Eigenvalue solutions Factor Analysis (FA) and Principal Components Analysis (PCA) Multi-dimensional scaling (MDS) LSA, <b>Topics</b> Pathfinder networks (PFNet) Self-organizing maps (SOM) includes SOM, ET-maps, etc.	INTERACTION Browse Pan Zoom Filter Query Detail on demand  ANALYSIS  CLUSTER ANALYSIS  SCALAR Triangulation Force-directed placement (FDP)
BROADENING By citation By terms					

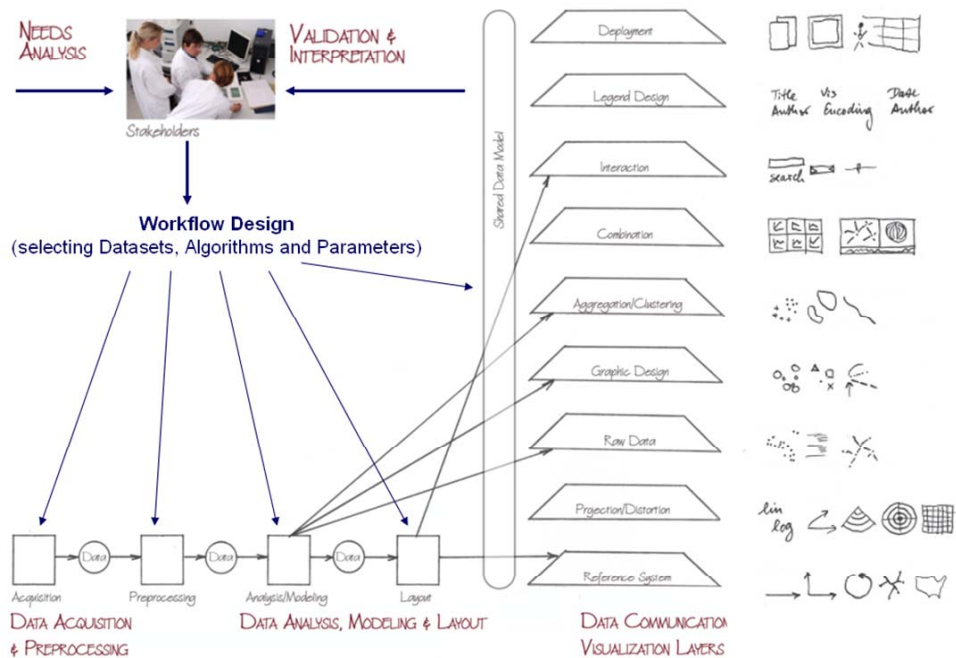
Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003) *Visualizing Knowledge Domains*. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology, Volume 37*. Medford, NJ: Information Today, Inc./ American Society for Information Science and Technology, chapter 5, pp. 179-255.

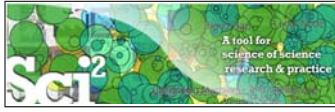
79



## Needs-Driven Workflow Design

using a modular data acquisition/analysis/modeling/visualization pipeline as well as modular visualization layers.





## Sci<sup>2</sup> Tool: Algorithms

See <https://nwb.slis.indiana.edu/community>

### Preprocessing

Extract Top N% Records  
Extract Top N Records  
Normalize Text  
Slice Table by Line

-----  
Extract Top Nodes  
Extract Nodes Above or Below Value  
Delete Isolates

-----  
Extract top Edges  
Extract Edges Above or Below Value  
Remove Self Loops  
Trim by Degree  
MST-Pathfinder Network Scaling  
Fast Pathfinder Network Scaling

-----  
Snowball Sampling (in nodes)  
Node Sampling  
Edge Sampling

-----  
Symmetrize  
Dichotomize  
Multipartite Joining

-----  
Geocoder  
-----  
Extract ZIP Code

### Modeling

Random Graph  
Watts-Strogatz  
Small World  
Barabási-Albert Scale-Free  
TARL

### Analysis

Network Analysis Toolkit (NAT)  
Unweighted & Undirected

Node Degree  
Degree Distribution

-----  
K-Nearest Neighbor (Java)  
Watts-Strogatz Clustering Coefficient  
Watts Strogatz Clustering Coefficient over K

-----  
Diameter  
Average Shortest Path  
Shortest Path Distribution  
Node Betweenness Centrality

-----  
Weak Component Clustering  
Global Connected Components

-----  
Extract K-Core  
Annotate K-Coreeness

-----  
HTTS

### Weighted & Undirected

Clustering Coefficient  
Nearest Neighbor Degree  
Strength vs Degree  
Degree & Strength  
Average Weight vs End-point Degree  
Strength Distribution  
Weight Distribution  
Randomize Weights

-----  
Blondel Community Detection

### HTTS

### Unweighted & Directed

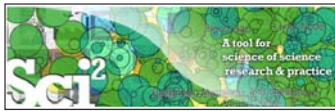
Node Indegree  
Node Outdegree  
Indegree Distribution  
Outdegree Distribution

-----  
K-Nearest Neighbor  
Single Node in-Out Degree Correlations

-----  
Dyad Reciprocity  
Arc Reciprocity  
Adjacency Transitivity

-----  
Weak Component Clustering  
Strong Component Clustering

81



## Sci<sup>2</sup> Tool: Algorithms cont.

See <https://nwb.slis.indiana.edu/community>

-----  
Extract K-Core  
Annotate K-Coreeness

-----  
HTTS  
PageRank  
Weighted & Directed  
HTTS  
Weighted PageRank

### Textual

Burst Detection

### Visualization

GnuPlot  
GUESS  
Image Viewer

-----  
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 Annotation Style)

### Geo Map (Colored-Region Annotation Style)

### \*Science Map (Circle Annotation)

### Scientometrics

Remove ISI Duplicate Records  
Remove Rows with Multitudinous Fields  
Detect Duplicate Nodes  
Update Network by Merging Nodes

### Extract Directed Network

Extract Paper Citation Network  
Extract Author Paper Network

### Extract Co-Occurrence Network

Extract Word Co-Occurrence Network  
Extract Co-Author Network  
Extract Reference Co-Occurrence  
(Bibliographic Coupling) Network

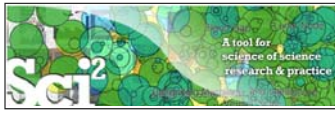
-----  
Extract Document Co-Citation Network

\* Requires permission from UCSD  
All four+ save into Postscript files.

[General Network extraction](#)

82

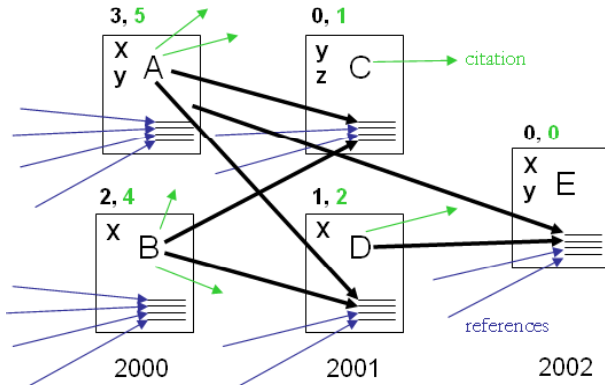




## Network Extraction

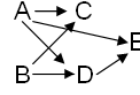
Sample paper network (left) and four different network types derived from it (right).  
From ISI files, about 30 different networks can be extracted.

Papers A-E written by authors x, y, z over 3 years.  
Each paper happens to have 4 references.



### Paper-Paper Citation Network

Papers are connected via direct citation links.  
Arrows represent information flow from older papers to younger papers.



### Author-Author (Co-Author) Network

x and y co-author papers A and E together  
y and z co-author papers A and E



### Document Co-Citation (DCA) Network

A and B are co-cited by C and D  
A and D are co-cited by E



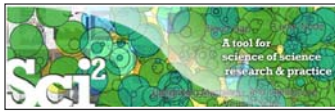
### Reference Co-Occurrence (Bibliographic Coupling) Network

C and D are bibliographically coupled as they both cite/reference A and B.



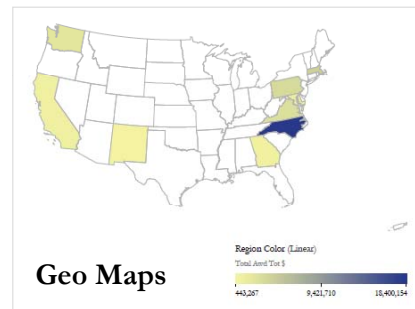
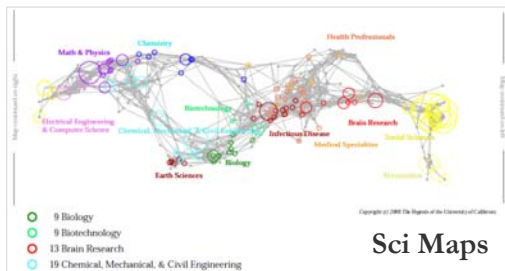
Local citation counts (within this dataset) are given in **black** and global citation counts (ISI times cited) are given in **green** above each paper.

83



## Sci² Tool

Plugins that render into Postscript files:



### Horizontal Time Graphs



Börner, Katy, Huang, Weixia (Bonnie), Linnemeier, Micah, Dubon, Russell Jackson, Phillips, Patrick, Ma, Nianli, Zoss, Angela, Guo, Hanning & Price, Mark. (2009). *Retz-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.*

84





## Overview

- Macroscopes and the Changing Scientific Landscape 15 mins
  - Introduction to network science with sample maps and insights 15 mins
  - Introduction to the Network Workbench Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - Introduction to science studies with sample maps and insights 15 mins
  - Introduction to the Sci<sup>2</sup> Tool 15 mins
  - **Demo and hands-on data analysis and visualization by participants** 60 mins
  - **Overview of validation approaches for science studies**
  - Plug-and-play tool design using OSGi/CIshell 30 mins
  - Scholarly Marketplaces 15 mins
- 240 mins**

85



## Type of Analysis vs. Scale of Level of Analysis

	<i>Micro/Individual</i> (1-100 records)	<i>Meso/Local</i> (101-10,000 records)	<i>Macro/Global</i> (10,000 < records)
<b>Statistical Analysis/Profiling</b>	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of USA, all of science.
<b>Temporal Analysis (When)</b>	Funding portfolio of one individual	Mapping topic bursts in 20-years of PNAS	113 Years of physics Research
<b>Geospatial Analysis (Where)</b>	Career trajectory of one individual	Mapping a states intellectual landscape	PNAS publications
<b>Topical Analysis (What)</b>	Base knowledge from which one grant draws.	Knowledge flows in Chemistry research	VxOrd/Topic maps of NIH funding
<b>Network Analysis (With Whom?)</b>	NSF Co-PI network of one individual	Co-author network	NSF's core competency

86

## Sci<sup>2</sup> Tool Demo and Hands-on Data Analysis and Visualization

### Micro/Individual (1-100 records)

- Mapping Collaboration, Publication and Funding Profiles of One Researcher (EndNote and NSF Data) (*section 5.1.1*)
- Studying Four Major NetSci Researchers (ISI Data) using Database (*section 5.1.5*)

### Meso/Local (101–10,000 records)

- Mapping CTSA Centers (NIH RePORTER Data) (*section 5.2.3*)
- Biomedical Funding Profile of NSF (NSF Data) (*section 5.2.4*)
- Mapping the Field of RNAi Research (SDB Data) (*section 5.2.7*)

### Macro/Global (10,000 < records)

- Geo USPTO (SDB Data) (*section 5.3.1*)

87

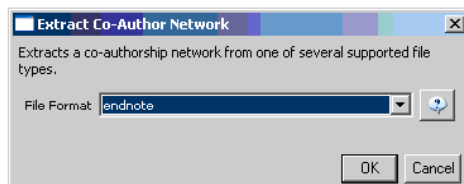


### Mapping Collaboration, Publication and Funding Profiles of One Researcher (*section 5.1.1*)

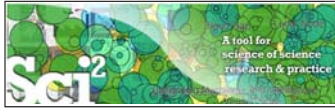
<b>KatyBorner.enw</b>	
<b>Time frame:</b>	1992-2010
<b>Region(s):</b>	Indiana University, University of Technology in Leipzig, University of Freiburg, University of Bielefeld
<b>Topical Area(s):</b>	Network Science, Library and Information Science, Informatics and Computing, Statistics, Cyberinfrastructure,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

Many researchers/publishers use EndNote to organize bibliographies.

To analyze an individual researcher's collaboration and publication profile, load an EndNote file into the Sci2 Tool, e.g., load the Katy Borner's EndNote file at '*\*yoursci2directory\*/sampledata/scientometrics/endnote/KatyBorner.enw*' and run '*Data Preparation > Text Files > Extract Co-Author Network*' using the parameter:



88



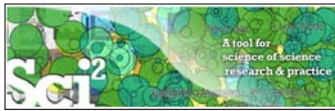
## Mapping Collaboration, Publication and Funding Profiles of One Researcher (section 5.1.1)

Next, run 'Analysis > Networks > Unweighted & Undirected > Node Degree' to append degree information to each node. To visualize the network, run 'Visualization > Networks > GUESS' and select 'GEM' in the 'Layout' menu once the graph is fully loaded. Optimize layout design using the following workflow:

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

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

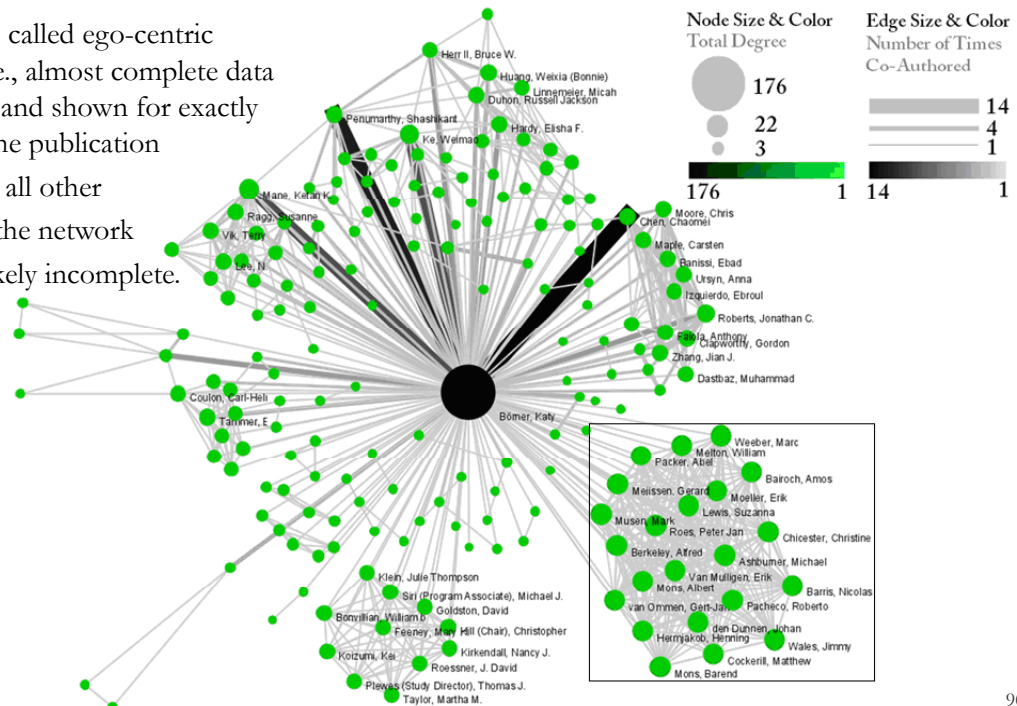
89



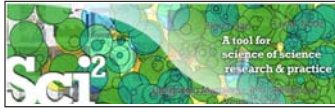
## Mapping Collaboration, Publication and Funding Profiles of One Researcher (section 5.1.1)

### Co-Author Network

This is a so called ego-centric network, i.e., almost complete data is available and shown for exactly one ego. The publication records for all other authors in the network are most likely incomplete.



90



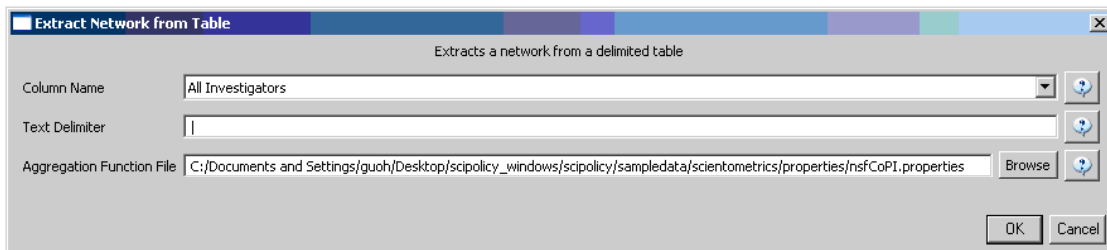
## Mapping Collaboration, Publication and Funding Profiles of One Researcher (section 5.1.1)

### Funding Data Analysis

Free online services such as NSF’s Award Search (See [Section 4.2.2.1 NSF Award Search](#)) support the retrieval of ego-centric funding profiles. Here, a search was exemplarily conducted for “Katy Borner” in the “Principal Investigator” field while keeping the “Include CO-PI” box checked.

The resulting data is available at

*‘\*yoursci2directory\*/sampledata/scientometrics/nsf/KatyBorner.nsf.’* Load the data using *‘File > Load’*, select the loaded dataset in the Data Manager window, and run *‘Data Preparation > Text Files > Extract Co-Occurrence Network’* using these parameters:







91



## Mapping Collaboration, Publication and Funding Profiles of One Researcher (section 5.1.1)

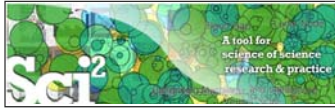
Select the “*Extracted Network on Column All Investigator*” network and run *‘Analysis > Networks > Network Analysis Toolkit (NAT)’* to reveal that there are 13 nodes and 28 edges in the network without isolates. Select *‘Visualization > Networks > GUESS’* to visualize the resulting Co-PI network. Select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

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

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

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

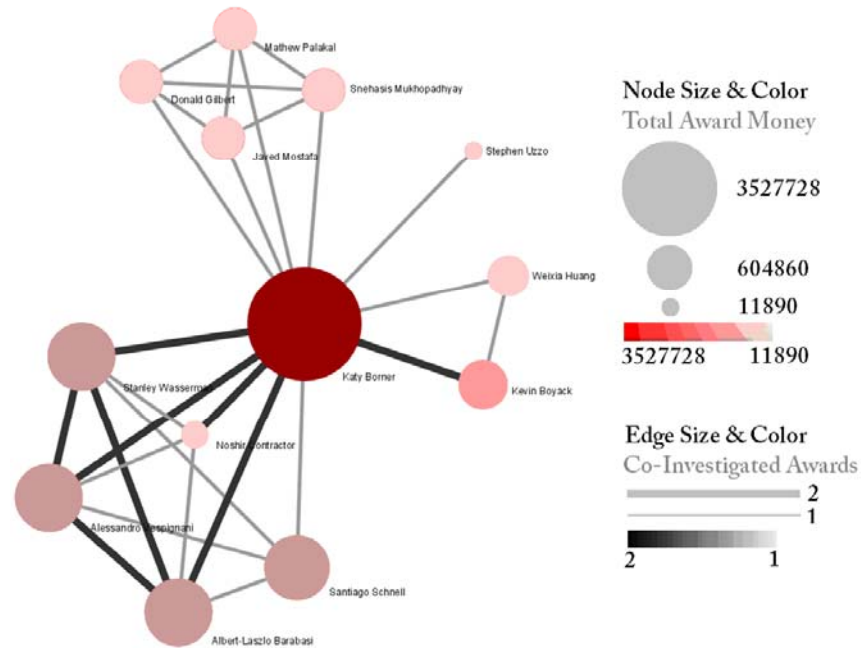
92



## Mapping Collaboration, Publication and Funding Profiles of One Researcher (section 5.1.1)

### Co-PI Network

This is a so called ego-centric network, i.e., almost complete data is available and shown for exactly one ego. The funding records for all other people in the network are most likely incomplete.



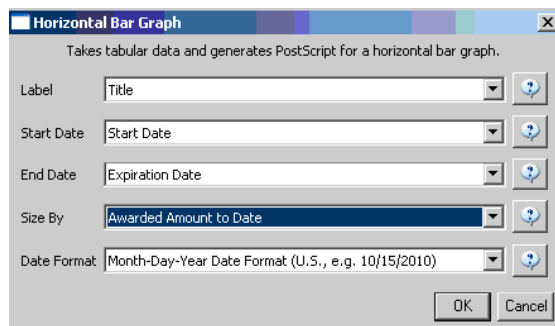
93



## Mapping Collaboration, Publication and Funding Profiles of One Researcher (section 5.1.1)

### Award Durations and Totals

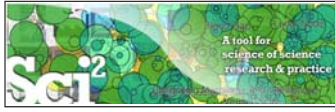
For a summary of the grants themselves, with a visual representation of their award amount, select the NSF csv file in the Data Manager and run '*Visualization > Temporal > Horizontal Bar Graph*', entering the following parameters:



The generated postscript file can be viewed using Adobe Distiller or GhostViewer (see Section [2.4 Saving Visualizations for Publication](#)).

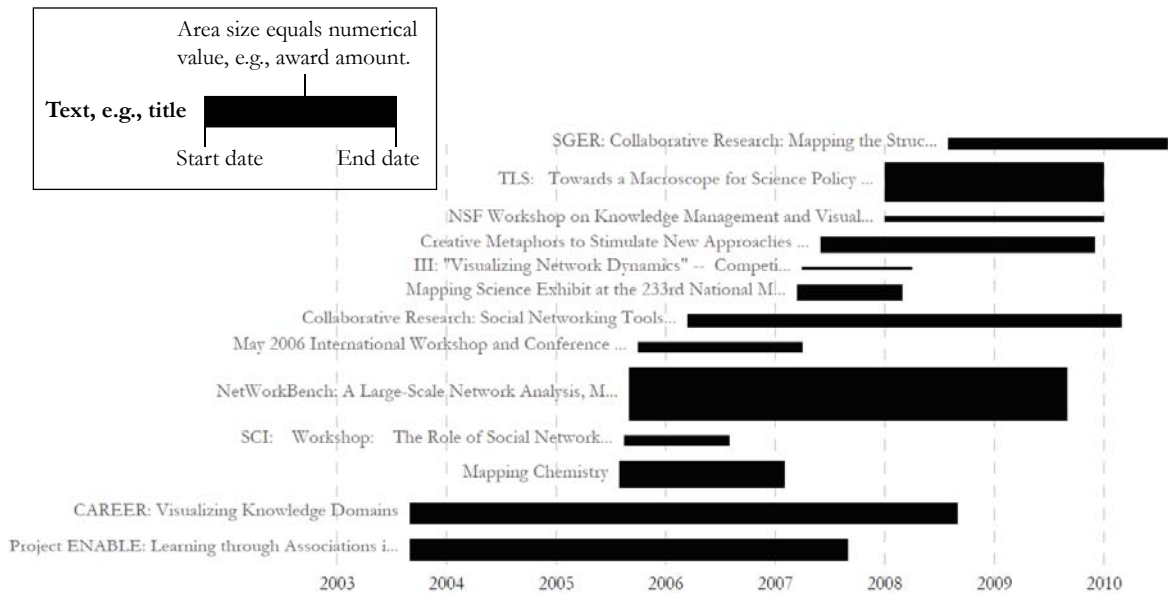
94





## Mapping Collaboration, Publication and Funding Profiles of One Researcher (*section 5.1.1*)

### Award Durations and Totals



95



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

FourNetSciResearchers.isi	
<b>Time frame:</b>	1955-2007
<b>Region(s):</b>	Miscellaneous
<b>Topical Area(s):</b>	Network Science
<b>Analysis Type(s):</b>	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.

96



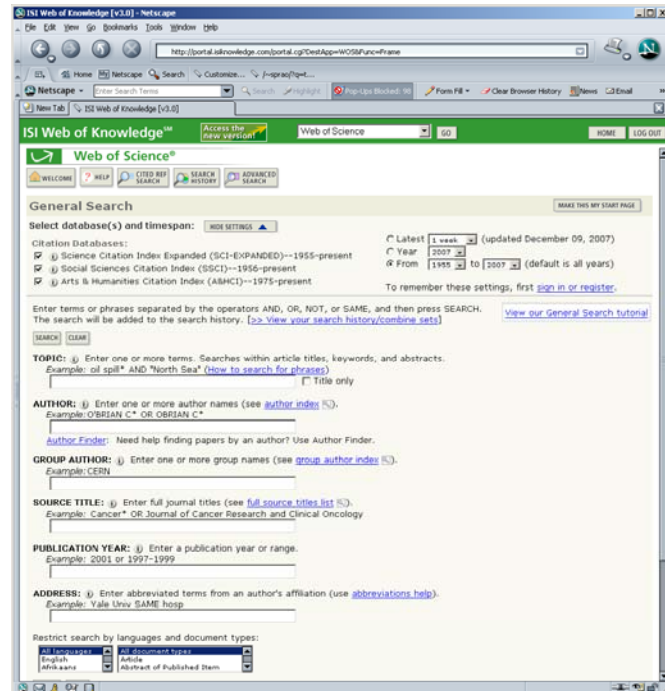
## Data Acquisition from Web of Science

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



97



## 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>

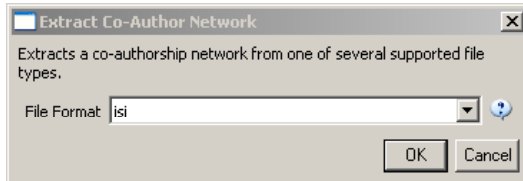
98



## Extract Co-Author Network

Load *\*yournwbdirectory\*/sampledata/scientometrics/isi/FourNetSciResearchers.isi'* using *'File > Load and Clean ISI File'*.

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

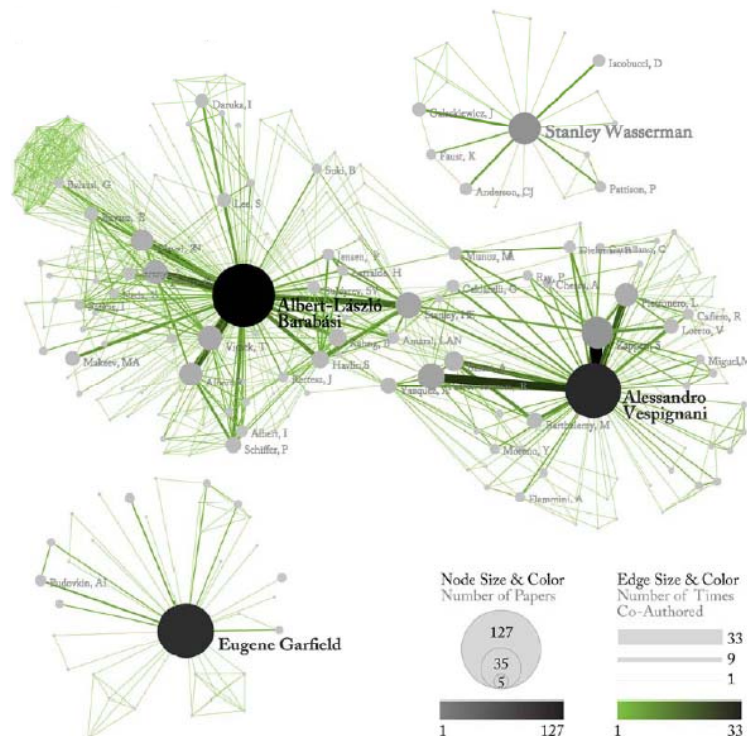


The result is an undirected network of co-authors in the Data Manager. It has 247 nodes and 891 edges.

To view the complete network, select the network and run *'Visualization > GUESS > GEM'*. Run *Script > Run Script... . And select Script folder > GUESS > co-author-nw.py*.

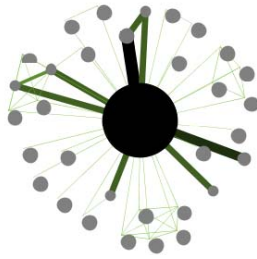


## Joint Co-Author Network of all Four NetsSci Researchers

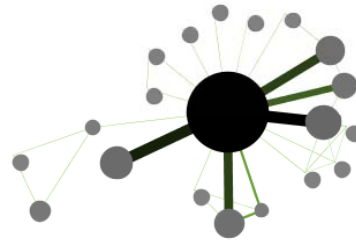




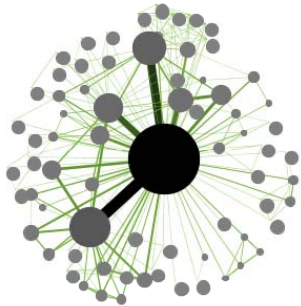
## Individual Co-Author Networks (Read/map 4 files separately)



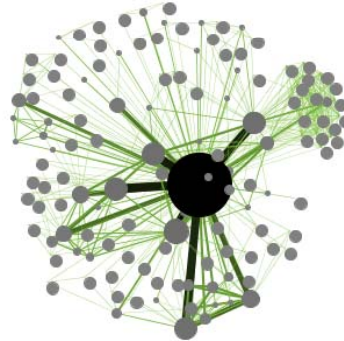
Eugene Garfield



Stanley Wasserman



Alessandro Vespignani

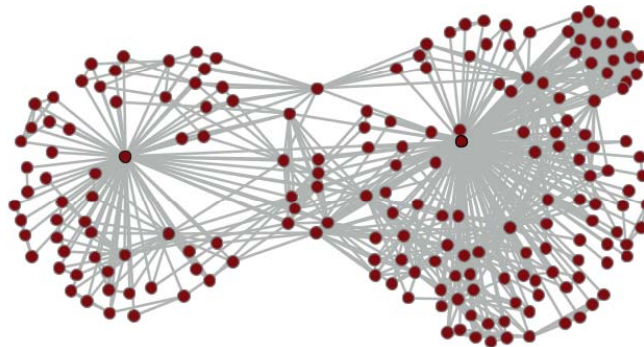


Albert-László Barabási

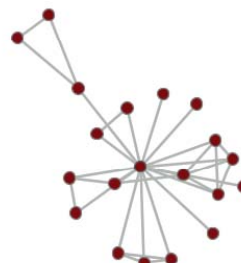
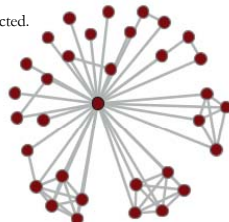


## Network Visualization: Node Layout

*Network Science researchers*



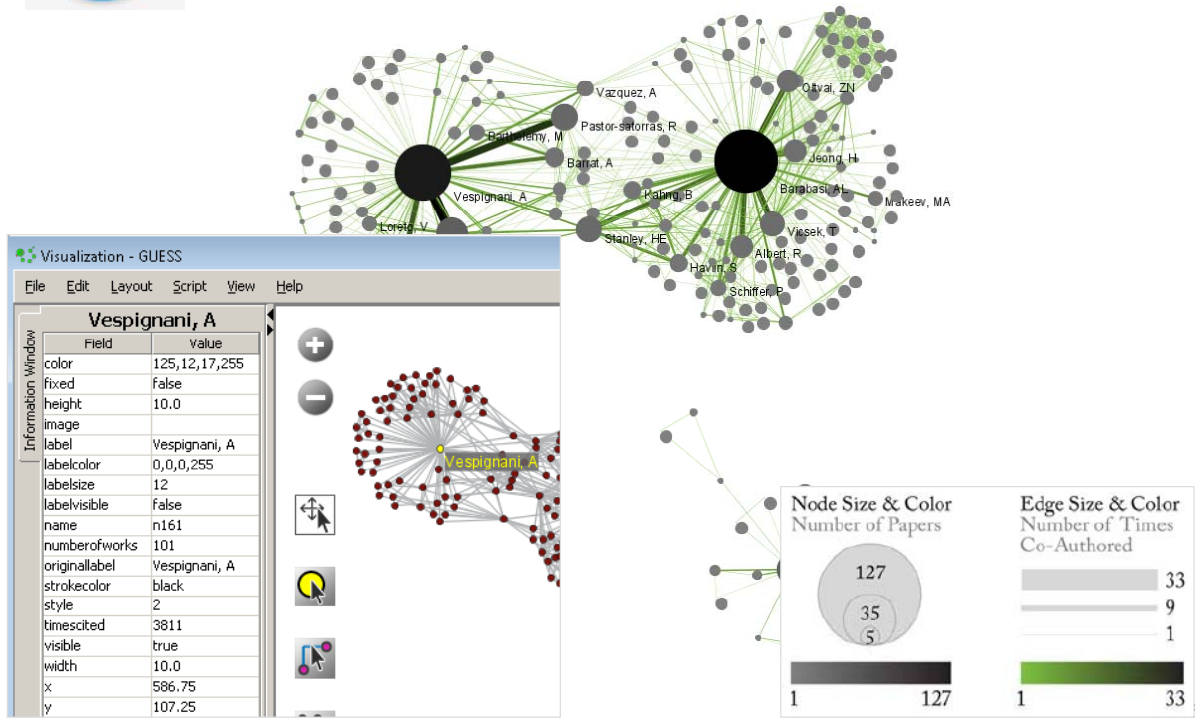
Load and Clean ISI File was selected.  
 Loaded 361 records.  
 Removed 0 duplicate records.  
 Author names have been normalized.  
 361 records with unique ISI IDs are available via Data Manager.  
 .....  
 Extract Co-Author Network was selected.  
 Input Parameters:  
 File Format: isi  
 ..... 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





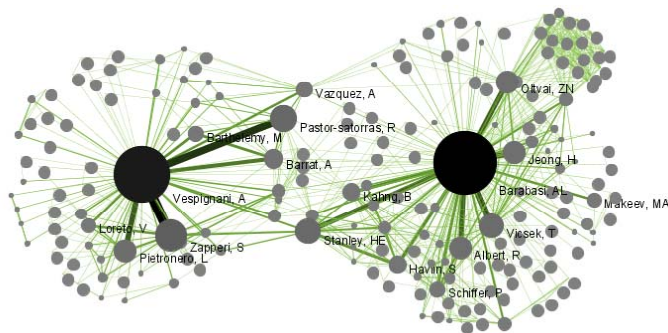
## Network Visualization: Color/Size Coding by Data Attribute Values

*Network Science researchers*



## Network Visualization: Giant Component

*Network Science researchers*



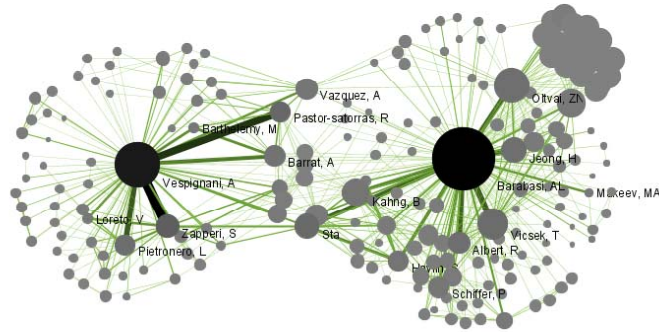
.....  
Weak Component Clustering was selected.  
Implementer(s): Russell Duhon  
Integrator(s): Russell Duhon

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

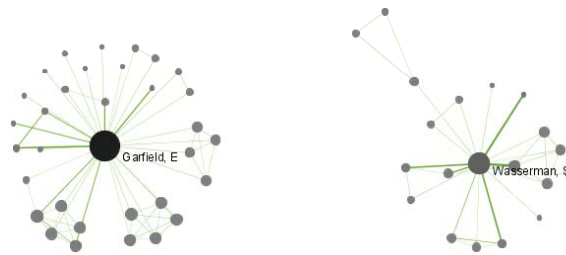




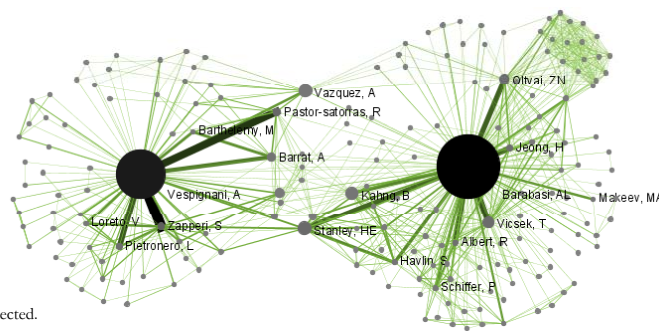
**Network Visualization:**  
**Color/Size Coding by Degree**  
*Network Science researchers*



.....  
 Node Degree was selected.  
 Documentation:  
<https://nwb.slis.indiana.edu/community/?n=AnalyzeData.NodeDegree>  
 .....

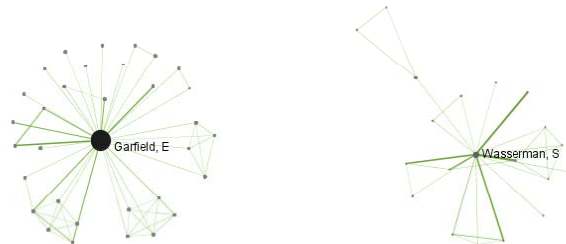


**Network Visualization:**  
**Color/Size Coding by Betweenness Centrality**  
*Network Science researchers*



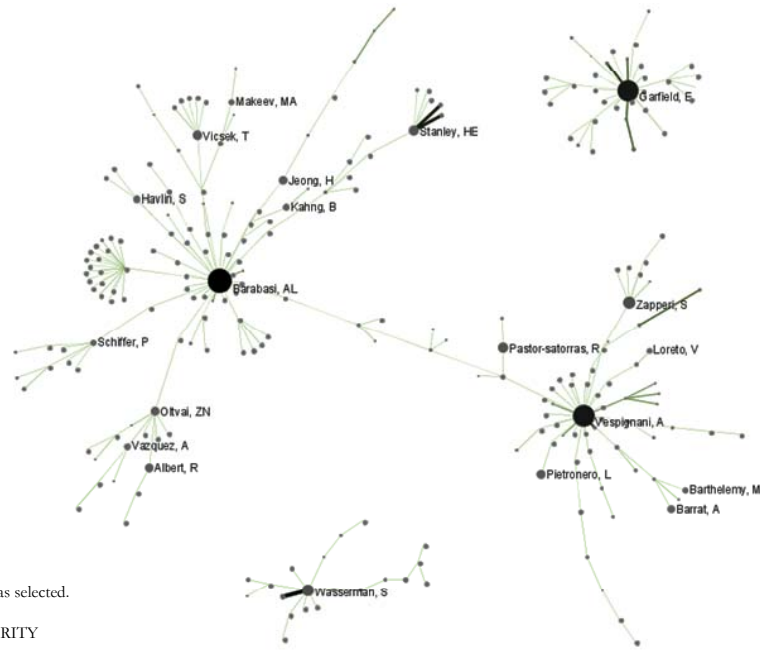
.....  
 Node Betweenness Centrality was selected.  
 Author(s): L. C. Freeman  
 Implementer(s): Santo Fortunato  
 Integrator(s): Santo Fortunato, Weixia Huang  
 Reference: Freeman, L. C. (1977). A set of measuring centrality based on betweenness.  
 Sociometry. 40:35-41.

Input Parameters:  
 Number of bins: 10  
 umber of bins: 10  
 .....





## Network Visualization: Reduced Network After Pathfinder Network Scaling *Network Science researchers*



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

107



## Paper-Citation Network Layout

Load *'\*yournwbdirectory\*/sampledata/scientometrics/isi/FourNetSciResearchers.isi'* using *'File > Load and Clean ISI File'*.

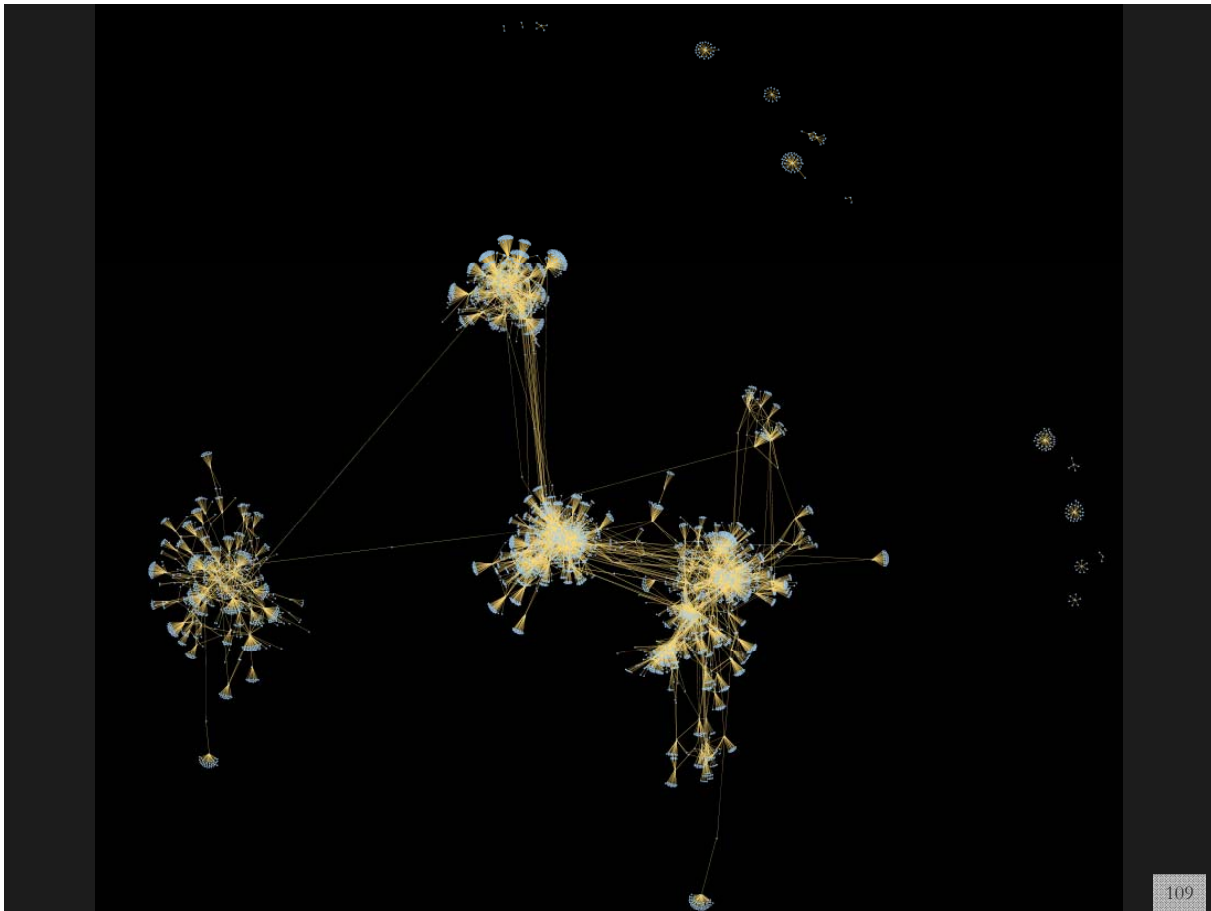
To extract the paper-citation network, select the *'361 Unique ISI Records'* table and run *'Scientometrics > Extract Directed Network'* using the parameters:

Source Column	Cited References
Target Column	Cite Me As
Text Delimiter	
Aggregate Function File	C:\Documents and Settings\kaly\Desktop\nwb\sampledata\scientometrics\properties\isiPaperCitation.properties

The result is a directed network of paper citations in the Data Manager. It has 5,335 nodes and 9,595 edges.

To view the complete network, select the network and run *'Visualization > GUESS'*. Run *'Script > Run Script ...'* and select *'yournwbdirectory\*/script/GUESS/paper-citation-nw.py'*.

108



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

#### Replicate Studies Using Database Support

Load *\*yoursci2directory\*/sampledata/scientometrics/isi/FourNetSciResearchers.isi*, using *'File > Load'* instead of *'File > Load and Clean ISI File'*.

Run *'File > Load Into Database > Load ISI File Into Database'*. View the database schema by right-clicking on the loaded database in the Data Manager and clicking "View".

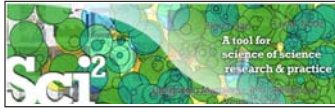
The screenshot shows the Sci2 Tool interface with the Data Manager window open. The Data Manager shows a list of databases, with 'ISI Data' selected. A context menu is open over 'ISI Data' with 'View' selected. Below the Data Manager is a Scheduler window with a table of tasks:

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

To the right is a Notepad window showing the database schema:

```

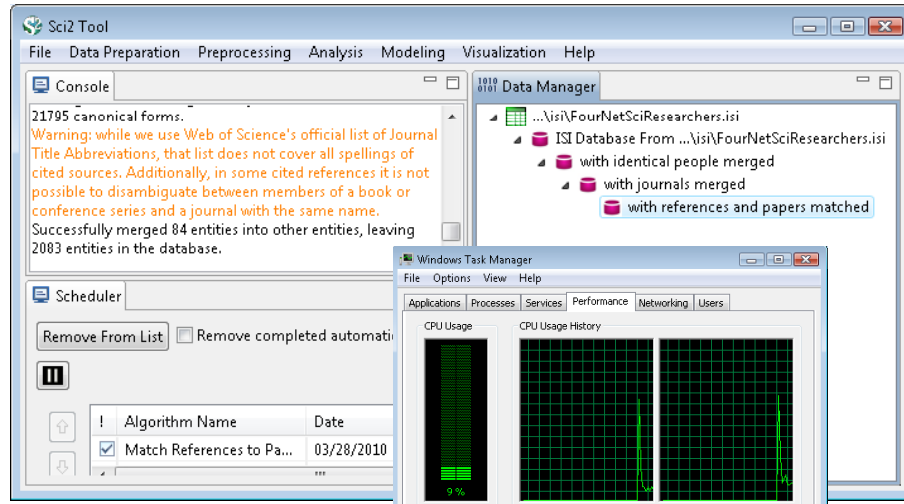
ADDRESS ( PK INTEGER, ADDRESS_CITY VARCHAR, ADDRESS_COUNTF
AUTHORS ( AUTHORS_DOCUMENT_FK INTEGER, AUTHORS_PERSON_FK I
AUTHORS_DOCUMENT_FK ----> DOCUMENT.PK
AUTHORS_PERSON_FK ----> PERSON.PK
CITED_PATENTS ( CITED_PATENTS_DOCUMENT_FK INTEGER, CITED_F
CITED_PATENTS_DOCUMENT_FK ----> DOCUMENT.PK
CITED_PATENTS_PATENT_FK ----> 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
  
```



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

### Replicate Studies Using Database Support – Unification

Run ‘Data Preparation > Database > ISI > Merge Identical ISI People’, followed by ‘Data Preparation > Database > ISI > Merge Journals’ 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.



111



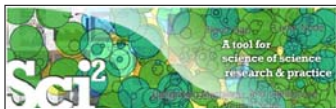
## Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

### 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	UNSPLIT_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		

112



## 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												
1	A	B	C	D	E	F	G	H	I	J	K	L
YR	DOCUMENTS	REFERENCES	TOTAL_REFERENCES	DISTINCT_REFERENCES	DISTINCT_AUTHORS	DISTINCT_SOURCES	DISTINCT_OTHER_KEYWORDS	DISTINCT_ISI_KEYWORDS	DISTINCT_AUTHOR_KEYWORDS	DISTINCT_SOURCE_KEYWORDS	DISTINCT_OTHER_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		

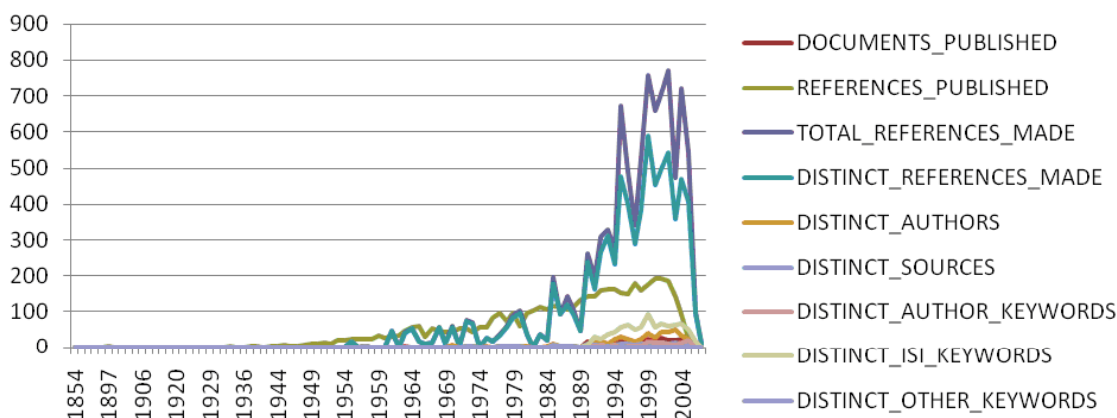
113



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

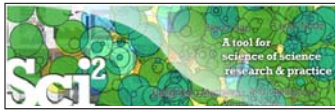
### Using Database Support – Records over time

Aggregate data by year by running ‘Data Preparation > Database > ISI > Extract Authors > 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.



114





## 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

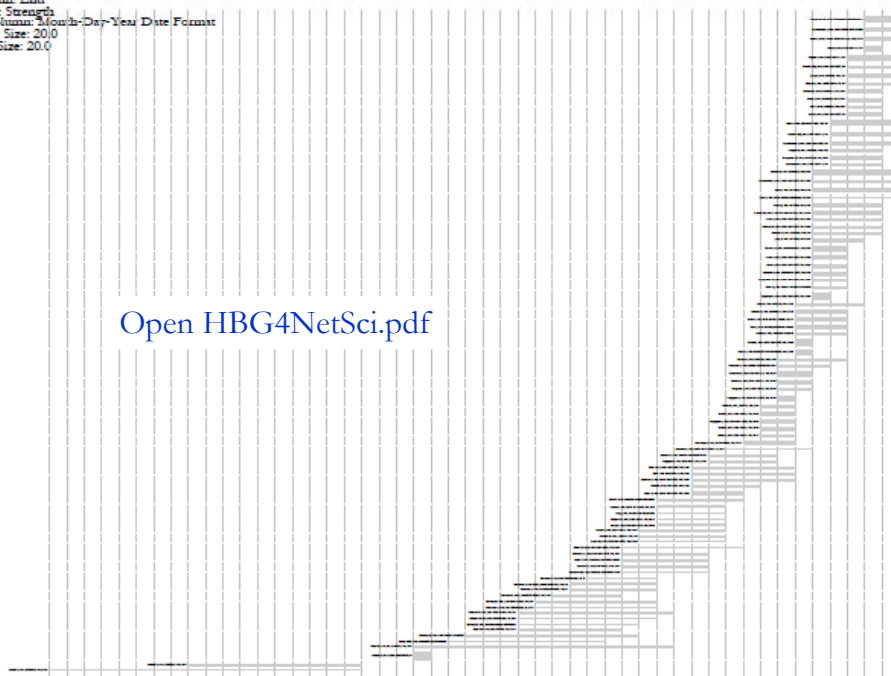
115

Date and time of analysis: March 28, 2010 7:43:48 PM EDT  
Input data: Burst detection analysis (Year, Reference): maximum burst level 1

#### Horizontal Bar Graph for maximum burst level 1

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

[Open HBG4NetSci.pdf](#)

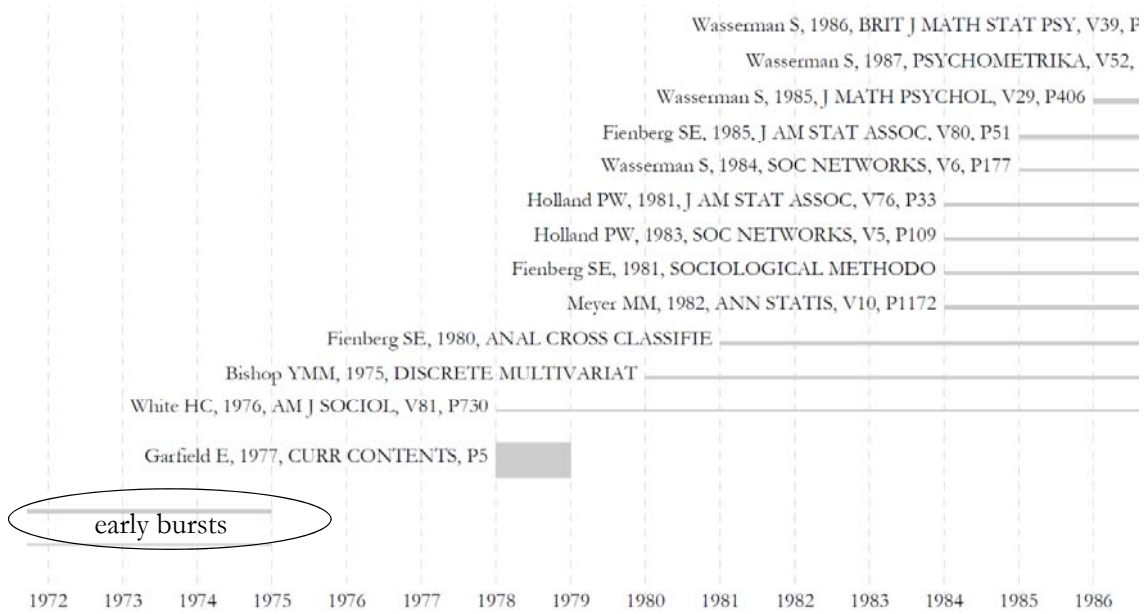


116

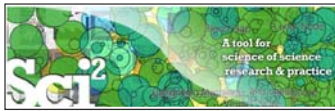


## Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

### Using Database Support – Burst Analysis Result



117



## 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.



JournalsScienceMap-FourNetSciResearchers.pdf

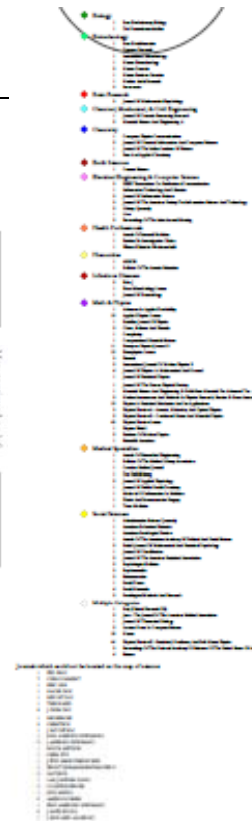
Science Map

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

Journal column: Journal Name (Abbreviated)

Dataset display name: FourNetSciResearchers.isi

OK Cancel



118



## Mapping CTSA Centers (NIH RePORTER Data) (section 5.2.3)

CTSA2005-2009.xls	
<b>Time frame:</b>	2005-2009
<b>Region(s):</b>	Miscellaneous
<b>Topical Area(s):</b>	Clinical and Translational Science
<b>Analysis Type(s):</b>	PI-Institution Network, Co-Authorship Network

A study of all NIH Clinical and Translational Science Awards (CTSA) awards and resulting publications from 2005-2009, requires advanced data acquisition and manipulation to prepare the required data. Data comes from the union of NIH RePORTER downloads (see Section 4.2.2.2 NIH RePORTER) and NIH ExPORTER data dumps (<http://projectreporter.nih.gov/exporter/>). CTSA Center grants were identified first and then matched with resulting publications using a project-specific ID. The result file is available as an Excel file in *\*yoursci2directory\*/sampledata/scientometrics/nih*. The file contains two spreadsheets, one with publication data and one with grant data. Save each spreadsheet out as *grants.csv* and *publications.csv*.

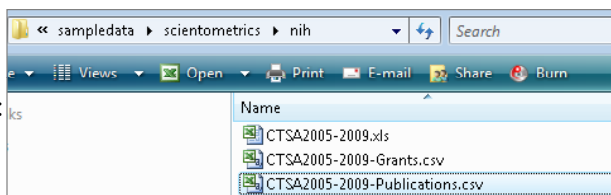
119



## Mapping CTSA Centers (NIH RePORTER Data) (section 5.2.3)

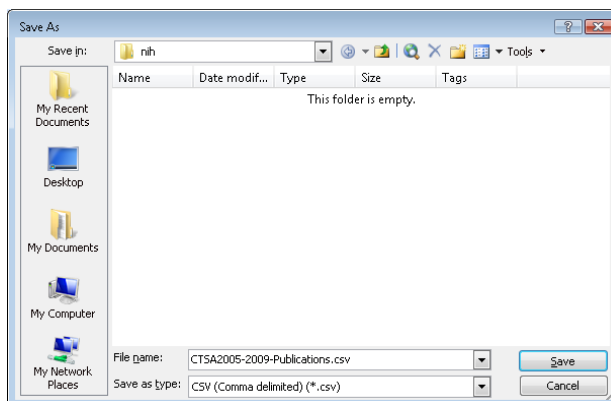
### Data Preparation

Open *../nih/CTSA2005-2009.xls* in MS Excel. It contains two worksheets: 'Grants' and 'Publications'.

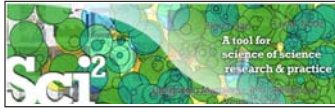


Save both worksheets separately as CSV (comma delimited) files, e.g.,

- *CTSA2005-2009-Grants.csv*
- *CTSA2005-2009-Publications.csv*



120



## Mapping CTSA Centers (NIH RePORTER Data) (section 5.2.3)

### Bimodal Network of Funded Institutions and PIs

Extract Directed Network was selected.

Input Parameters:

Source Column: Organization Name

Text Delimiter: |

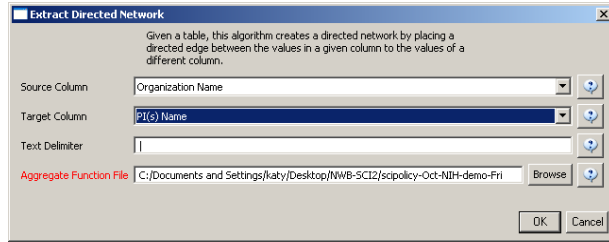
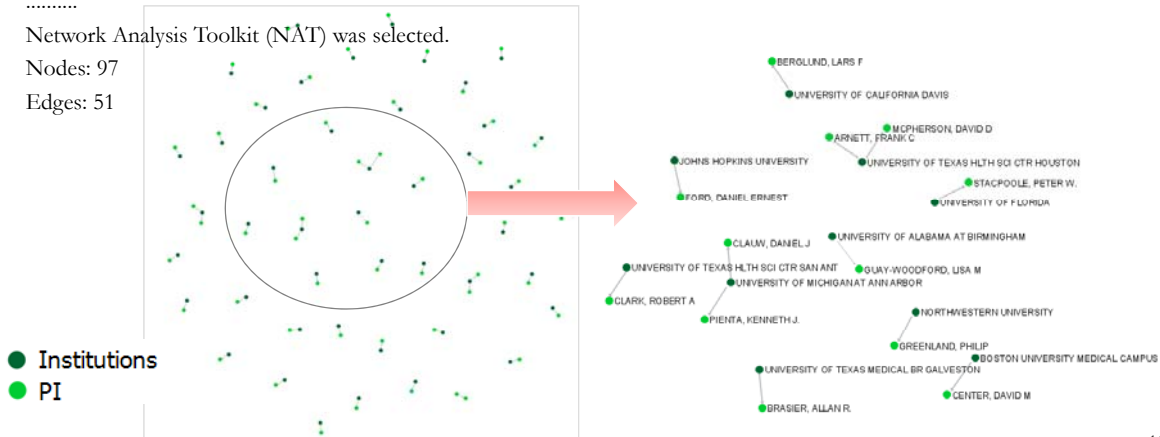
Target Column: PI(s) Name

.....

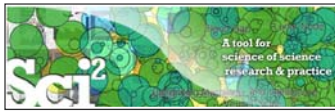
Network Analysis Toolkit (NAT) was selected.

Nodes: 97

Edges: 51



121



## Mapping CTSA Centers (NIH RePORTER Data) (section 5.2.3)

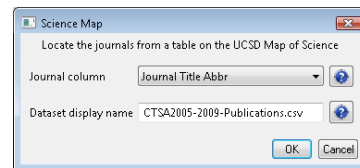
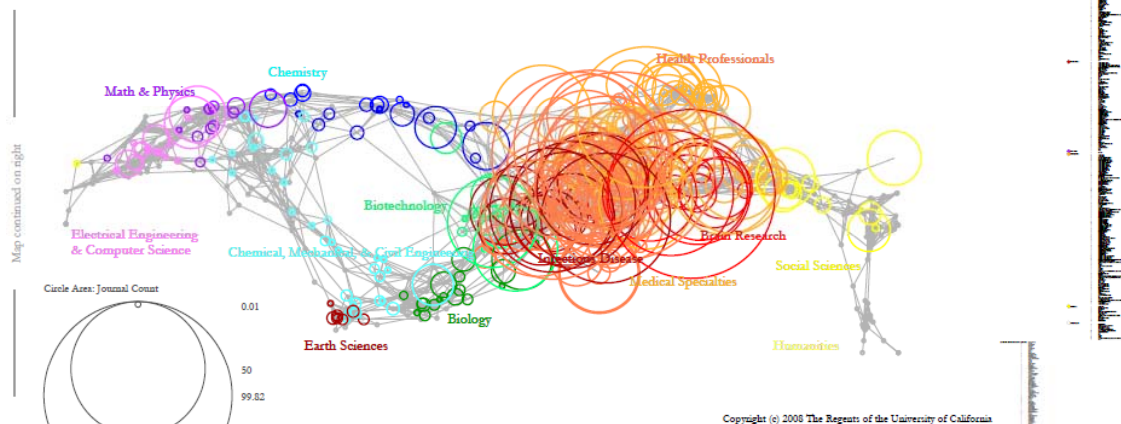
### Topic Coverage of Publications

*Visualization > Topical > Science Map via Journals'*

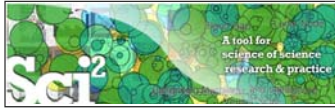
Science Map via Journals for CTS2005-2009-Publications.csv

2,226 journal references matched out of 2,456 found.

These 2,226 references are associated with 12 of 13 disciplines of science and 303 of 554 research specialties in the UCSD Map of Science



122



## Mapping CTSA Centers (NIH RePORTER Data) (section 5.2.3)

### NIH CTSA Grants: Publication Co-Author Network

Extract Co-Occurrence Network was selected.

Input Parameters:

Text Delimiter: ;

Column Name: Authors

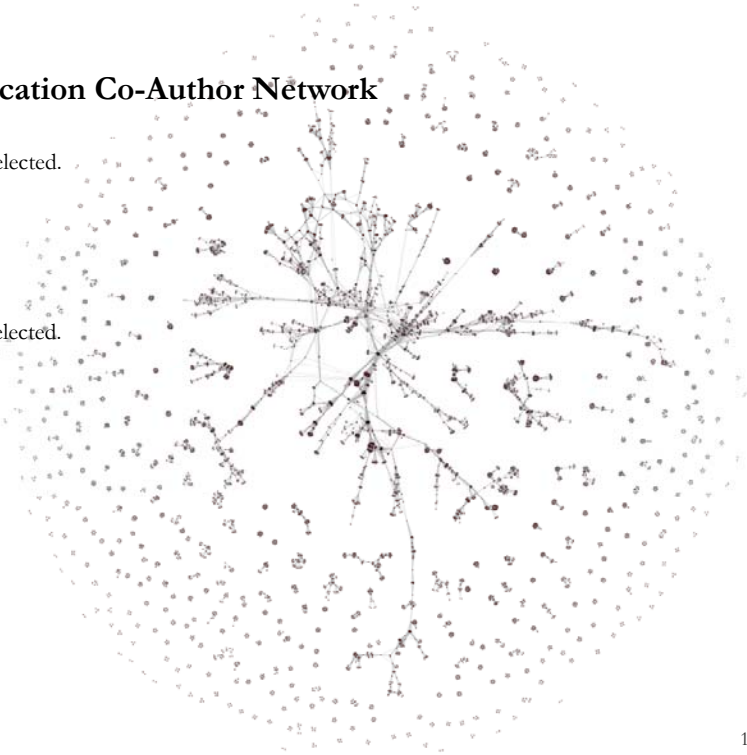
.....

Network Analysis Toolkit (NAT) was selected.

Nodes: 8680

Isolated nodes: 27

Edges: 50160



123



## Mapping CTSA Centers (NIH RePORTER Data) (section 5.2.3)

### Extract Largest (Giant) Component

Weak Component Clustering

was selected.

Input Parameters:

Number of top clusters: 10

535 clusters found

.....

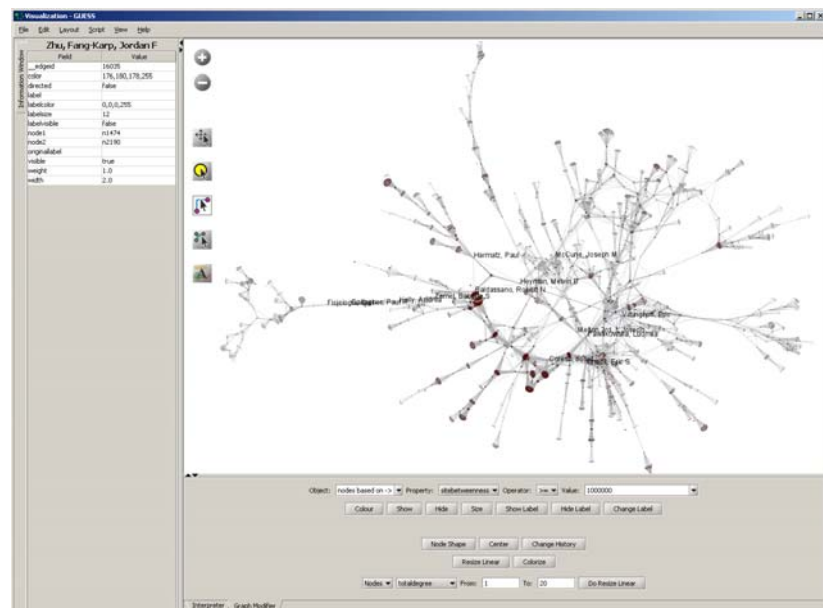
Network Analysis Toolkit (NAT)

was selected.

Nodes: 3239

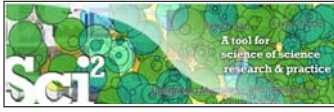
Edges: 24969

See Poster



124





## Biomedical Funding Profile of NSF (NSF Data) (section 5.2.4)

MedicalAndHealth.nsf	
<b>Time frame:</b>	2003-2010
<b>Region(s):</b>	Miscellaneous
<b>Topical Area(s):</b>	Biomedical
<b>Analysis Type(s):</b>	NSF Organization-Program Network

What organizations and programs at the National Science Foundation support projects that deal with medical and health related topics? Data was downloaded from the NSF Awards Search SIRE (<http://www.nsf.gov/awardsearch>) on Nov 23rd, 2009, using the query “medical AND health” in the title, abstract, and awards field, with “Active awards only” checked (see section 4.2.2.1 [NSF Award Search](#) for data retrieval details).

125



## Biomedical Funding Profile of NSF (NSF Data) (section 5.2.4)

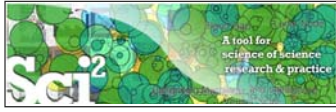
Using NSF Awards Search:  
<http://www.nsf.gov/awardsearch>  
download relevant NSF awards that have “medical” AND “health” in title, abstract, and awards. Active awards only.

Number of awards: 283 awards  
Total awarded amount to date:  
\$152,015,288

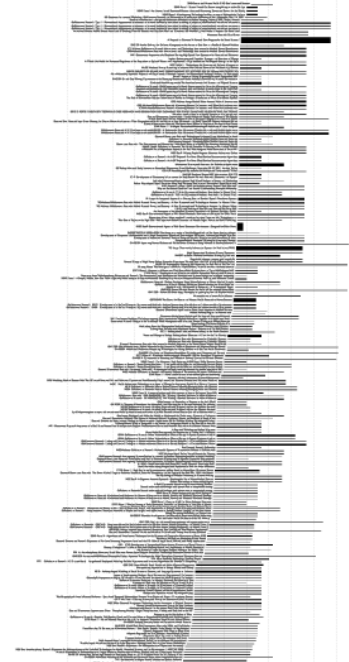
*Retrieved on Oct 18, 2009*

NSF - Award Search - Award Information - Mozilla Firefox  
http://www.nsf.gov/awardsearch/awardsearch.do?searchType=PCA  
NSF - Award Search  
National Science Foundation  
WHERE DISCOVERIES BEGIN  
HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT | FastLane  
Award Search  
Send Comments | Award Search Help  
Awarder Information | Program Information | Search All Free-Text | Search All Fields | Show  
Hint: The text field below 'Search Award For' searches the title, abstract, and award number fields.  
Search Award For: "medical" and "health"  
Restrict to Title Only:   
Awarder Information  
Principal Investigator  
First Name:   
Last Name:  PI Lookup  
Hint: Including CO-PI will result in slower searches.  
Include CO-PI:   
Organization:  Organization Lookup  
State:   
ZIP Code:   
Country:   
Hint: Historical data is from prior to 1976. This data may not be as complete as recent data.  
Historical Awards:   
Active Awards Only:   
Expired Awards Only:   
Search Reset

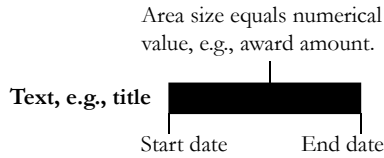
126



## Biomedical Funding Profile (section 5.2.4)

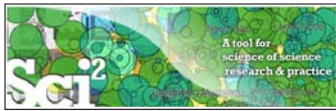


### Horizontal Bargraph

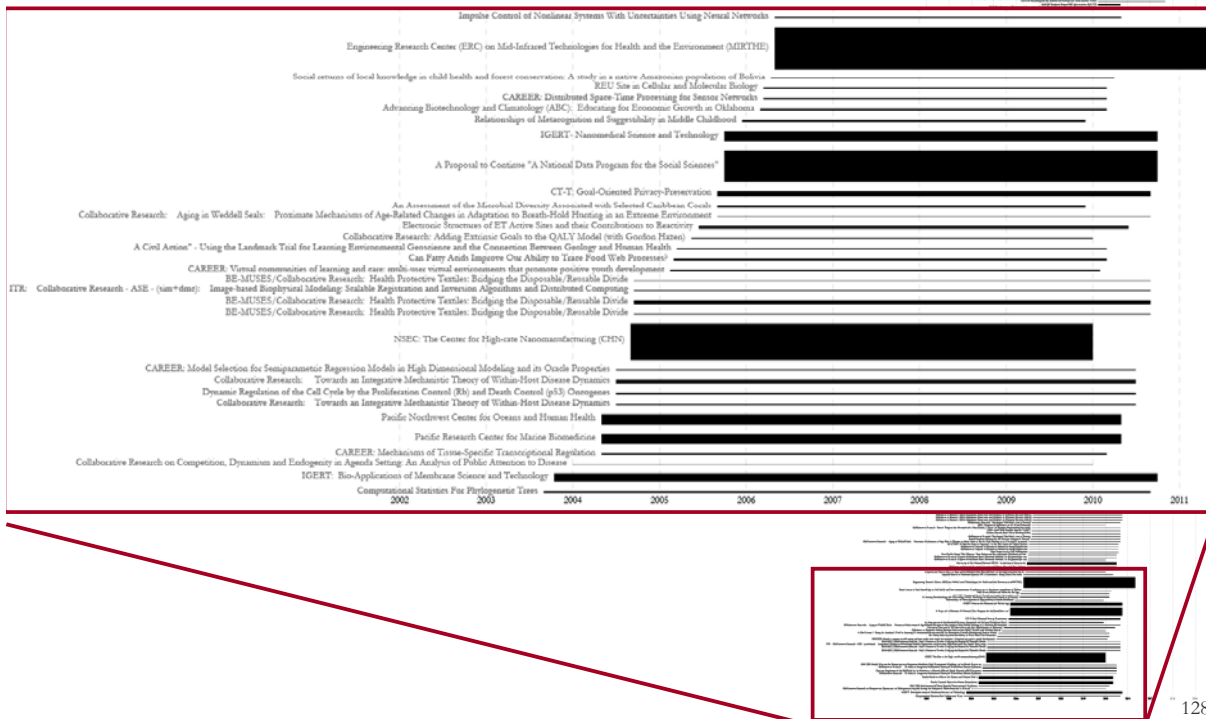


### Top-10 grants with highest \$Awarded to Date:

Title	NSF Org.	Program(s)	PI	State	Organization	\$ Awarded to Date
University of New Mexico/Harvard PREM: Leadership in Biomedical Research	DMR	PREMI	MATERIALS Lopez, Gabriel	NM	University of New Mexico	2,037,500
TC: Large: Trustworthy Information Systems for Healthcare	CNS	TRUSTWORTHY	Information Systems Kotz, David	NH	Dartmouth College	2,999,999
IGERT: Nanomedical Science and Technology	DGE	IGERT FULL	PF Sridhar, Srinivas	MA	Northeastern University	3,323,891
IGERT: Bio-Applications of Membrane Science and Technology	DGE	HUMAN RESOURCE	Applications of Membrane Science and Technology Fried, Joel	OH	University of Cincinnati Main Campus	3,644,410
Pacific Research Center for Marine Biomedicine	OCE	CHEMICAL OCEAN	Law, Edward	HI	University of Hawaii	3,816,943
Pacific Northwest Center for Oceans and Human Health	OCE	CHEMICAL OCEAN	Human Health Faustman, Elaine	WA	University of Washington	4,026,968
A Proposal to Continue "A National Data Program for the Social Sciences"	SES	SCIENCE & ENGINEERING	Smith, Tom	IL	National Opinion Research Center	5,835,140
A Proposal to Continue "A National Data Program for the Social Sciences"	SES	SCIENCE & ENGINEERING	Davis, James	IL	National Opinion Research Center	10,053,668
NSEC: The Center for High-rate Nanomanufacturing (CHN)	EEC	Studies of Policy	Manufacturing Busnaina, Ahmed	MA	Northeastern University	13,047,758
Engineering Research Center (ERC) on Mid-Infrared Technologies for Health and the Environment (MIRTHE)	EEC	COLLABORATIVE	Technologies for Health and the Environment Gmachl, Claire	NJ	Princeton University	13,681,994



## Biomedical Funding Profile (section 5.2.4)





## Biomedical Funding Profile of NSF (NSF Data)

(section 5.2.4)

### Bimodal Network of NSF Organization to Program(s)

Extract Directed Network was selected.

Source Column: NSF Organization

Text Delimiter: |

Target Column: Program(s)

Nodes: 167

Isolated nodes: 0

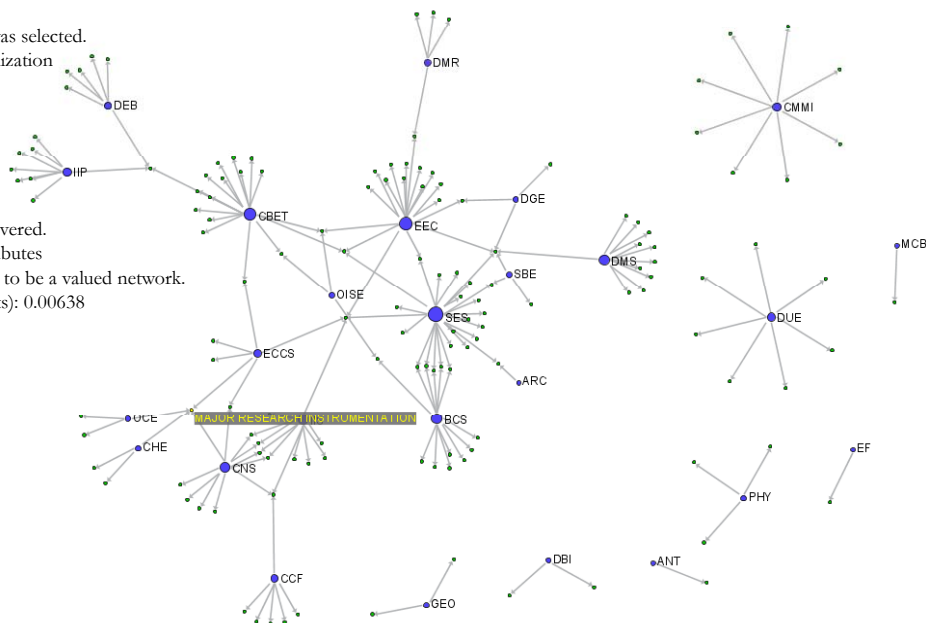
Edges: 177

No parallel edges were discovered.

Did not detect any edge attributes

This network does not seem to be a valued network.

Density (disregarding weights): 0.00638



129



## Mapping the Field of RNAi Research (SDB Data)

(section 5.2.7)

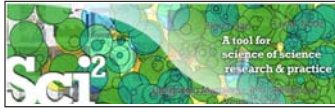
RNAi	
<b>Time frame:</b>	1865-2008
<b>Region(s):</b>	Miscellaneous
<b>Topical Area(s):</b>	RNAi
<b>Analysis Type(s):</b>	Co-Author Network, Patent-Citation Network, Burst Detection

How many papers, patents, and funding awards exist on a specific topic?

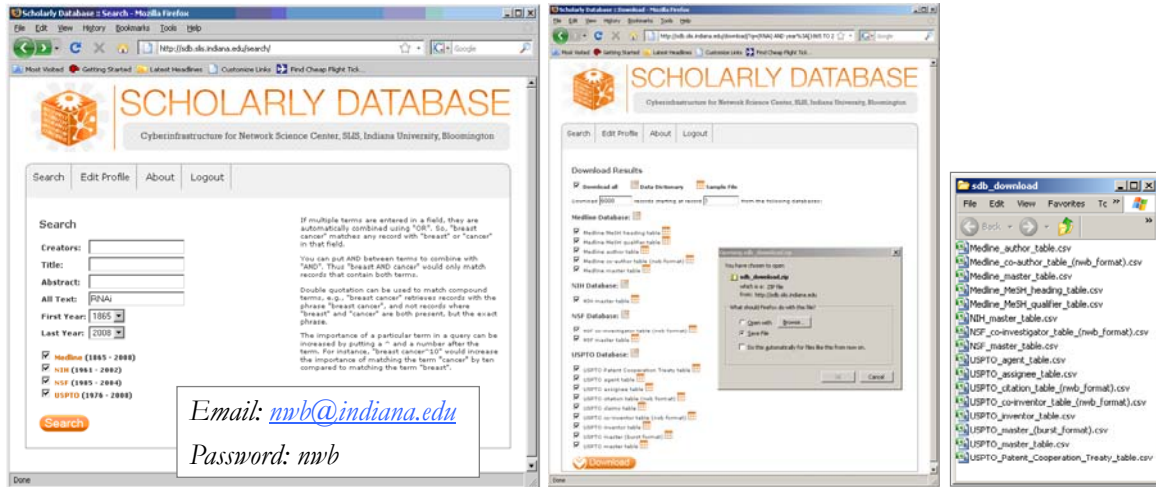
Here we selected research on RNA interference (RNAi) is a system within living cells that helps to control which genes are active and how active they are.

The data for this analysis comes from a search of the Scholarly Database (SDB) (<http://sdb.slis.indiana.edu/>) for “RNAi” in “All Text” from MEDLINE, NSF, NIH and USPTO. A copy of this data is available in ‘\*yoursci2directory\*/sampledata/scientometrics/sdb/RNAi’. The default export format is .csv, which can be loaded in the Sci2 Tool directly.

130



## Mapping the Field of RNAi Research (SDB Data) (section 5.2.7)



Email: [mwb@indiana.edu](mailto:mwb@indiana.edu)  
Password: mwb

The **Scholarly Database** at Indiana University provides free access to 23,000,000 papers, patents, and grants. Since March 2009, users can also download networks, e.g., co-author, co-investigator, co-inventor, patent citation, and tables for burst analysis. For more information and to register, visit <http://sdb.slis.indiana.edu>.



## Mapping the Field of RNAi Research (SDB Data) (section 5.2.7)

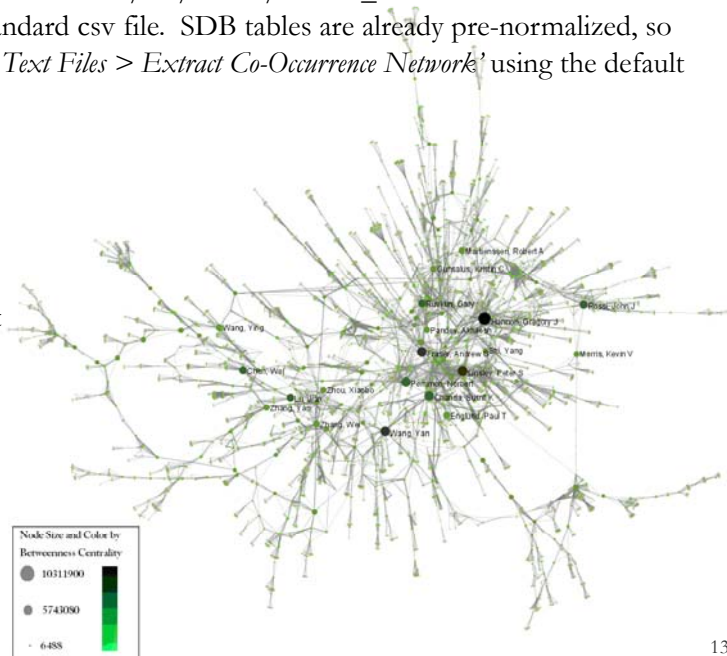
### Co-Author Network

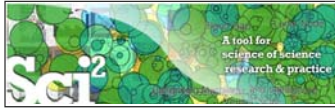
Load *\*yoursci2directory\*/sampledata/scientometrics/sdb/RNAi/Medline\_co-author\_table\_(mwb\_format).csv* as a standard csv file. SDB tables are already pre-normalized, so now simply run *Data Preparation > Text Files > Extract Co-Occurrence Network* using the default parameters.

*Network Analysis Toolkit (NAT):*  
21,578 nodes with 131 isolates,  
77,739 edges.

Extract only the largest component by running *Analysis > Networks > Unweighted and Undirected > Weak Component Clustering.*

Visualize with *GUESS* using *Layout > GEM*.  
Use a custom python script to color and size the network.





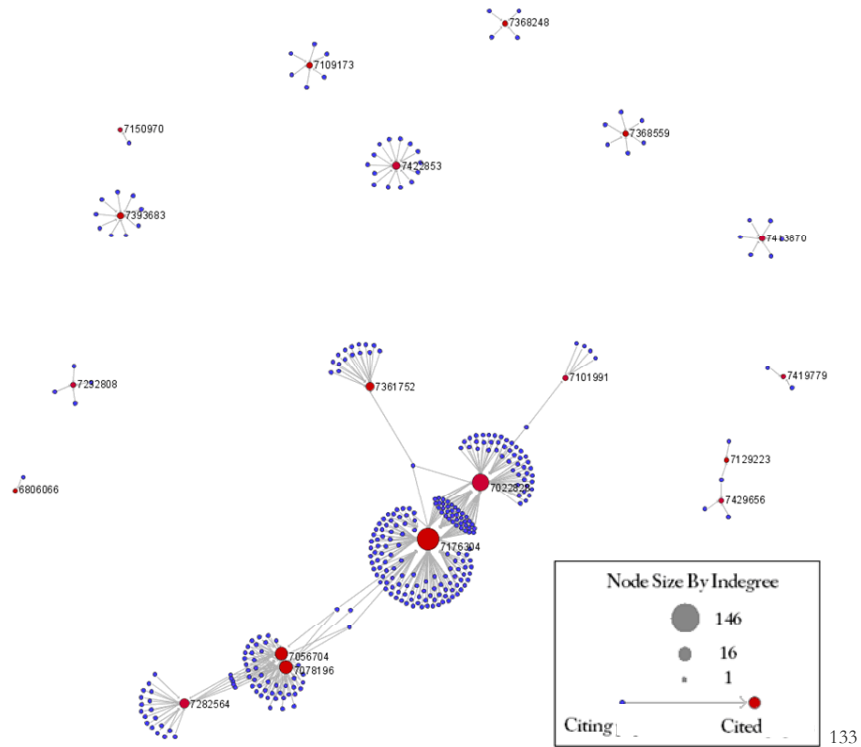
## Mapping the Field of RNAi Research (SDB Data) (section 5.2.7)

### Patent Citation Network

To visualize the citation patterns of patents on RNAi, load

```
*yoursci2directory*/sampledata/scientometrics/sdb/RNAi/USPTO_citation_table_(nwb_format).csv'
```

as a standard csv file and follow the instructions in the tutorial.



## Mapping the Field of RNAi Research (SDB Data) (section 5.2.7)

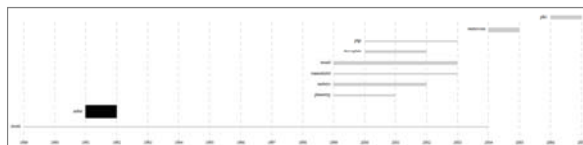
### Topic Bursts

Load `*yoursci2directory*/sampledata/scientometrics/sdb/RNAi/Medline_master_table.csv'`. This table includes full records of MEDLINE papers, and can be used to find bursting terms from MEDLINE abstracts dealing with RNAi.

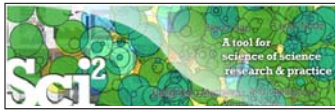
Load the file as a standard csv and run `Preprocessing > Topical > Normalize Text` with the default separator and the “abstract” box checked. Run `Analysis > Topical > Burst Detection` with “date\_cr\_year” in the Date Column and “abstract” in the Text Column, leaving the rest of the values default.

Right click on “Burst detection analysis (date\_cr\_year, abstract): maximum burst level 1” in the Data Manager and view the file. There are more words than can easily be viewed with the horizontal bar graph, so sort the list by “Strength” and prune all but the strongest 10 words. Save the file as a new .csv and load it into the Sci2 Tool as a standard csv file.

Select the new table in the data manager and visualize it using `Visualize > Temporal > Horizontal Bar Graph`.







## Geo USPTO (SDB Data) (section 5.3.1)

usptoInfluenza.csv	
<b>Time frame:</b>	1865-2008
<b>Region(s):</b>	Miscellaneous
<b>Topical Area(s):</b>	Influenza
<b>Analysis Type(s):</b>	Geospatial Analysis

The file ‘*usptoInfluenza.csv*’ was generated with an SDB search for patents containing the term “Influenza”, and was heavily modified to produce a simple geographic table. Load it using ‘*File > Load > sampledata > geo > usptoInfluenza.csv*’ and then select ‘Standard csv format’. See the data format in Figure 5.30 (left). Once loaded, select the dataset in data manager and click ‘*Visualization > Geo Map (Circle Annotation Style)*’, inputting the parameters Figure 5.30 (right). The tool will output a PostScript visualization which can be viewed using GhostView (see section [2.4 Saving Visualizations for Publication](#) and Figure 5.31).

135



## Geo USPTO (SDB Data) (section 5.3.1)

The screenshot shows the Sci² Tool interface with the following components:

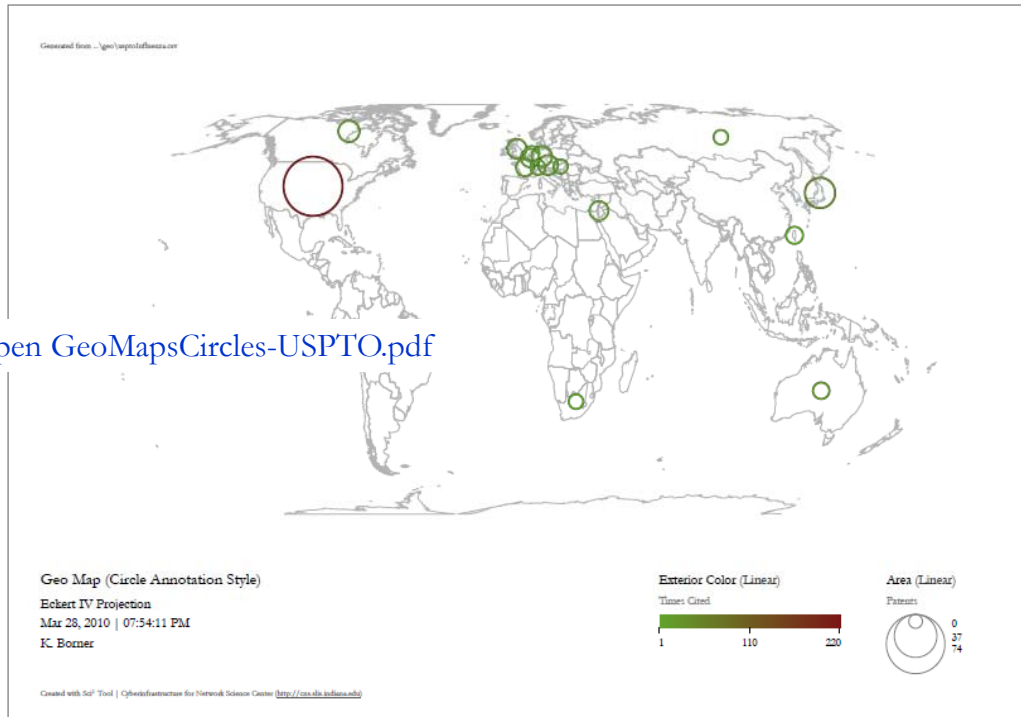
- Data Manager:** Shows the loaded CSV file: `C:\Users\User\Desktop\scipolicy\sampladata\geo\usptoInfluenza.csv` and the generated PostScript file: `C:\Users\User\Desktop\scipolicy\sampladata\geo\usptoInfluenza.csv`.
- Console:** Shows the selected visualization options: `Geo Map (Circle Annotation Style) was selected.`, `Author(s): Joseph R. Biberstine`, `Implementer(s): Joseph R. Biberstine`, and `Integrator(s): Joseph R. Biberstine`.
- Geo Maps (circles) dialog box:** Shows the map settings:
  - Map: Countries
  - Author Name: K. Borner
  - Latitude: Latitude
  - Longitude: Longitude
  - Size Circles By: Patents
  - Size Scaling: Linear
  - Color Circle Exteriors By: Times Cited
  - Exterior Color Scaling: Linear
  - Exterior Color Range: Green to Red
  - Color Circle Interiors By: None (no inner color)
  - Interior Color Scaling: Linear
- Microsoft Excel:** Shows the data table with columns: Country, Latitude, Longitude, Patents, Times Cited.

	A	B	C	D	E	F
	Country	Latitude	Longitude	Patents	Times Cited	
1	Hungary	47.16116	19.50496	0.083333	4	
2	Belgium	50.50099	4.47677	3.017857	11	
3	Germany	51.09094	10.45424	4.783333	4	
4	Canada	62.35873	-96.8821	5.530266	21	
5	Russia	59.46148	108.8318	0.266667	2	
6	Austria	47.69651	13.34577	4.2	17	
7	Netherlan	52.10809	5.33033	1	2	
8	Switzerlan	46.81309	8.22414	0.507576	6	
9	Taiwan	23.59975	121.0238	2	3	
10	Australia	-24.9162	133.3931	1.617857	23	
11	United Sta	39.83	-96.58	73.99639	220	
12	France	46.71245	1.71932	2.201166	9	
13	South Afric	-28.4832	24.67699	0.333333	1	
14	Japan	37.4876	139.8363	15.99167	39	
15	Israel	31.3893	35.36124	3.5	3	
16	United Kin	54.31392	-2.23218	3.85	12	
17						
18						
19						

136



## Geo USPTO (SDB Data) – Circle Coding (section 5.3.1)



Open GeoMapsCircles-USPTO.pdf

137



## Geospatial Maps – Region Coding

Sci² Tool

File Preprocessing Modeling Analysis Visualization Scientometrics Help

Console  
Geo Map (Colored-Region Annotation Style) was selected.  
Author(s): Joseph R. Biberstine  
Implementer(s): Joseph R. Biberstine  
Integrator(s): Joseph R. Biberstine  
Documentation:

Scheduler

Microsoft Excel - NWB-Session-xxx-Session-185659073315...  
File Edit View Insert Format Tools Data Window Help  
Adobe PDF

Data Manager  
CSV file: C:\Users\User\Desktop\scipolicy\sampladata\geo\usptoInfluenza.csv  
PostScript: CSV file: C:\Users\User\Desktop\scipolicy\sampladata\geo\usptoInfluenza.csv  
PostScript: CSV file: C:\Users\User\Desktop\scipolicy\sampladata\geo\usptoInfluenza.csv.2

	A	B	C	D	E	F
1	Country	Latitude	Longitude	Patents	Times Cited	
2	Hungary	47.16116	19.50496	0.083333	4	
3	Belgium	50.50099	4.47677	3.017857	11	
4	Germany	51.09084	10.45424	4.783333	4	
5	Canada	62.35973	-96.5921	5.539286	21	
6	Russia	59.46148	108.63219	0.268667	2	
7	Austria	47.59651	13.34377	4.2	17	
8	Netherlands	52.10809	5.33033	1	2	
9	Switzerland	46.81309	8.22414	0.507576	6	
10	Taiwan	23.59975	121.0238	2	3	
11	Australia	-24.9162	133.3931	1.617857	23	
12	United Sta	39.83	-98.58	73.99839	220	
13	France	46.71245	1.71832	2.201166	9	
14	South Afric	-28.4832	24.67699	0.333333	1	
15	Japan	37.4876	139.8383	15.99167	39	
16	Israel	31.3693	35.36124	3.5	3	
17	United Kgn	54.31392	-2.23218	3.85	12	
18						
19						

Geo Maps (region coloring)  
Creates a map with colored-region annotations. Regions are identified and colored according to columns in the input table. The table data can be log-scaled before processing.

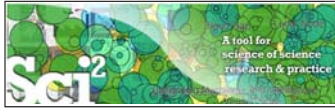
Map: Countries  
Projection: Mercator  
Author Name: Katy Borner  
Region Name: Country  
Color By: Patents  
Color Scaling: Linear  
Color Range: Green to Red

Geo Maps (region coloring)  
Creates a map with colored-region annotations. Regions are identified and colored according to columns in the input table. The table data can be log-scaled before processing.

Map: Countries  
Projection: Mercator  
Author Name: Katy Borner  
Region Name: Country  
Color By: Times Cited  
Color Scaling: Logarithmic  
Color Range: Green to Red

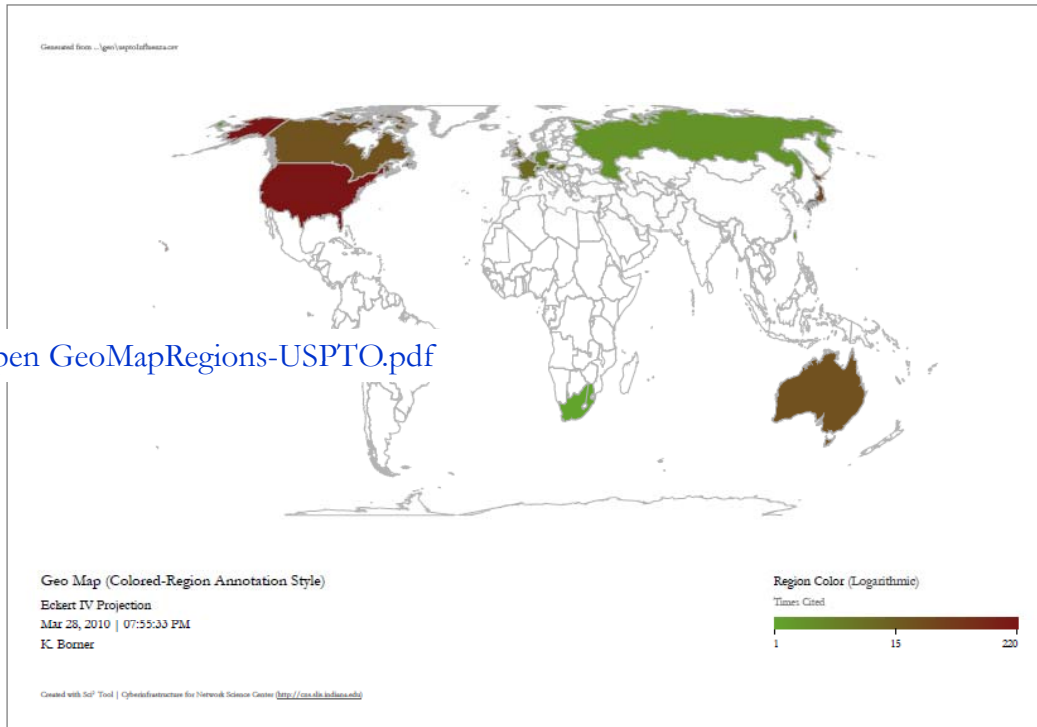
OK Cancel

138



# Geo USPTO (SDB Data) – Region Coding (section 5.3.1)

Open GeoMapRegions-USPTO.pdf



## Science of Science Cyberinfrastructure — P O R T A L —

Provided by the [Cyberinfrastructure for Network Science Center](#) at Indiana University.

**Introduction**  
E. O. Wilson writes in *Consilience: The Unity of Knowledge* (1998): "Features that distinguish science from pseudoscience are repeatability, economy, mensuration, heuristics, and consilience." Please see Börner's [recent presentation](#) at the *A Deeper Look at the Visualization of Scientific Discovery* NSF Workshop for a general introduction of the needs and the resources provided here.

**Needs Analysis**  
As part of the "TLS: Towards a Macroscopic for Science Policy Decision Making" NSF SBE-0738111 award, interviews with science policy makers are conducted to identify what science of science research results and tools might be most desirable and effective. So far, 30 formal, one-hour interviews have been conducted with science policy makers at university campus level, program officer level, and division director level for governmental, state, and private foundations. Data compilation will start in October 2008 and resulting report can be ordered by sending a request to Mark Price ([maaprice@indiana.edu](mailto:maaprice@indiana.edu)).

**Conceptualization of Science**  
A science of science requires a theoretically grounded and practically useful conceptualization of the structure and evolution of science. A special journal issue entitled "*Science of Science: Conceptualizations and Models of Science*" edited by [Katy Börner](#), Indiana University & [Andrea Scharnhorst](#), Royal Netherlands Academy of Arts and Sciences invites contributions on this topic. It will be published in the *Journal of Informetrics* 3(1) in January 2009.

**Scholarly Database**  
The [Scholarly Database \(SDB\)](#) at Indiana University aims to serve researchers and practitioners interested in the analysis, modeling, and visualization of large-scale scholarly datasets. The database currently provides access to over 20 million papers, patents and grants. Resulting datasets can be downloaded in bulk. Register for free access at <https://sdb.slis.indiana.edu/>.

**Cyberinfrastructures**  
The Scientometrics filling of the [Network Workbench \(NWB\) Tool](#) provides a unique distributed, shared resources environment for large-scale network analysis, modeling, and visualization. Thomson Scientific/ISI, Scopus and Google Scholar data, EndNote and Bibtext files, or NSF awards can be read and diverse networks can be extracted and studied. Download [User Manual with focus on Scientometrics](#).

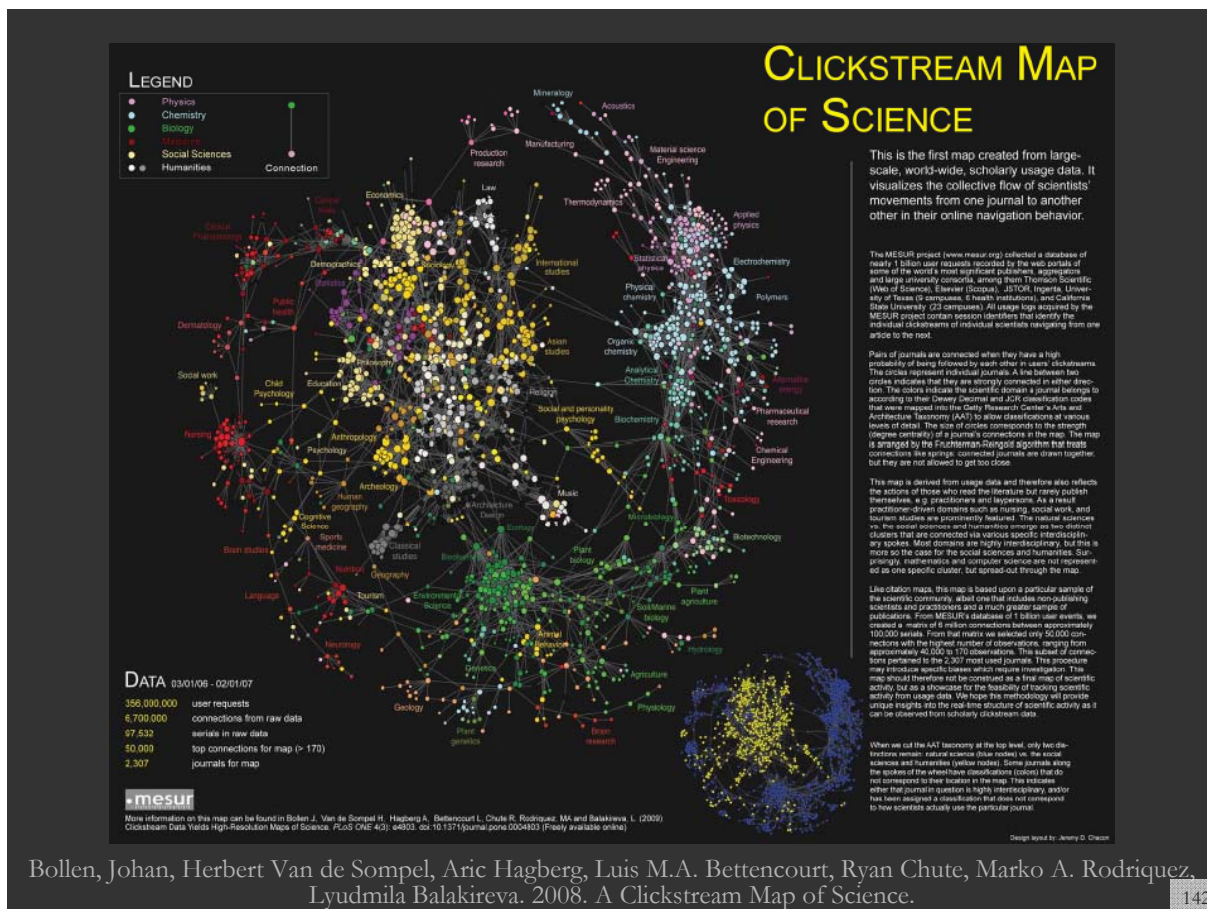
<http://sci.slis.indiana.edu>



## Overview

- Introduction to the Sci<sup>2</sup> Tool
- Demo and hands-on data analysis and visualization by participants
- Outlook
- Overview of validation approaches for science studies

141



142



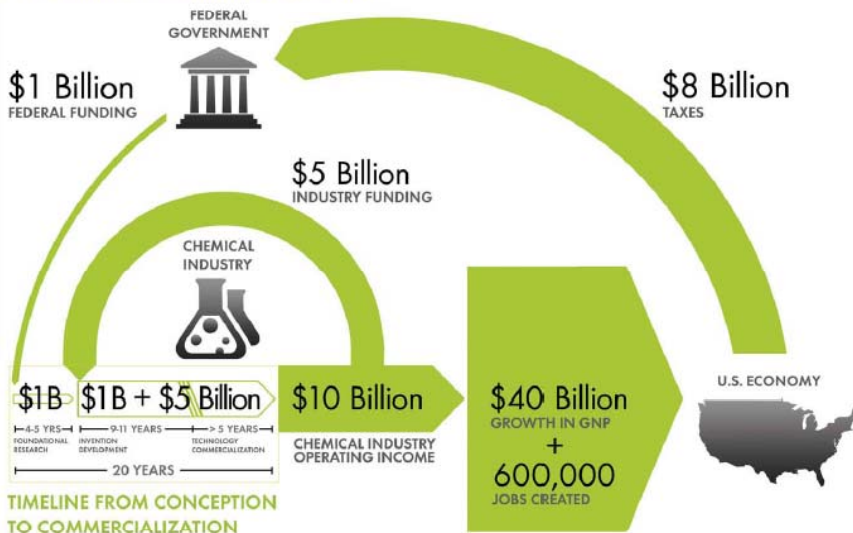
# Chemical Research & Development Powers the U.S. Innovation Engine

Macroeconomic Implications of Public and Private R&D Investments in Chemical Sciences



has provided the U.S. Congress and government policy makers with important results regarding the impact of Federal Research & Development (R&D) investments on U.S. innovation and global competitiveness through its commissioned 5-year two phase study. To take full advantage of typically brief access to policy makers, CCR developed the graphic below as a communication tool that distills the complex data produced by these studies in direct, concise and clear terms.

## INVESTMENT IN CHEMICAL SCIENCE R&D



The design shows that an input of \$1B in federal investment, leveraged by \$5B industry investment, brings new technologies to market and results in \$10B of operating income for the chemical industry, \$40B growth in the Gross National Product (GNP) and further impacts the US economy by generating approximately 600,000 jobs, along with a return of \$8B in taxes. Additional details, also reported in the CCR studies, are depicted in the map to the left. This map clearly shows the two R&D investment cycles; the shorter industry investment at the innovation stage to commercialization cycle; and the longer federal investment cycle which begins in basic research and culminates in national economic and job growth along with the increase tax base that in turn is available for investment in basic research.

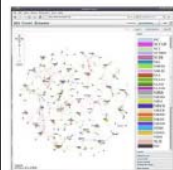
Council for Chemical Research. 2009. Chemical R&D Powers the U.S. Innovation Engine. Washington, DC. Courtesy of the Council for Chemical Research.

## A Topic Map of NIH Grants 2007

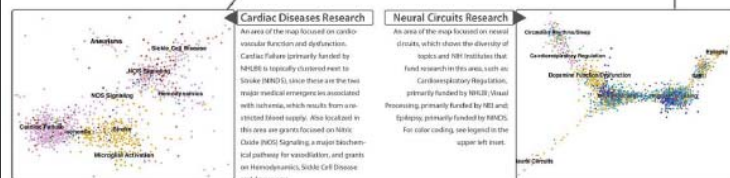
Bruce W. Herr II (Chalklabs & IU), Gully Burns (IS), David Newman (UCI), Edmund Talley (NIH)



The National Institutes of Health (NIH) is organized as a multitude of Institutes and Centers whose missions are primarily focused on distinct diseases. However, disease etiologies and therapies blur scientific boundaries, and thus there is tremendous overlap in the kinds of research funded by each institute. This creates a daunting landscape for decisions on research directions, funding allocations, and policy formulations. Shown here is devised an interactive topic map for navigating this landscape, online at [www.nih.gov/td](http://www.nih.gov/td). Institute abbreviations can be found at [www.nih.gov/td](http://www.nih.gov/td).



Topic modeling, a statistical technique that automatically learns semantic categories, was applied to assess projects in terms used by researchers to describe their work, without the biases of keywords or subject headings. Grant similarities were derived from their topic mixtures, and grants were then clustered on a two-dimensional map using a force-directed simulated annealing algorithm. This analysis creates an interactive environment for assessing grant relevance to research categories and to NIH institutes in which grants are localized.



### National Cancer Institute (NCI)

- TOP 10 TOPICS
1. Oncology/Clinical Trials
  2. Cancer Treatment
  3. Cancer Therapy
  4. Carcinogenesis
  5. Risk Factor Analysis
  6. Cancer Chemotherapy
  7. Metastasis
  8. Leukemia
  9. Prevention/Prognosis
  10. Cancer Chemoprevention

### National Institute of General Medical Sciences (NIGMS)

- TOP 10 TOPICS
1. Bioactive Organic Synthesis
  2. X-ray Crystallography
  3. Protein-DB
  4. Computational Models
  5. Yeast Biology
  6. Metabolism
  7. Enzymatic Mechanisms
  8. Protein Complexes
  9. Invertebrate/Zebrafish Genetics
  10. Cell Division

### National Heart, Lung, and Blood Institute (NHLBI)

- TOP 10 TOPICS
1. Cardiac Failure
  2. Pulmonary Injury
  3. Genetic Linkage Analysis
  4. Cardiovascular Disease
  5. Atherosclerosis
  6. Hemostasis
  7. Blood Pressure
  8. Arteriosclerosis, Atherosclerosis
  9. Gene Association
  10. Lipoproteins

### National Institute of Mental Health (NIMH)

- TOP 10 TOPICS
1. Mood Disorders
  2. Schizophrenia
  3. Behavioral/Intervention Studies
  4. Mental Health
  5. Depression
  6. Cognitive Behavior Therapy
  7. AIDS Prevention
  8. Genetic Linkage Analysis
  9. Adolescence
  10. Child/Adol

Herr II, Bruce W., Gully Burns, David Newman, Edmund Talley. 2007. A Topic Map of NIH Grants 2007. Bloomington, IN.





Science Maps in "Expedition Zukunft" science train visiting 62 cities in 7 months  
 12 coaches, 300 m long  
 Opening was on April 23<sup>rd</sup>, 2009 by German Chancellor Merkel  
<http://www.expedition-zukunft.de>



### Validating Science Maps

Boyack, Kevin W., Klavans, Richard & Börner, Katy. (2005). *Mapping the Backbone of Science*. *Scientometrics*. Vol. 64(3), 351-374.

Eight alternative measures of journal similarity were applied to a data set of 7,121 journals covering over 1 million documents in the combined Science Citation and Social Science Citation Indexes. For each journal similarity measure we generated two-dimensional spatial layouts using the force-directed graph layout tool, VxOrd. Next, mutual information values were calculated for each graph at different clustering levels to give a measure of structural accuracy for each map. The best co-citation and inter-citation maps according to local and structural accuracy were selected and are presented and characterized. These two maps are compared to establish robustness. The inter-citation map is then used to examine linkages between disciplines.

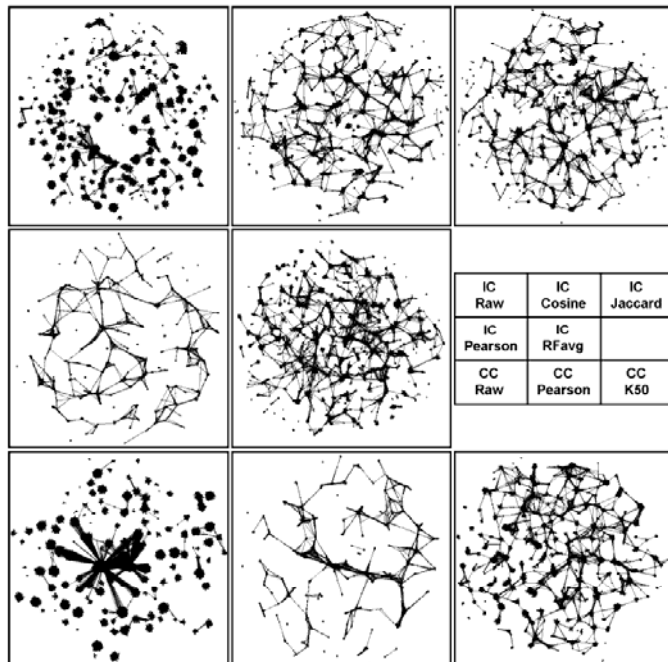


Table 1. Summary of validation results for maps based on eight similarity measures.

Measure	Local accuracy @ 95% coverage <sup>1</sup>	Scalability <sup>1</sup>	Z-score for 200 clusters	Clustering (qualitative)
IC-Raw	60.1%	High	360.0	Too few, loose
IC-Cosine	80.2%	High	381.3	Good balance
IC-Jaccard	79.5%	High	387.1	Good balance
IC-Pearson	71.7%	Low	386.5	Too tight
IC-RFavg	80.2%	High	373.3	Good balance
CC-Raw	25.6%	High	294.9	Too few, loose
CC-Pearson	65.3%	Low	377.0	Too tight
CC-K50	71.4%	High	376.6	Good balance



Figure 4. Map of science generated using the IC-Jaccard similarity measure. The map is comprised of 7,121 journals from year 2000. Large font size labels identify major areas of science. Small labels denote the disciplinary topics of nearby large clusters of journals

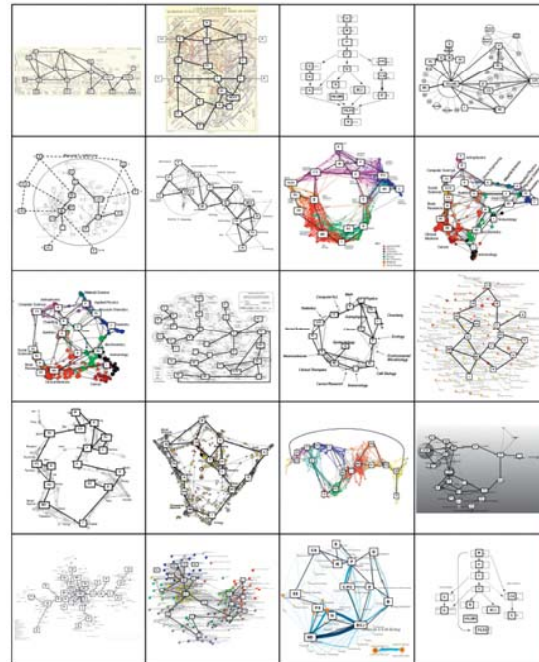


## Validating Science Maps

Klavans, R., & Boyack, K. W. (2009). *Toward a consensus map of science. Journal of the American Society for Information Science and Technology*, 60(3), 455-476.

A consensus map of science is generated from an analysis of twenty existing maps of science. These twenty maps occur in three basic forms: hierarchical, centric, and non-centric (or circular). The consensus map, generated from consensus edges that occur in at least half of the input maps, emerges in a circular form. The ordering of areas is as follows: mathematics is (arbitrarily) placed at the top of the circle, and is followed clockwise by physics, physical chemistry, engineering, chemistry, earth sciences, biology, biochemistry, infectious diseases, medicine, health services, brain research, psychology, humanities, social sciences, and computer science.

The circular map of science is found to have a high level of correspondence with the twenty existing maps, and has a variety of advantages over hierarchical and centric forms.



149

**Table 1: Characteristics of twenty comprehensive maps of science. Abbreviations SC, SS, AH, and PR refer to Thomson Scientific's Science, Social Science, Arts & Humanities, and Proceedings Citation databases, respectively.**

Researcher(s) & Reference	Map Name	Method	Elements	# Clust	Database & Year	Form
(Bernal, 1939)	Bernal	Expert		14, 110		Hierarchical
(Ellingham, 1948)	Ellingham	Expert		13, 51, 130		Hierarchical & Non-centric
(Balaban & Klein, 2006)	Balaban-I	Expert	16 fields	16		Hierarchical & Centric
(Griffith, Small, Stonehill, & Dey, 1974)	Small74	Reference papers	1,150 pap	41	SC, 1972 Q1	Centric
(Small & Garfield, 1985)	Small85	Reference papers	~11,000 pap	51	SC+SS, 1983	Hierarchical & Centric
(Small, 1999)	Small99	Reference papers	36,720 pap	35	SC+SS, 1995	Hierarchical
(Klavans & Boyack, 2008) <sup>a</sup>	KB-Para	Reference papers	800k pap	776	SC+SS, 2003	Non-centric
(Klavans & Boyack, 2007)	KB06-TS	Reference papers	1.9M pap	283	SC+SS, 2004	Non-centric
(Klavans & Boyack, 2007)	KB06-SC	Reference papers	2.1M pap	554	Scopus, 2004	Non-centric
(Bassecouard & Zitt, 1999)	B-Z	Journals	~2,000 jnl	29	SC/JCR, 1993	Hierarchical & Centric
Klavans, unpublished, 2002	K02	Journals	5,647 jnl	69	SC+SS+AH, 2000	Non-centric
(Boyack, Klavans, & Börner, 2005)	Backbone	Journals	7,121 jnl	205	SC+SS, 2000	Non-centric
(Boyack et al., 2009)	BBK02-S	Journals	7,227 jnl	671	SC+SS, 2002	Non-centric
(Boyack, 2009)	B03-ST	Journals	8,667 jnl	852	SC+SS+PR, 2003	Non-centric
(Klavans, Boyack, & Patek, 2008) <sup>b</sup>	UCSD	Journals	16,235 jnl	554	SC/SS/AH + Scopus, 2001-05	Non-centric
(Rosvall & Bergstrom, 2008) <sup>c</sup>	Rosvall	Journals	6,116 jnl	87	SC+SS, 2004	Non-centric
(Moya-Aneón et al., 2004)	Scimago-I	Journal categories	25 categ	25	SC+SS+AH, 2000 Spanish papers	Non-centric
(Moya-Aneón et al., 2007) <sup>d</sup>	Scimago-II	Journal categories	219 categ	219	SC+SS+AH, 2002	Centric
(Leydesdorff & Rafols, 2008) <sup>e</sup>	L-R	Journal categories	6,164 jnl; 172 categ	172	SC, 2006	Mixed
(Balaban & Klein, 2006)	Balaban-II	Course prerequisites		11	Texas A&M undergraduate	Centric

150

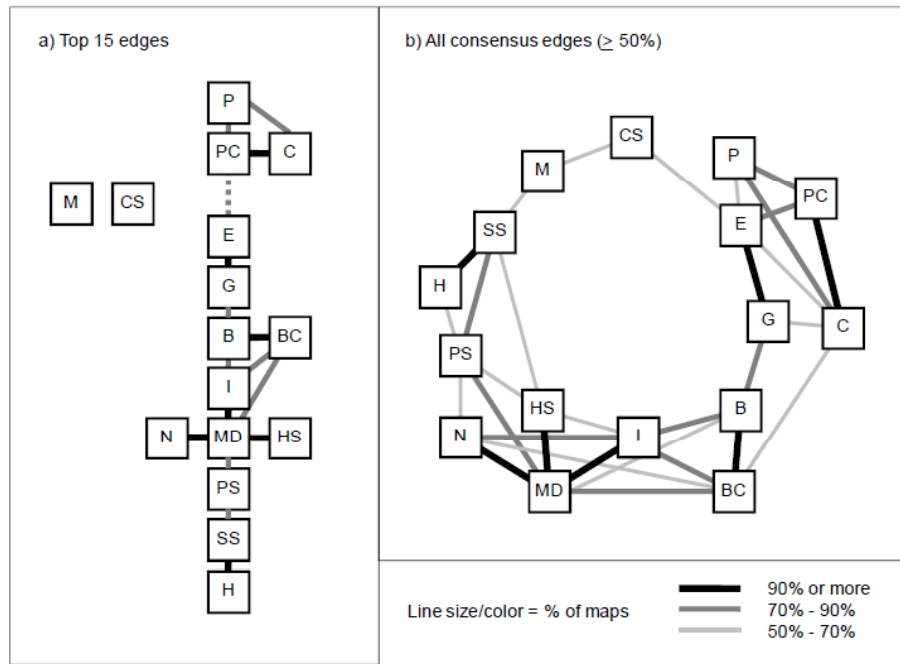


Figure 5: Two-dimensional consensus maps of science from twelve of the twenty input maps, excluding the eight input maps from Klavans, Boyack, and Börner.



## Validating Science Maps

*Accuracy of Models for Mapping the Medical Sciences*  
 Kevin W. Boyack, Richard Klavans, SciTech Strategies Inc.  
 Katy Börner, Russell J. Dubon, Nianli Ma, Indiana University,  
 Bob Schijvenaars, Aaron Sorensen, Collexis Holdings Inc.,  
 André Skupin, San Diego State University

This project aims to provide a highly accurate interactive map of medical research that can be easily used by both technical and non-technical users. Phase I of this project compares and determines the relative accuracies of maps of medical research based on commonly used text-based and citation-based similarity measures at a scale of over two million documents.

All work is documented in real time at <http://sci.slis.indiana.edu/sts> and at a level of detail that supports the exact replication of work.

The screenshot shows the project website with a map of the United States at the top. Below the map, there is a section titled "Accuracy of Models for Mapping the Medical Sciences" with an abstract, team list, project history, datasets, and analysis result data. The analysis result data includes links for various models and data sources.

**Abstract**  
 This project aims to provide a highly accurate interactive map of medical research that can be easily used by both technical and non-technical users. Phase I of this project compares and determines the relative accuracies of maps of medical research based on commonly used text-based and citation-based similarity measures at a scale of over two million documents.

**Team**  
 The project is led by Richard Klavans, Inc. (<http://www.scitechstrategies.com>) in collaboration with the CyberInfrastructure for Network Science center at Indiana University (<http://sci.slis.indiana.edu>). There are subcontractors to different researchers and one company. The full team comprises:  
 • Kevin W. Boyack, Richard Klavans, SciTech Strategies Inc.  
 • Katy Börner, Russell J. Dubon, Nianli Ma, Indiana University  
 • Bob Schijvenaars, Aaron Sorensen, Collexis Holdings Inc.  
 • André Skupin, San Diego State University

The following people, although not part of the formal team, will also contribute to the project:  
 • Edmund Tolley, National Institute of Health  
 • Dean Stensrud, University of California, Irvine

**Project History:** [sci.slis.indiana.edu](http://sci.slis.indiana.edu)

**Datasets**  
 It was decided that all work will be documented in real time and at a level of detail that supports the exact replication of work. All documentation will have access to this documentation as well as to interactive data results. Where the desktop data cannot be made available, all Medline based desktop data will be made freely available from this page. Data compilation and statistics are documented in [Data Documentation](#).

**Raw data**

List of PDBs	<a href="#">slis-pdb4.txt.gz</a>
List of stop words	<a href="#">slis-stop-words.txt.gz</a>

**Analysis Input Data**

File/Abstract term adjacency list	<a href="#">slis-text-adj.txt</a> (40MB)
MSH adjacency list	<a href="#">slis-mesh-adj.txt</a> (97MB)

Analysis results will also be made available from this site. They will comprise:

**Analysis Result Data**

File/Abstract term analysis	
Coincidence	<a href="#">slis-terms.txt</a>
LSA	<a href="#">slis-LSA.txt</a>
Topic model (L1)	<a href="#">slis-TA-topics-l1.txt</a>
Topic model (CC)	<a href="#">slis-TA-topics-cc.txt</a>
Self-organizing maps (SOM)	<a href="#">slis-TA-som.txt</a>
Collexis	<a href="#">slis-TA-collexis.txt</a>
MSH analysis	
Coincidence	<a href="#">slis-mesh.txt</a>
LSA	<a href="#">slis-mesh-LSA.txt</a>
Topic model (L1)	<a href="#">slis-mesh-topics-l1.txt</a>
Topic model (CC)	<a href="#">slis-mesh-topics-cc.txt</a>
Self-organizing maps (SOM)	<a href="#">slis-mesh-som.txt</a>
Collexis	<a href="#">slis-mesh-collexis.txt</a>
Collexis (full engine on raw MEDLINE data)	
Collexis full analysis	<a href="#">slis-collexis-full.txt</a>

**Validation**  
 coming soon

**Acknowledgements**  
 This project is funded by NIH SEIR Contract HHS(N0180000002).

**NATIONAL INSTITUTES OF HEALTH**





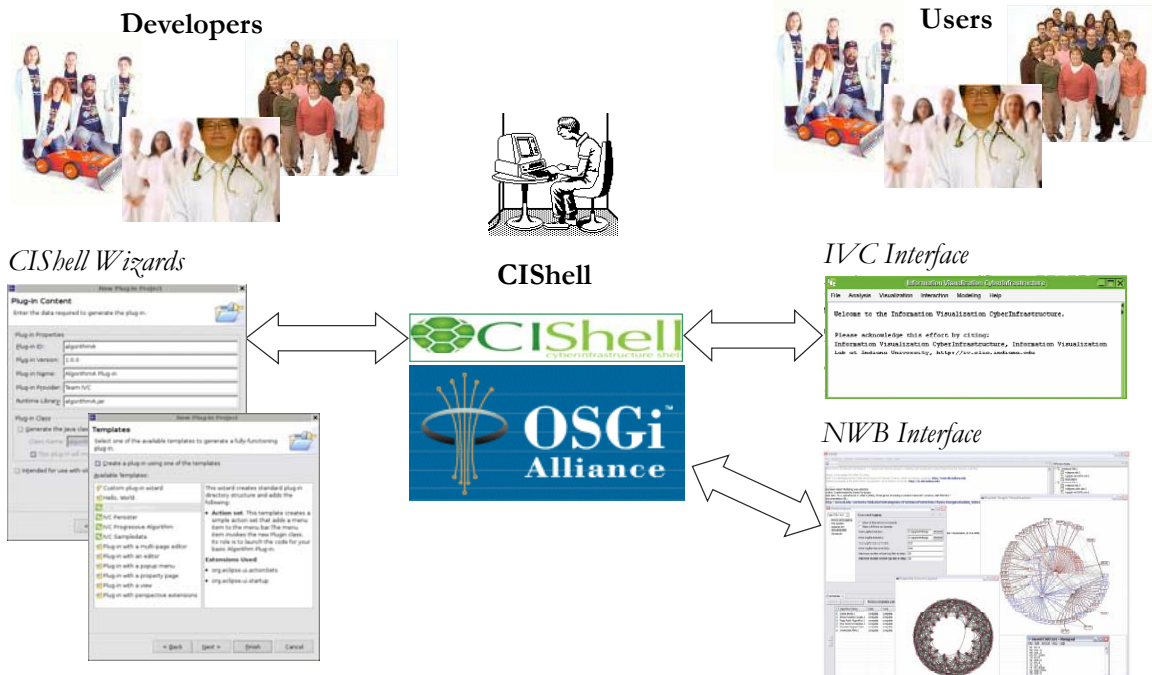
## Overview

- **Macrosopes and the Changing Scientific Landscape** 15 mins
  - **Introduction to network science with sample maps and insights** 15 mins
  - **Introduction to the Network Workbench Tool** 15 mins
  - **Demo and hands-on data analysis and visualization by participants** 60 mins
  - **Introduction to science studies with sample maps and insights** 15 mins
  - **Introduction to the Sci<sup>2</sup> Tool** 15 mins
  - **Demo and hands-on data analysis and visualization by participants** 60 mins
  - **Overview of validation approaches for science studies**
  - **Plug-and-play tool design using OSGi/CIshell** 30 mins
  - **Scholarly Marketplaces** 15 mins
- 240 mins**

153



## CIShell – Serving Non-CS Algorithm Developers & Users



154



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

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

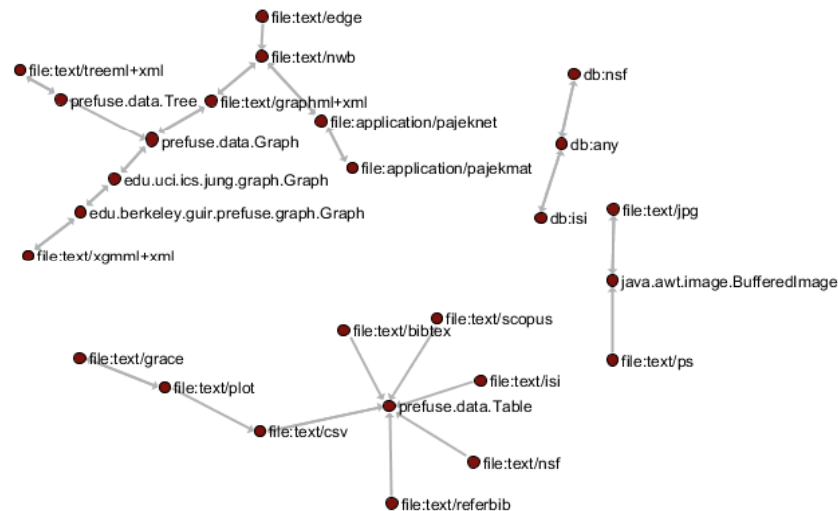
- A standardized, component oriented, computing environment for networked services.
- Successfully used in the industry from high-end servers to embedded mobile devices since 8 years.
- Alliance members include IBM (Eclipse), Sun, Intel, Oracle, Motorola, NEC and many others.
- Widely adopted in open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

- No central data format.
- Sci<sup>2</sup> 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).



- Not all code can be shared freely (yet).
- To make the UCSD Science Map and Cytoscape tool available via the Sci<sup>2</sup> menu, simply add

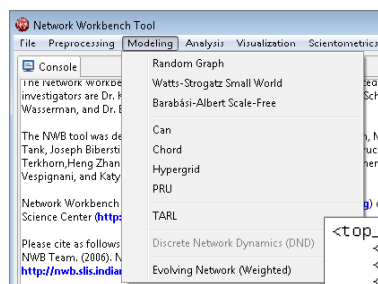
Name	Date modified	Type
edu.iu.scipolicy.visualization.scimap.fields_0.0.1.jar	3/26/2010 6:43 PM	Executable Jar File
edu.iu.scipolicy.visualization.scimap.journals_0.0.1.jar	3/26/2010 6:43 PM	Executable Jar File
edu.iu.scipolicy.visualization.scimap.references_0.0.1.jar	3/26/2010 6:43 PM	Executable Jar File
org.textrend.visualization.cytoscape_0.0.3.jar	3/26/2010 2:46 PM	Executable Jar File

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

- 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.

157

- 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:



```
<top_menu name="Modeling">
<menu pid="edu.iu.nwb.modeling.erdosrandomgraph"/>
<menu pid="edu.iu.nwb.modeling.smallworld"/>
<menu pid="edu.iu.nwb.modeling.barabasiAlbert"/>
<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.Tar1Algorithm"/>
<menu type="break"/>
<menu pid="edu.iu.nwb.modeling.discretenetworkdynamics.DNDAlgorithm"/>
<menu type="break"/>
<menu pid="edu.iu.nwb.modeling.weighted.evolvingnetwork"/>
</top_menu>
```

158

### 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>

159

CIShell/OSGi is at the core of different CIs and a total of 180 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 (Sc<sup>2</sup> 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 or macroscopes will expand.

160



## Overview

- Macroscopes and the Changing Scientific Landscape 15 mins
  - Introduction to network science with sample maps and insights 15 mins
  - Introduction to the Network Workbench Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - Introduction to science studies with sample maps and insights 15 mins
  - Introduction to the Sci<sup>2</sup> Tool 15 mins
  - Demo and hands-on data analysis and visualization by participants 60 mins
  - **Overview of validation approaches for science studies**
  - Plug-and-play tool design using OSGi/CIshell 30 mins
  - **Scholarly Marketplaces** 15 mins
- 240 mins**

161



## Network Workbench Community Wiki

Print | Search:  | Go

# NetworkWorkbench

A Workbench for Network Scientists Main / Home Page

<p><b>Main</b></p> <ul style="list-style-type: none"> <li><a href="#">People</a></li> <li><a href="#">NWB Tool</a></li> <li><a href="#">Update Sites</a></li> <li><b>Tutorials</b></li> <li><a href="#">Algorithms</a></li> <li><a href="#">Datasets</a></li> <li><a href="#">Data Formats</a></li> <li><a href="#">Glossary</a></li> <li><a href="#">FAQ</a></li> </ul> <p>Related Work</p> <p>Site Statistics</p> <p>     </p>	<p><b>About the Network Workbench Community Wiki</b></p> <p>The Network Workbench Community Wiki is the part of <a href="#">Network Workbench (NWB)</a> project. It provides descriptions for algorithms and datasets that have been integrated in the <a href="#">NWB Tool</a>. It is also a place for users of the <a href="#">NWB Tool</a>, the <a href="#">Cyberinfrastructure Shell</a>, or any other CShell based program to get, upload, and request algorithms &amp; datasets to be used in the tool. This site is a sounding board to be used by the community to work together and create a tool which will meet their needs and the needs of the scientific community at large.</p> <p>Check out the lists of available <a href="#">algorithms</a> and <a href="#">datasets</a>. Download the <a href="#">NWB Tool</a> and play with it.</p> <p>You are invited to add or edit your own dataset and algorithm descriptions (sign up <a href="#">here</a>), or post wanted algorithms and datasets.</p> <p>If you are interested in joining the NWB community, please sign up the <a href="#">NWB mailing list</a>, post your question there, or contact Weixia (Bonnie) Huang <a href="mailto:huangb@indiana.edu">huangb@indiana.edu</a> for more information.</p>
--	---

<https://nwb.slis.indiana.edu/community>

162



## Epidemics Marketplace

The screenshot shows the EpiC Marketplace website. At the top, there is a navigation bar with 'Browse', 'Upload', 'Request', and 'About' buttons, and a 'My Account' link. Below the navigation bar is a search bar and a 'Categories' section listing various topics like Demographics, Infectious Diseases, and Social Contagion. The main content area features a large heading 'The EpiC Marketplace' with a sub-heading 'A community to browse, request, and share epidemics data.' Below this are three main actions: 'Browse' (View & Download projects, datasets, and requests, by category and tags), 'Upload' (Upload your datasets or create a project with multiple datasets), and 'Request' (Can't find a dataset you're looking for? Make a Request!). A map titled 'Location of datasets' shows various locations marked with red pins across the globe. On the right side, there is a 'Recent Activity' section with 'Data Requests' and 'Data Uploads'.

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

163



## VIVO National Network of Researchers (see also Eagle-I National Network of Resources)

- Semantic web application + ontology editor developed at Cornell.
- Enables discovery of research and scholarship across multiple schools.
- Facilitates cross-disciplinary collaboration.

<http://vivoweb.org>

The screenshot shows the VIVO National Network of Researchers website. At the top, there is a map of the United States with several university logos and names marked on it, including Washington University in St. Louis, Cornell University, and Indiana University. Below the map is the VIVO logo and the text 'VIVO is a research-focused discovery tool'. The main content area features a search bar, a 'Making Headlines' section with a photo of a man, and a 'Faculty and Staff' section with a 'Manage your page' button. On the right side, there is a 'Upcoming Seminars' section with a list of events.

164



