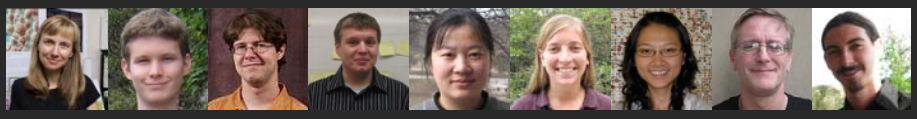


Plug-and-Play Macroscopes Tutorial

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

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 Sci² 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 Fickr 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.

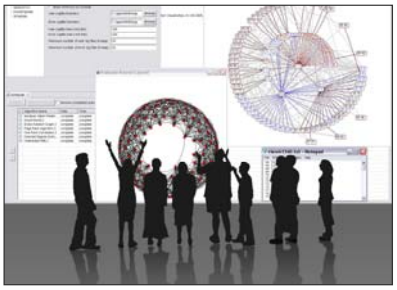
Modularity: The design of software modules with well defined functionality that can be flexibly combined helps reduce costs, makes it possible to have many contribute, and increases flexibility in tool development, augmentation, and customization.

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 Macrocopes





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

Desirable Features of 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.

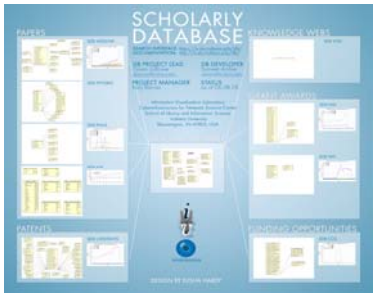
Plugin Content and Interfaces: Should a plugin represent one algorithm or an entire tool? What about data converters needed to make the output of one algorithm compatible with the input of the next? Should those be part of the algorithm plugin or should they be packaged separately?

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.

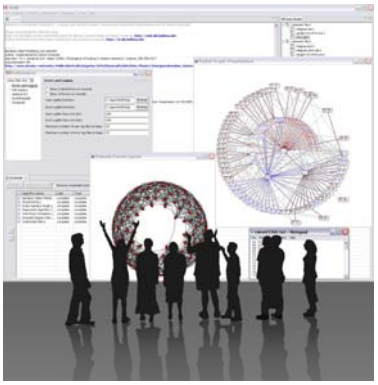
CIs Developed and Served by CNS



Scholarly Database: 23 million scholarly records
http://sdb.slis.indiana.edu

James S. McDonnell Foundation

Information Visualization Cyberinfrastructure
http://iv.slis.indiana.edu



Network Workbench Tool + Community Wiki
http://nwb.slis.indiana.edu

Sci² Tool and Science of Science CI Portal
http://sci.slis.indiana.edu

Epidemics Cyberinfrastructure
http://epic.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 Sci² 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

Computational Proteomics

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

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

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

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

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).

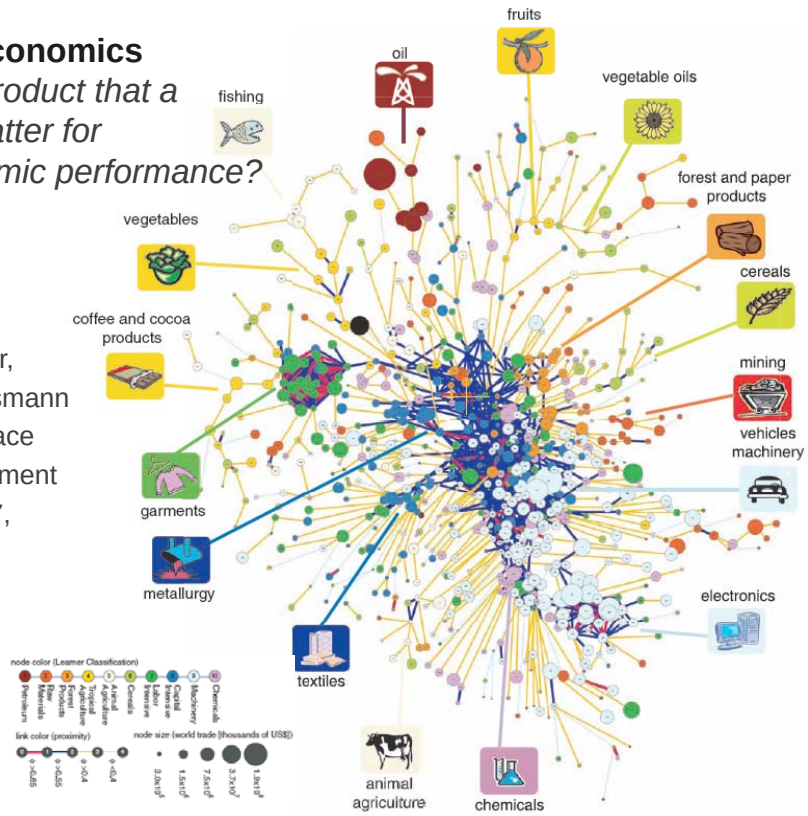


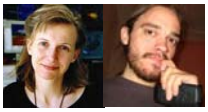
Fig. 1. The product space. (**A**) Hierarchically clustered proximity (φ) matrix representing the 775 SITC-4 product classes exported in the 1998–2000 period. (**B**) Network representation of the product space. Links are color coded with their proximity value. The sizes of the nodes are proportional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

Computational Social Science

Studying large scale social networks such as Wikipedia

Vizzards 2007 Entry

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



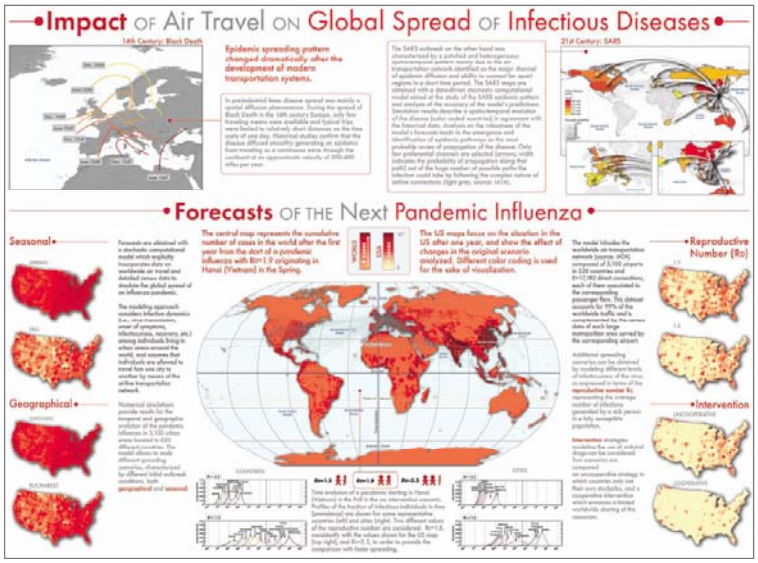
Computational Epidemics

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 Sci² 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

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.

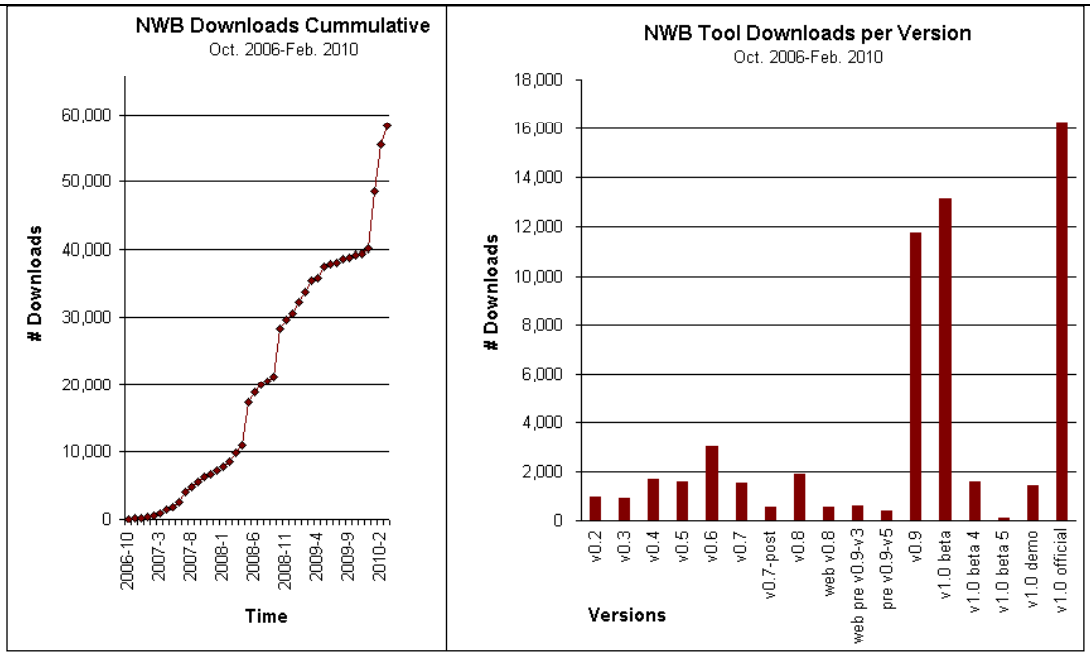
In Aug. 2009, the tool provides more 160 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

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

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.

Network Workbench Tool

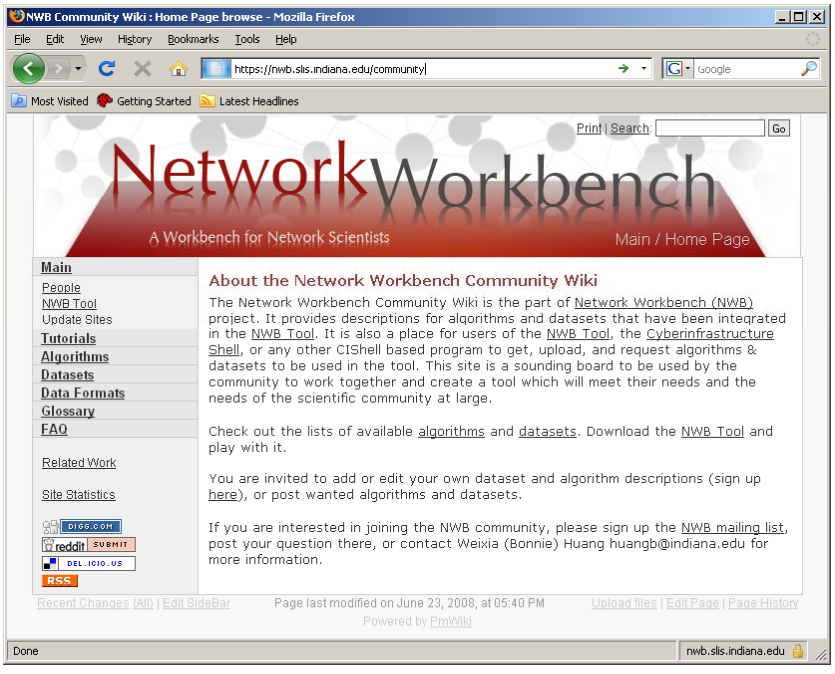
http://nwb.slis.indiana.edu



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.

Network Workbench Community Wiki

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



Project Details

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

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

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

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

Duration: Sept. 2005 - Aug. 2009

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

Project Details (cont.)

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/

NWB Tool: Supported Data Formats

Personal Bibliographies

- Bibtex (.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, Bibtex, 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)

NWB Tool: Algorithms (July 1st, 2008)

See https://nwb.slis.indiana.edu/community and handout for details.

Preprocessing Edit

- **Remove Nodes**
  - Extract Top Nodes
  - Extract Nodes Above or Below Val
  - Delete High Degree Nodes
  - Delete Random Nodes
  - Delete Isolates
- **Remove Edges**
  - Extract Top Edges
  - Extract Edges Above or Below Val
  - Remove Self Loops
  - Trim By Degree[2]
  - Pathfinder Network Scaling
- **Sampling**
  - Snowball Sampling (n nodes)
  - Node Sampling
  - Edge Sampling
- **Transformations**
  - Symmetrize
  - Dichotomize
  - Multipartite Joining

Modeling Edit

- **General**
  - Random Graph
  - Watts-Strogatz Small World
  - Barabási-Albert Scale-Free
- **Structured**
  - CAN
  - Chord
- **Unstructured**
  - Hypergrid
  - PRU
- **Other**
  - TARL
  - Discrete Network Dynamics

Analysis Edit

- **General Purpose**
  - Network Analysis Toolkit[2]
- **Unweighted & Undirected**
  - **Based on degree/**
    - Node Degree
    - Node Distribution
  - **Based on clustering**
    - k-Nearest Neighbor
    - Watts Strogatz Clustering Coefficient
    - Watts Strogatz Clustering Coefficient Over k
  - **Based on path**
    - Diameter
    - Average Shortest Path
    - Shortest Path Distribution
    - Node Betweenness Centrality
  - **Based on components**
    - Connected Components
    - Weak Component Clustering
  - **K-Core**
    - Extract K-Core[2]
    - Annotate K-Coreness[2]
- **Unweighted & Directed**
  - **Based on degree**
    - Node Indegree
    - Node Outdegree
    - Indegree Distribution
    - Outdegree Distribution
  - **Based on local graph structure**
    - k-Nearest Neighbor
    - Single Node In-Out Degree Correlations[2]
  - **Unnamed Category?**
    - Page Rank
  - **Based on local graph structure #2**
    - Dyad Reciprocity[2]
    - Arc Reciprocity[2]

tion Edit

- **Tools**
  - GUESS
  - GnuPlot[2]
- **Predefined Positions Layout**
  - DrL (VxOrd)
  - Pre-defined Positions (prefuse beta)[2]
- **Move**
  - Circular
- **Tree Layouts**
  - Radial Tree (prefuse alpha)
  - Radial Tree with Annotations (prefuse beta)[2]
  - Tree Map
  - Tree View
  - Balloon Graph (prefuse alpha)[2]
- **Network Layouts**
  - Force Directed with Annotation (prefuse beta)
  - Kamada-Kawai (JUNG)
  - Fruchterman-Reingold (JUNG)
  - Fruchterman-Reingold with Annotation (prefuse beta)
  - Spring (JUNG)
  - Small World (prefuse alpha)
- **Other Layouts**
  - Parallel Coordinates (demo)[2]
  - LaNet (k-Core Decomposition)

etrics Edit

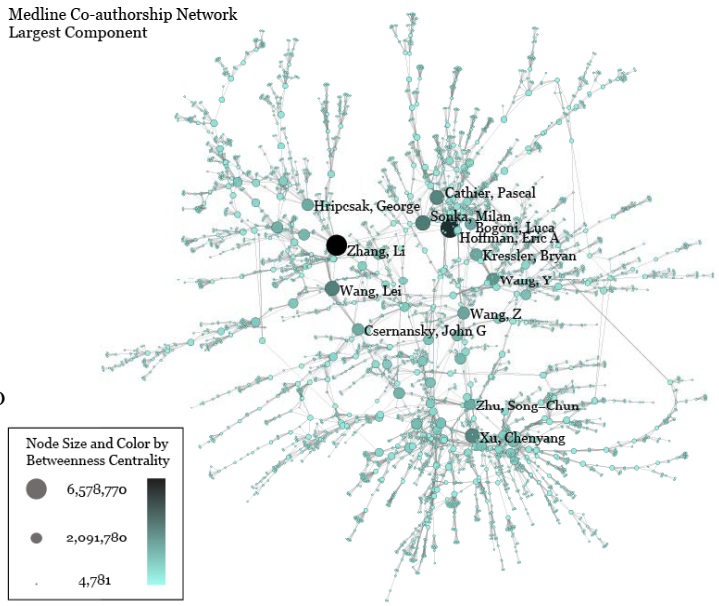
- **Extract Network From Table**
  - Extract Co-Authorship Network
  - Extract Co-Occurrence Network From Table[2]
  - Extract Directed Network From Table[2]
- **Extract Network From Another Network**
  - Extract Bibliographic Coupling Similarity Network
  - Extract Co-Citation Similarity Network[2]
- **Cleaning**
  - Remove ISI Duplicate Records

NWB Tool: Output Formats

- 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.



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^2 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 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.

Software on DVD

Name	Size	Type
I - P (6)		
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
Q - Z (7)		
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

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.

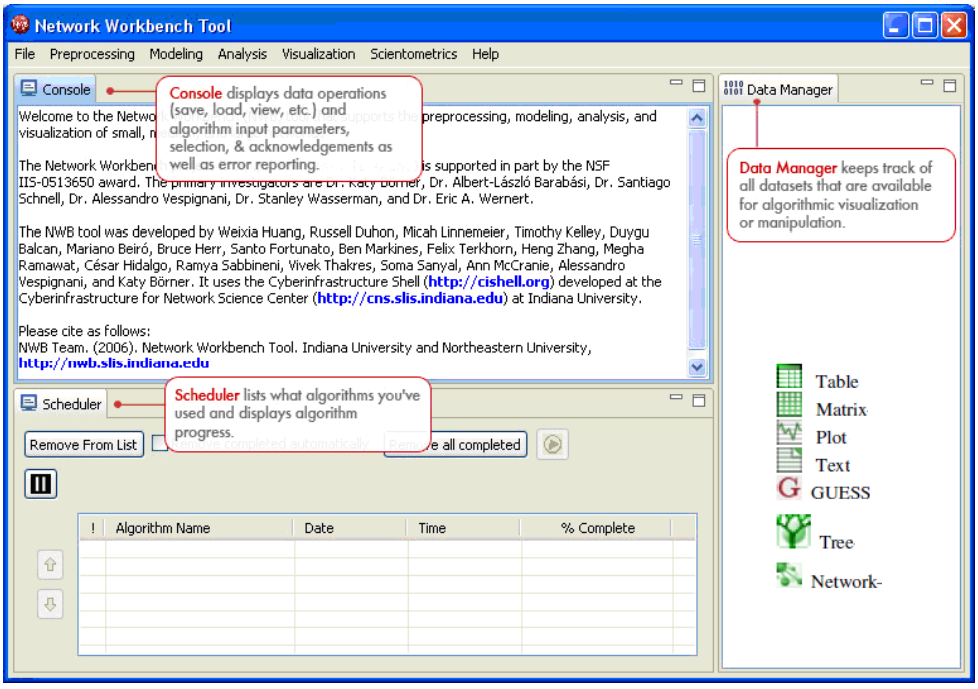
Save as *.jar file.

Install and run.

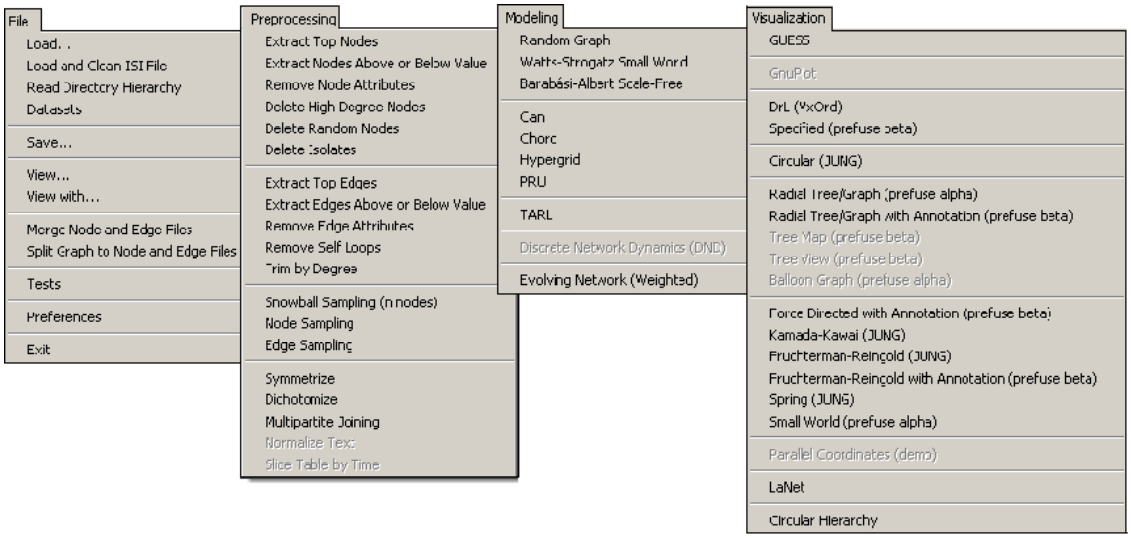
Session log files are stored in '**yournwbdirectory*/logs*' directory.

NWB Demo DVD has all installers.

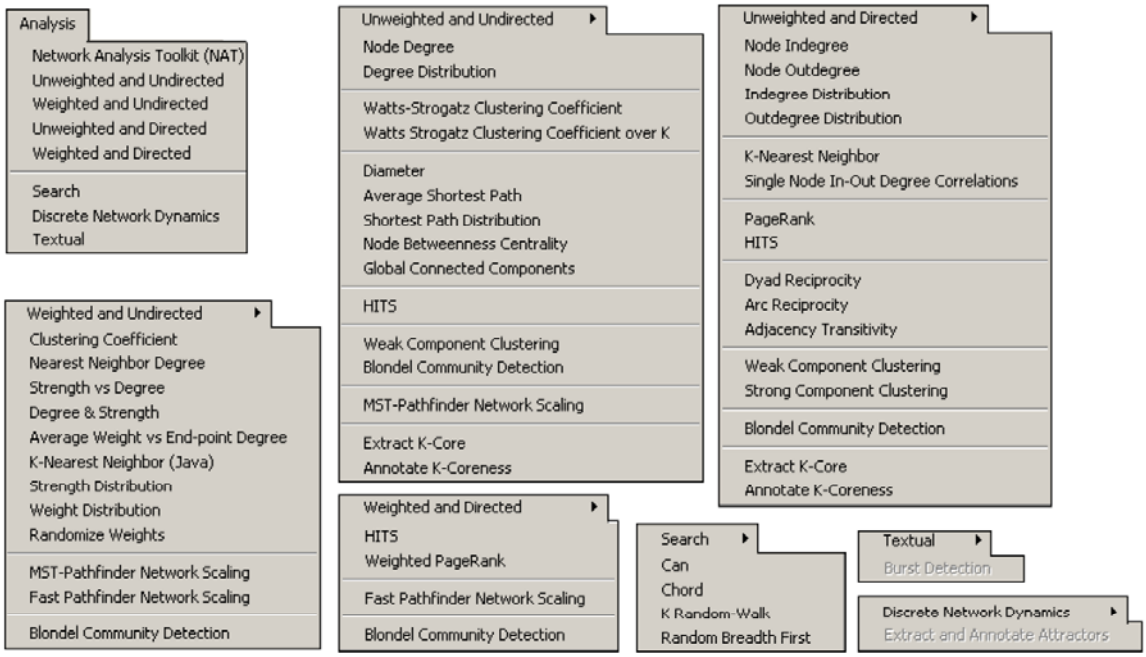
NWB Tool Interface Components



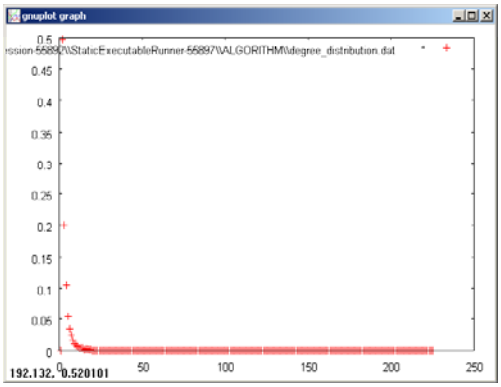
File, Preprocessing, Modeling, and Visualization Menus



Analysis Menu and Submenus

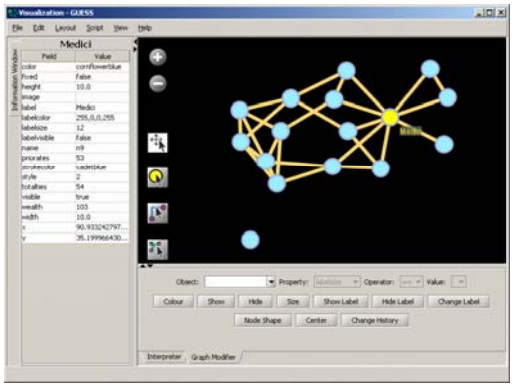


Integrated Tools



Gnuplot

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



GUESS

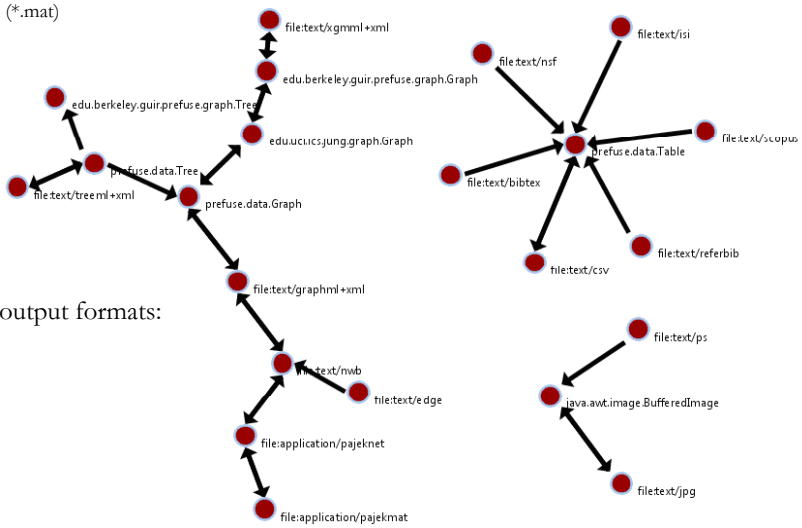
exploratory data analysis and visualization tool for graphs and networks.

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

Supported Data Formats

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)
- Bibtex (*.bib)
- Endnote (*.enw)



and the following network file output formats:

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

These formats are documented at https://nwb.slis.indiana.edu/community/?n=DataFormats.HomePage.

Sample Datasets

The '**yournwbdirectory*/sampledata'* directory provides sample datasets from the biology, network, scientometics, 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

Property Files and Python Scripts

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.

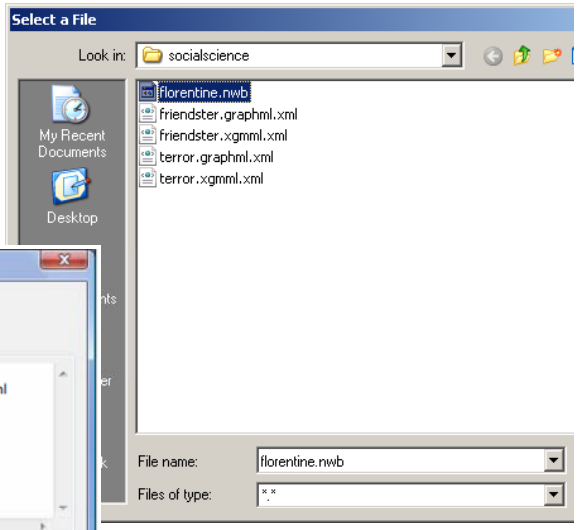
E. Advanced community detection and scalable visualizations.

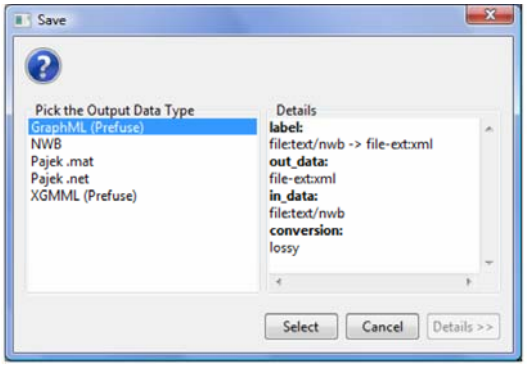
Load, View and Save (Convert) Data

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

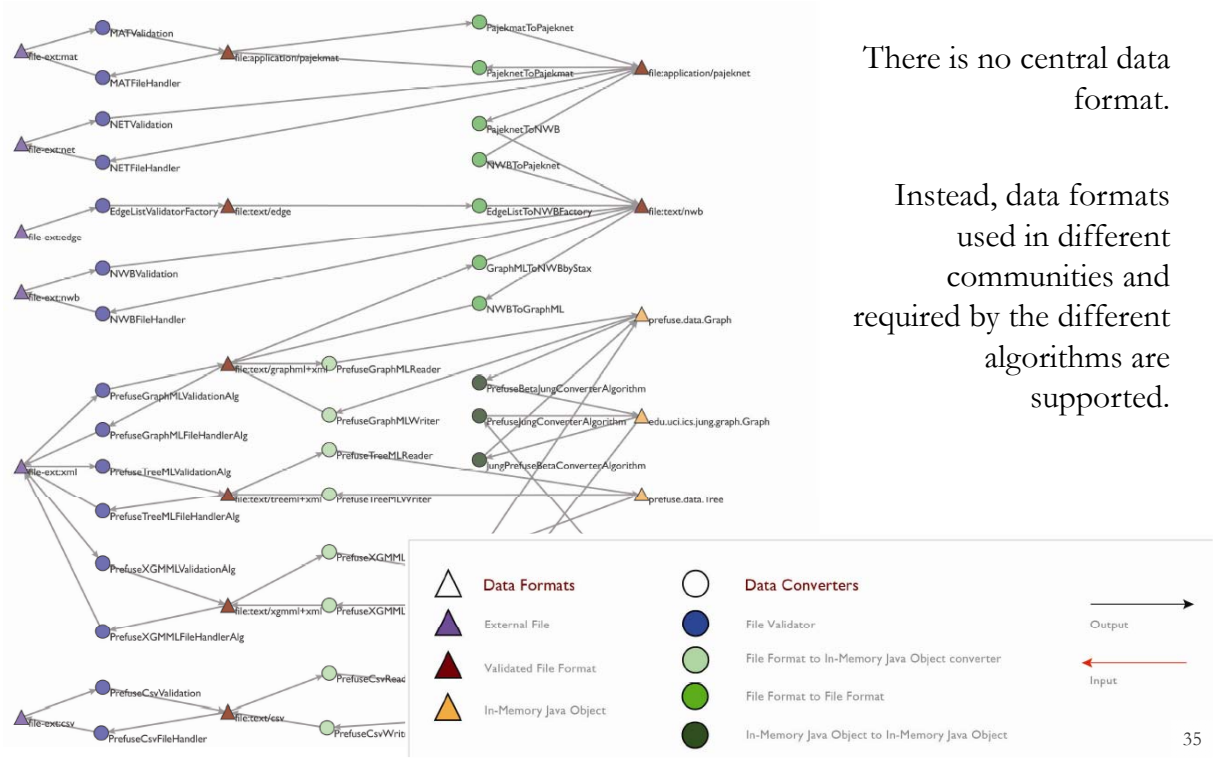
The loaded file will appear in the Data Manager window.

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





Data Converter Graph



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.

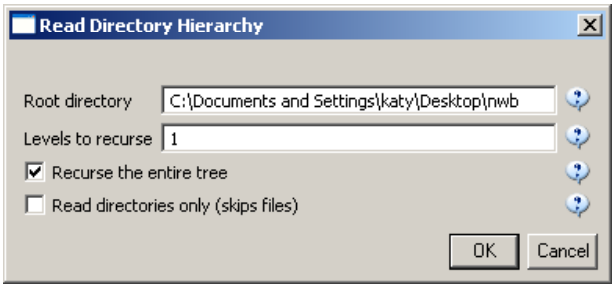
C. **Read and visualize a directory hierarchy.**

D. Load a network, compute its basic properties, and explore it in GUESS.

E. Advanced community detection and scalable visualizations.

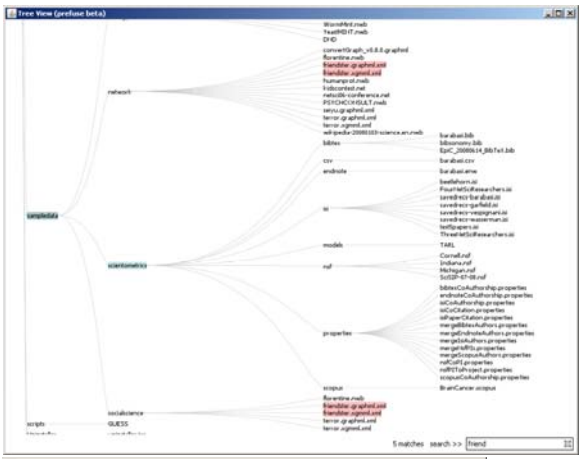
Reading and Visualizing a Directory Hierarchy

Use *'File > Read Directory Hierarchy'* with parameters



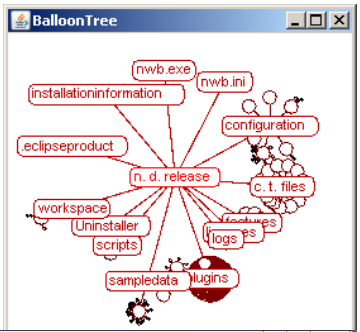
Visualize resulting *'Directory Tree - Prefuse (Beta) Graph'* using

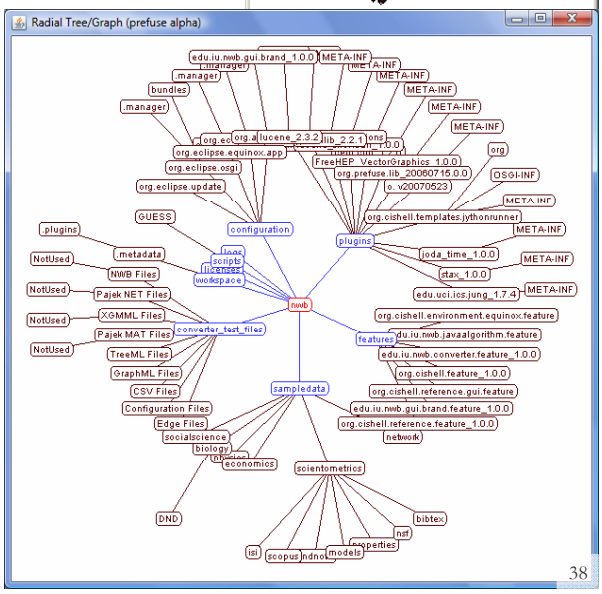
- *'Visualization > Tree View (prefuse beta)'*
- *'Visualization > Tree Map (prefuse beta)'*
- *'Visualization > Balloon Graph (prefuse alpha)'*
- *'Visualization > Radial Tree/Graph (prefuse alpha)'*



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.**

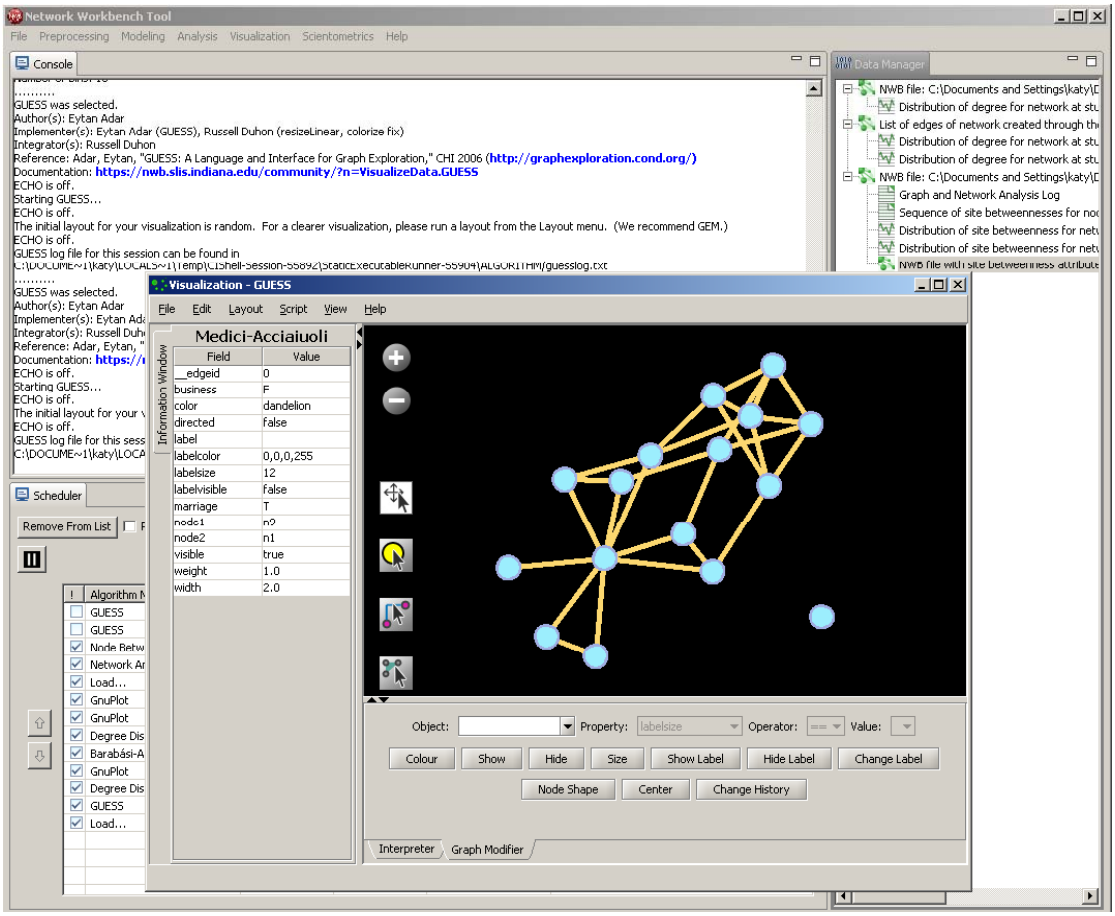
Compute Basic NW Properties & View in GUESS

Select *florentine.nwb* in Data Manager.

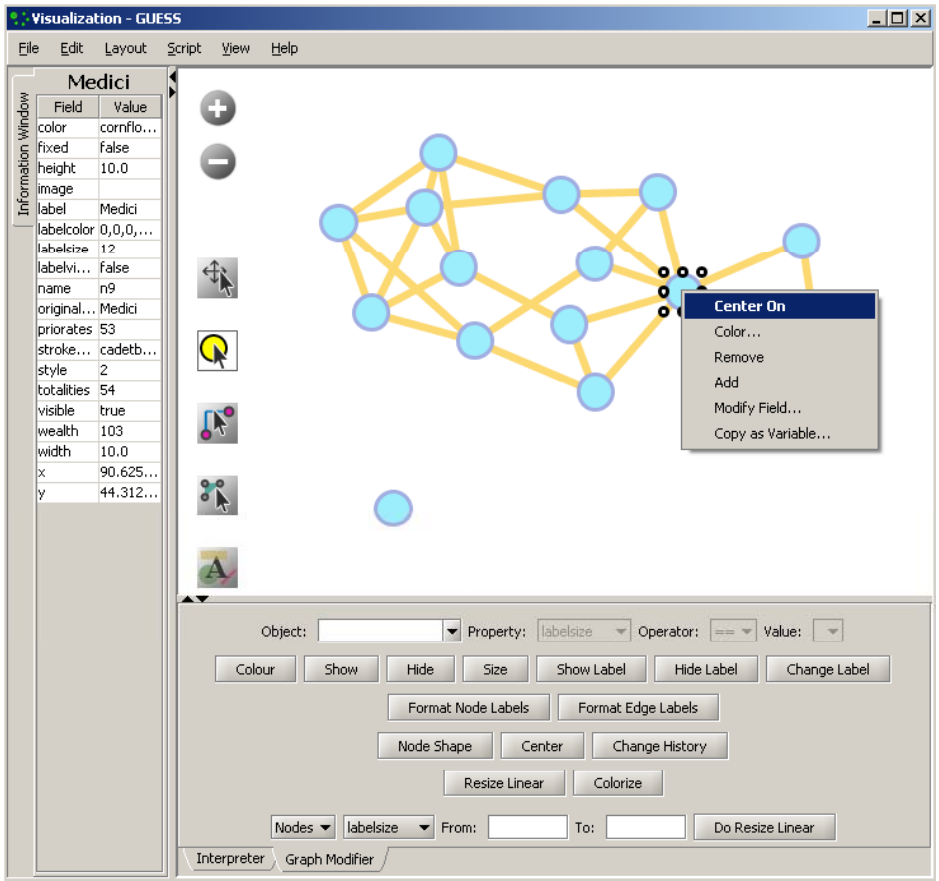
- Run *'Analysis > Network Analysis Toolkit (NAT)'* to get basic properties.

```
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:
                      Example value
        marriag...    T
        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'.



41



Pan:

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

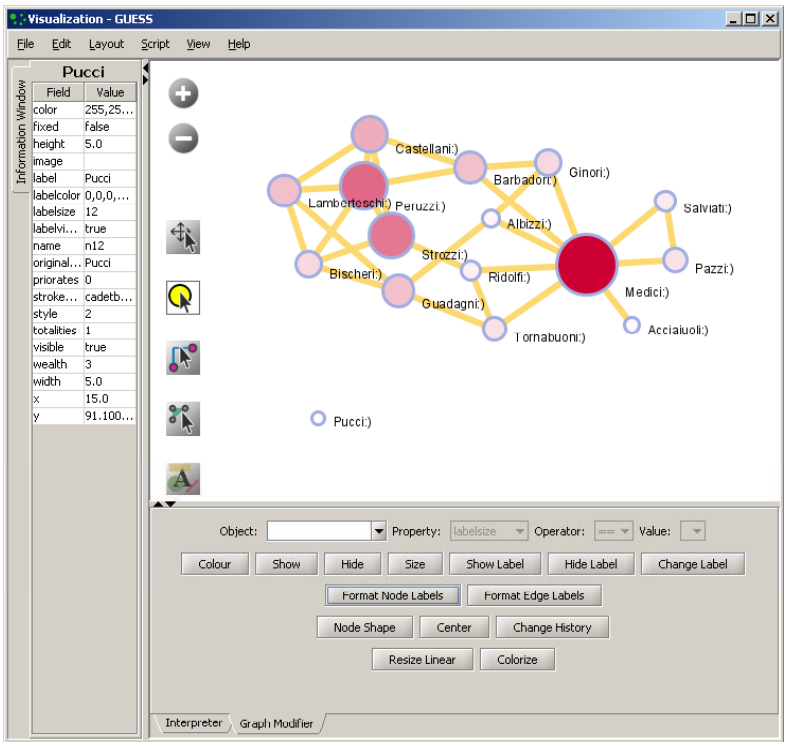
Zoom:

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



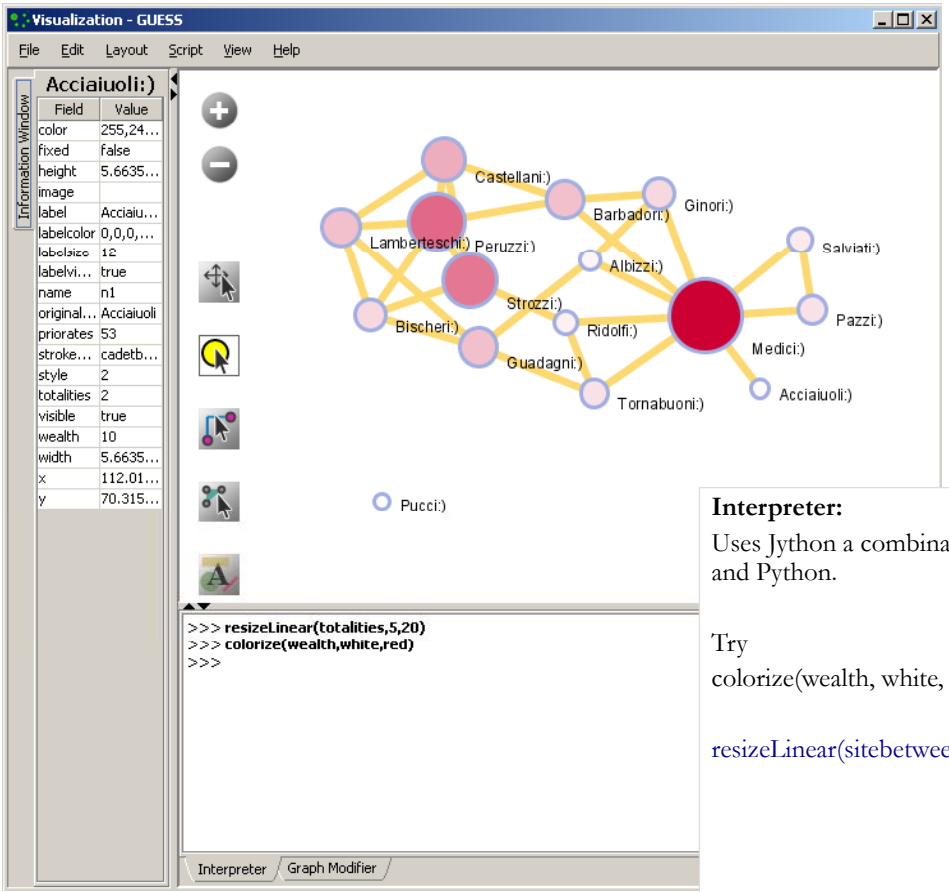
Graph Modifier:

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:

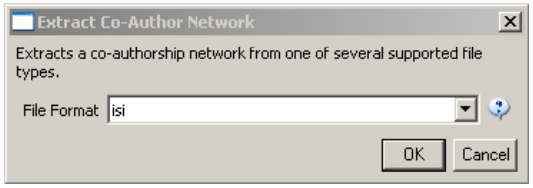
Uses Jython a combination of Java and Python.

Try
colorize(wealth, white, red)

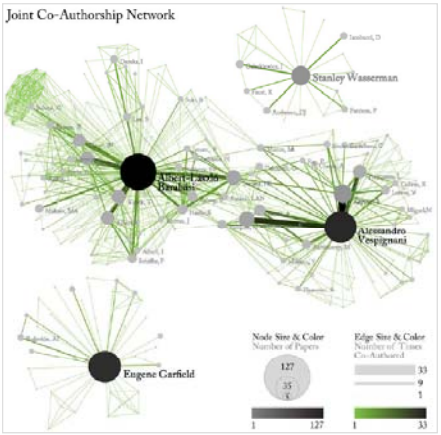
resizeLinear(sitebetweenness, 5, 25)

Extract and Visualize 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*.

NWB Tool Overview

A. Download, install, and run.

B. Load, view, convert, save data.

C. Read and visualize a directory hierarchy.

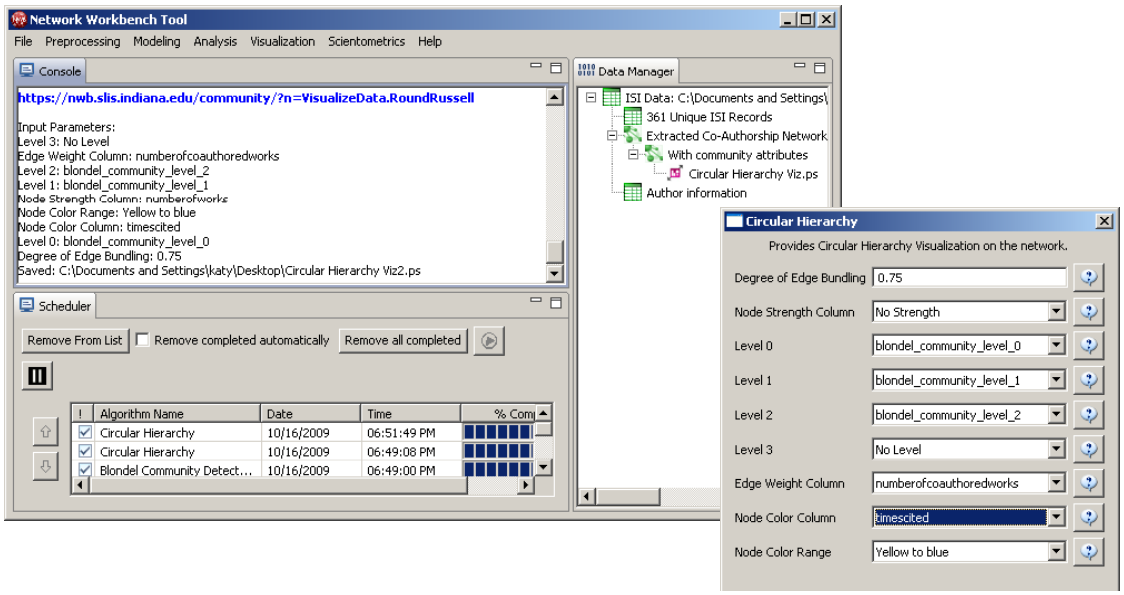
D. Load a network, compute its basic properties, and explore it in GUESS.

E. **Advanced community detection and scalable visualizations.**

Cluster Co-Author Network Hierarchically

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'*.

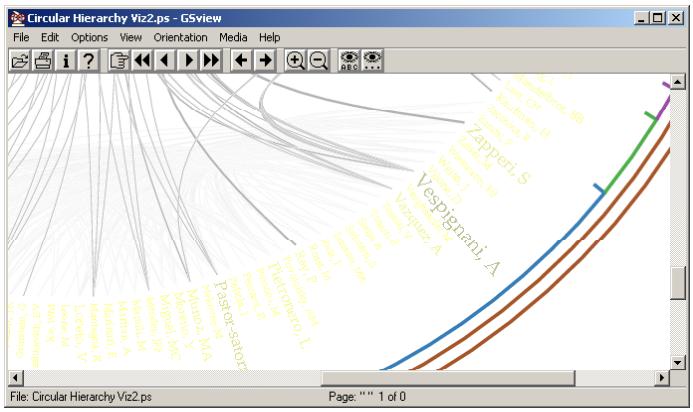


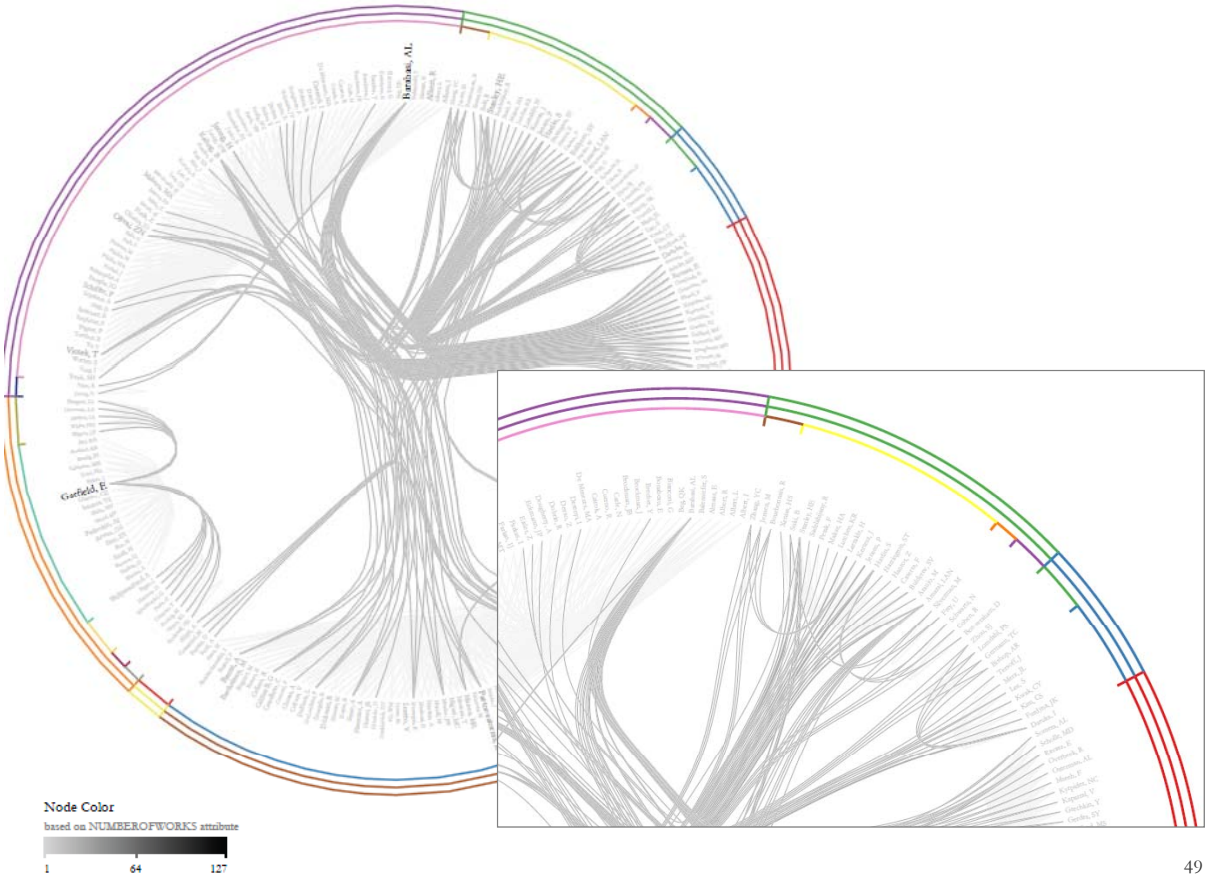
NWB Tool: GSView to View PS Files

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





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^2 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

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

Type of Analysis vs. Scale of Level of Analysis

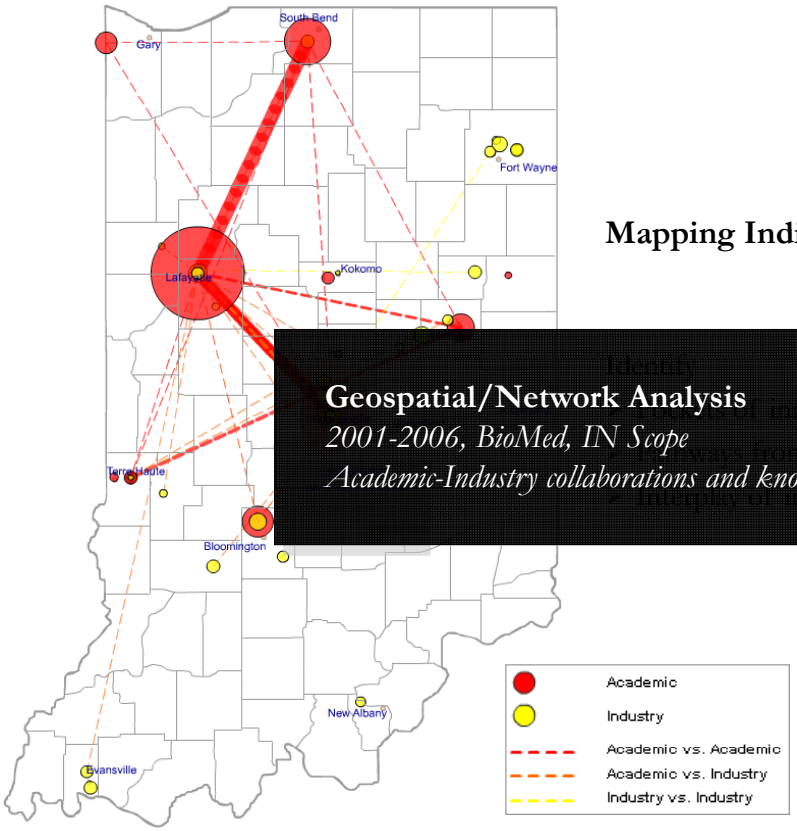
	Micro/Individual (1-100 records)	*Meso/Local (101–10,000 records)*	*Macro/Global (10,000 < records)*
Statistical Analysis/Profiling	Individual person and	Larger labs, centers,	All of NS SA, all of sci
Temporal Analysis (When)		bursts PNAS	113 Years of P Research
Geospatial Analysis (Where)			PNAS
Topical Analysis (What)		arch	VxOrd/Topic m NIH funding
Network Analysis (With Whom?)	NSF work of one		NIH's cy

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.

Mapping Indiana's Intellectual Space

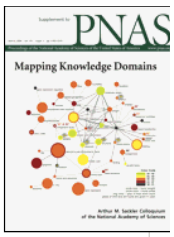


Geospatial/Network Analysis
2001-2006, BioMed, IN Scope
Academic-Industry collaborations and knowledge diffusion

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.



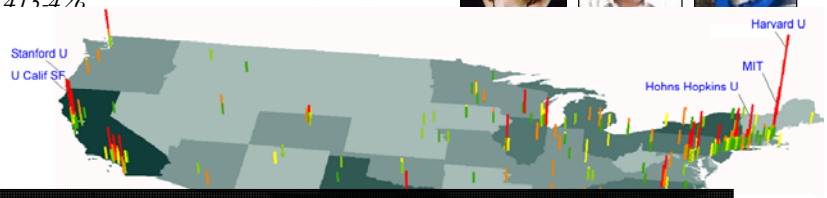
Temporal/Topical Analysis
1982-2001, US, PNAS (BioMed) Scope
Topic evolution and bursts

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

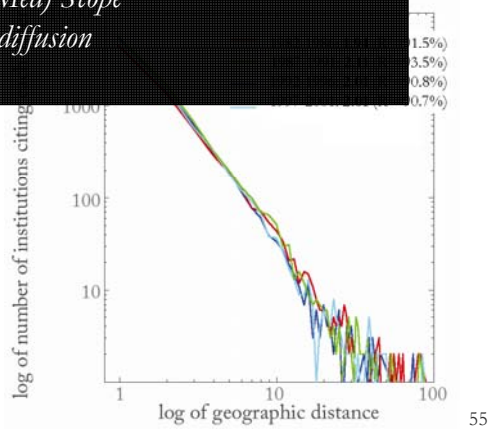
Börner, Katy, Penumarthy, Shashikant, Meiss, Mark and Ke, Weimao. (2006) Mapping the Diffusion of Scholarly Knowledge Among Major U.S. Research Institutions. Scientometrics. 68(3), pp. 415-426.

Research questions:

1. Does space still matter in the Internet age?
2. Does one still have to study and work at major research institutions in order to produce high quality data and high quality research?
3. Does the Internet change these patterns, i.e., more citation links between papers produced at geographically distant research instructions?



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

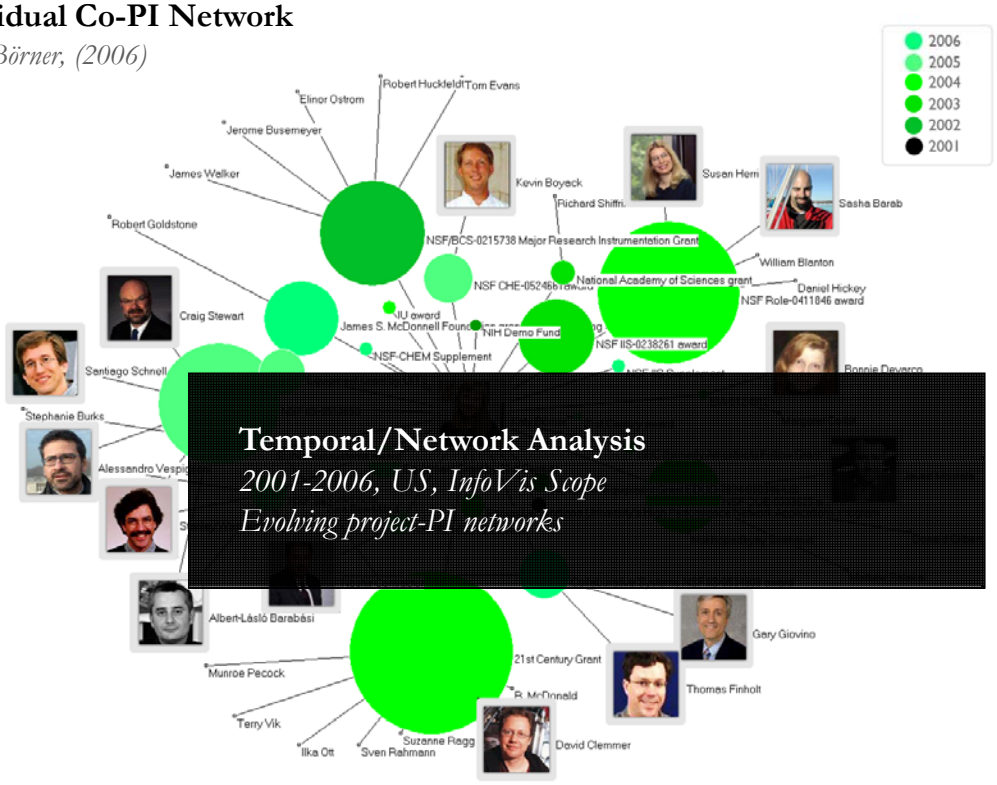
By Weixia (Bonnie) Huang, Russell J. Duhon, Elisha F. Hardy, Katy Börner, Indiana University, USA

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

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

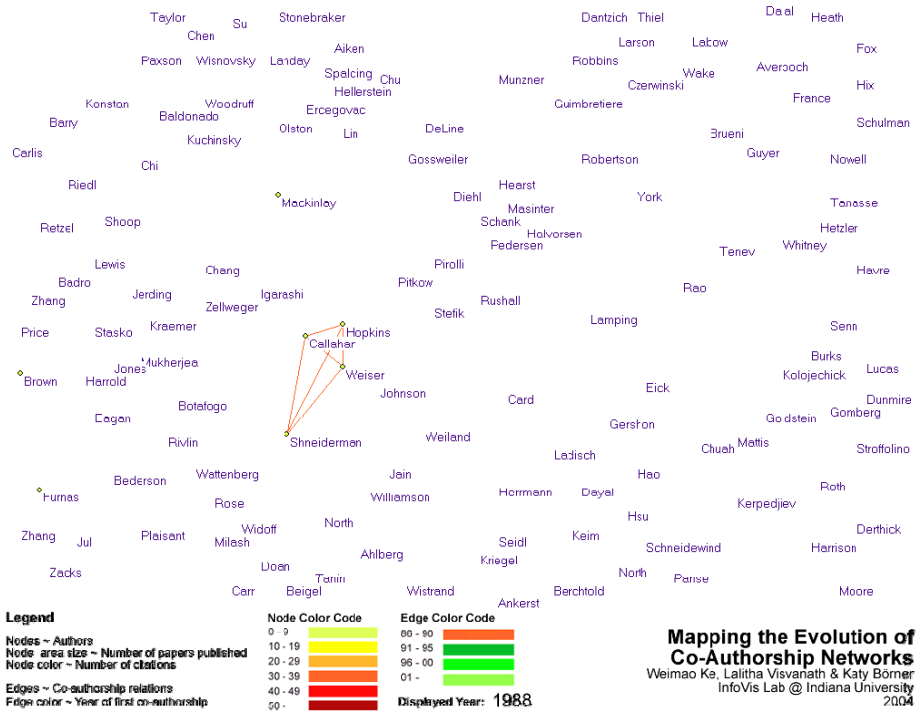
Individual Co-PI Network

Ke & Börner, (2006)



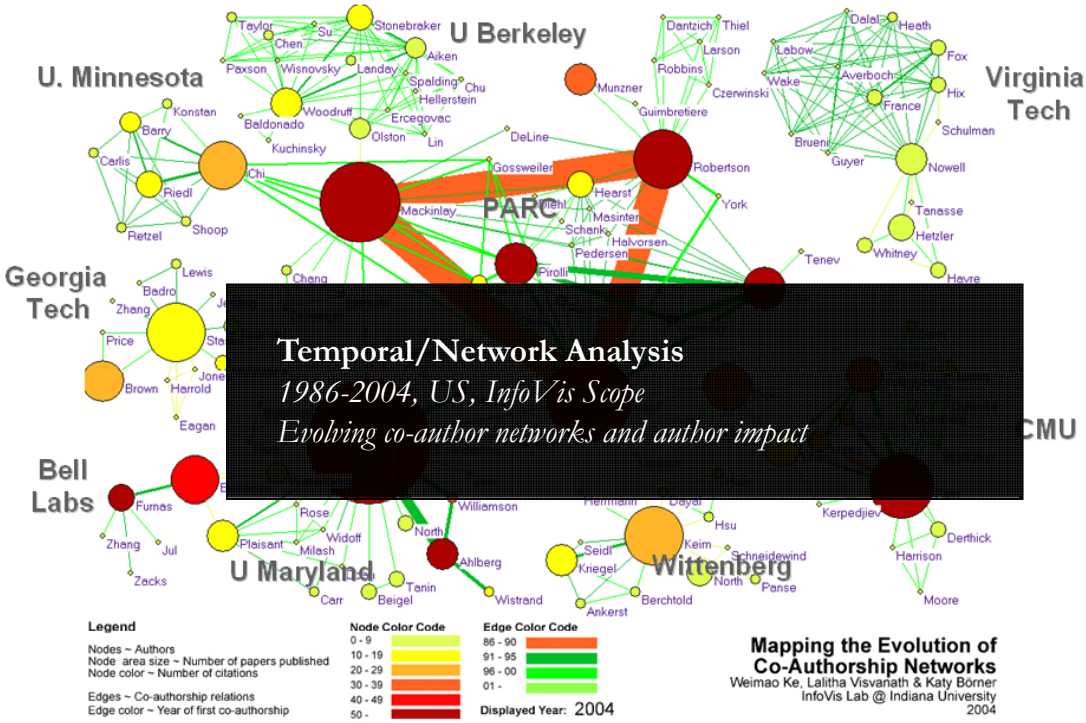
Mapping the Evolution of Co-Authorship Networks

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



Mapping the Evolution of Co-Authorship Networks

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

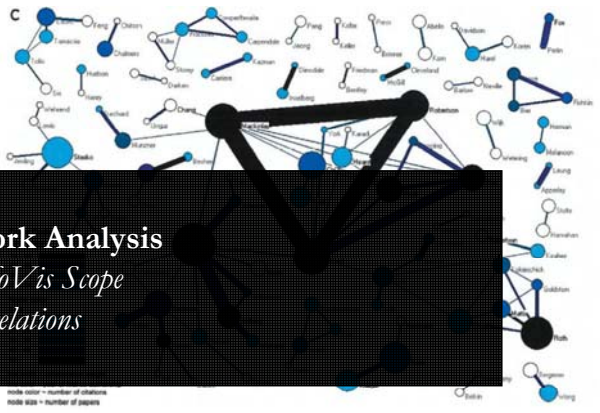


Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams

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

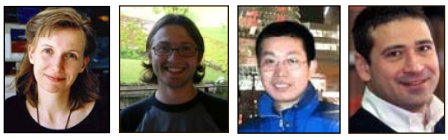
Research question:

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



Contributions:

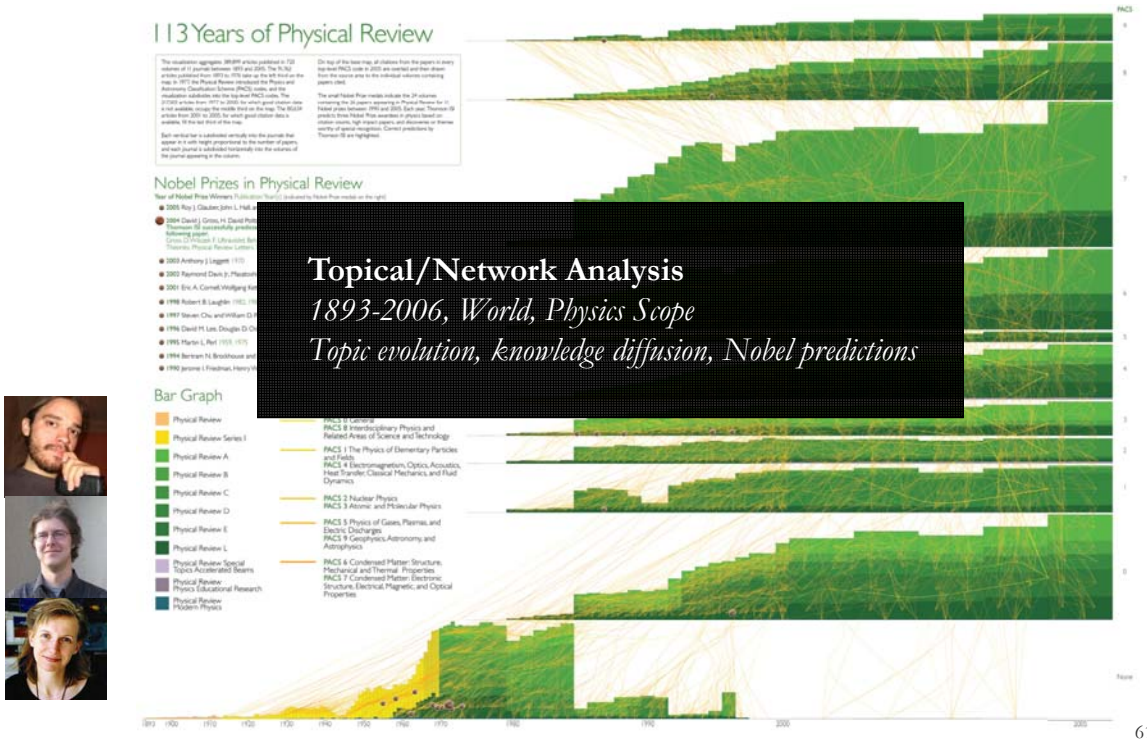
- New approach to allocate citational credit.
- Novel weighted gr
- Visualization of th
co-author network
- Centrality measure
impact.
- Global statistical analysis of paper production and citations in correlation with co-authorship team size over time.
- Local, author-centered entropy measure.



113 Years of Physical Review

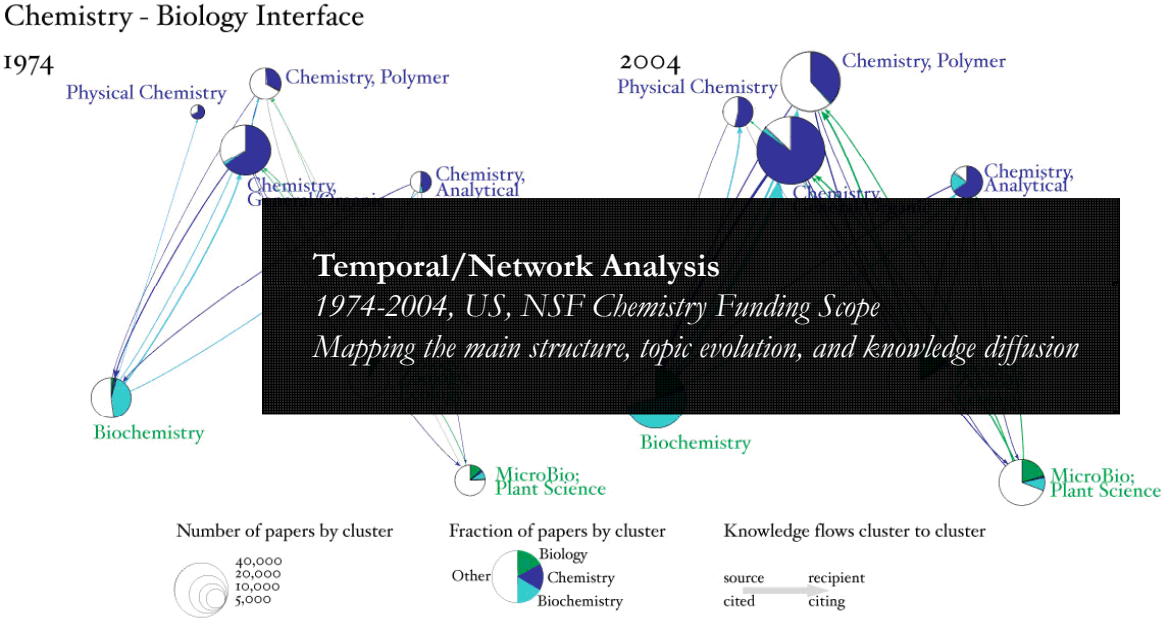
http://scimaps.org/dev/map_detail.php?map_id=171

Bruce W. Herr II and Russell Duhon (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

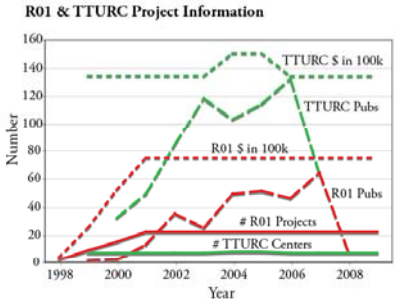
Kevin W. Boyack, Katy Börner, & Richard Klavans (2007)

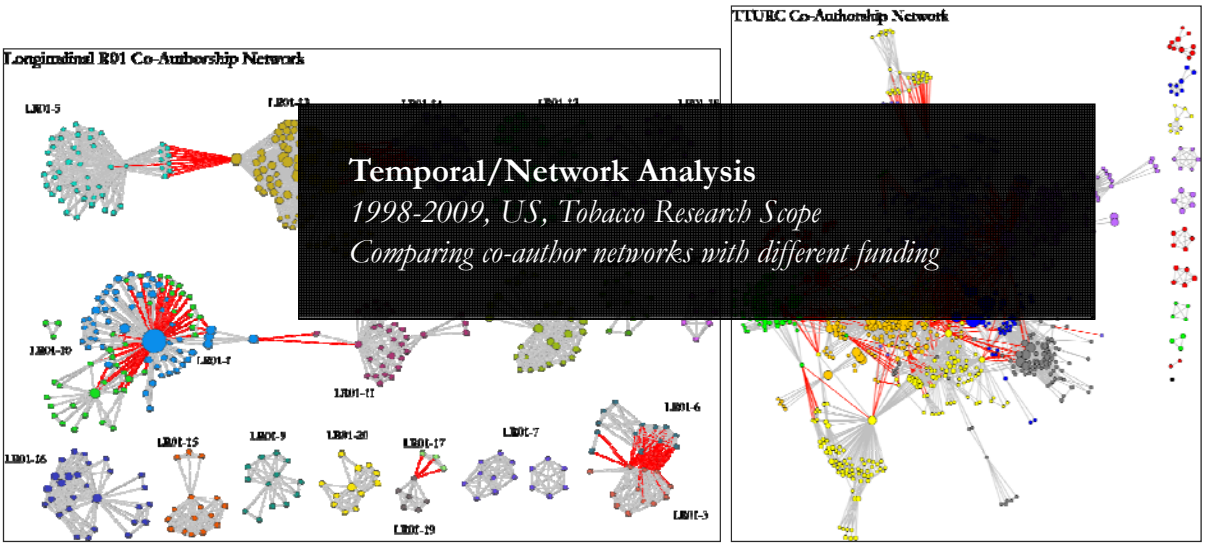


Mapping Transdisciplinary Tobacco Use Research Centers Publications

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

Zoss & Börner, forthcoming.

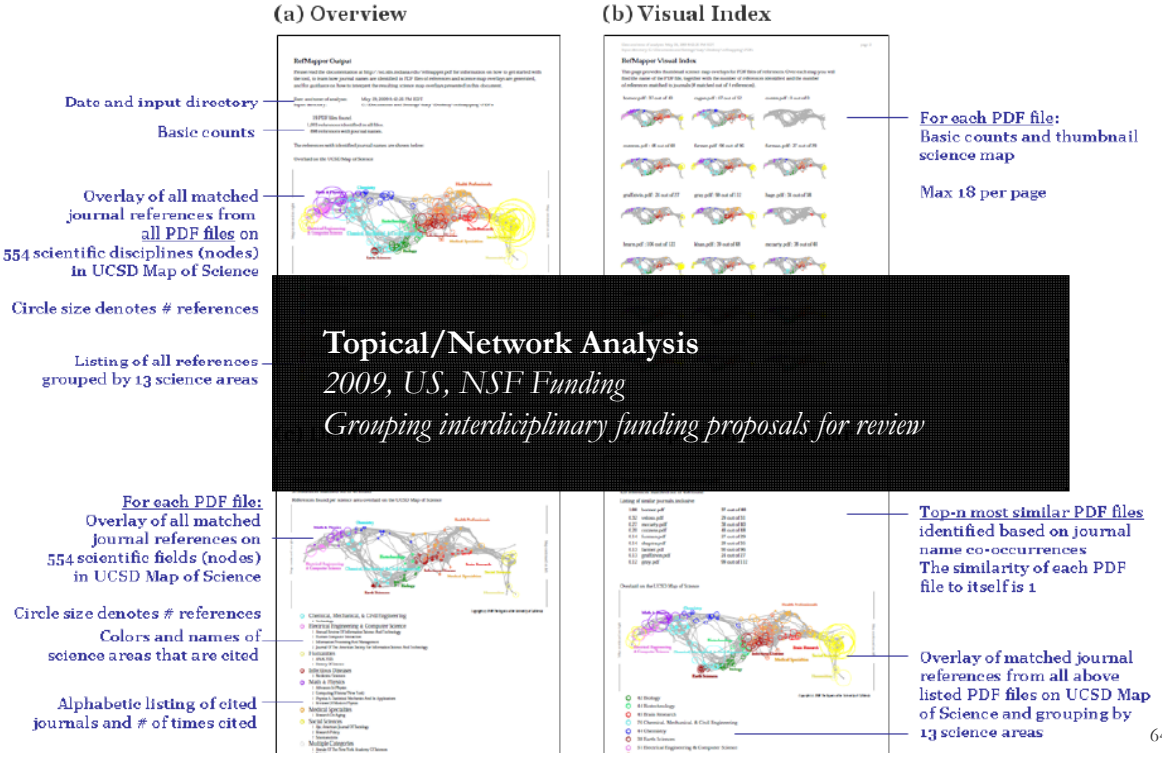




Temporal/Network Analysis
1998-2009, US, Tobacco Research Scope
Comparing co-author networks with different funding

Reference Mapper

Duhon & Börner, forthcoming.

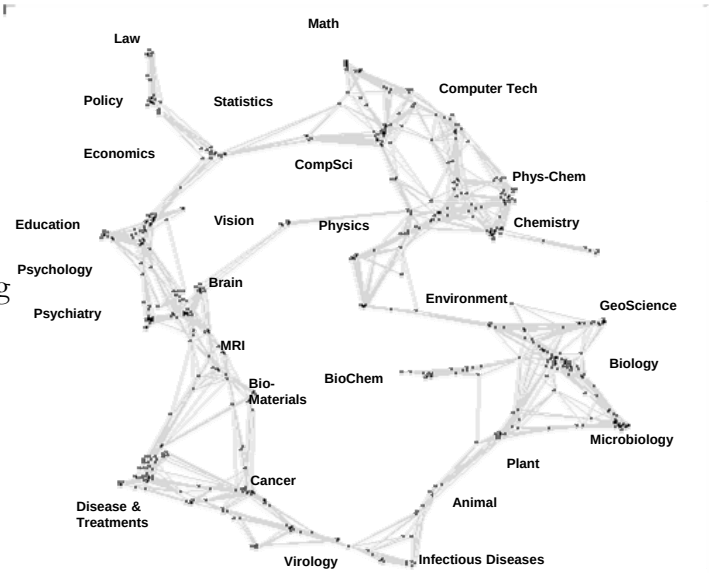


Topical/Network Analysis
2009, US, NSF Funding
Grouping interdiciplinary funding proposals for review

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.

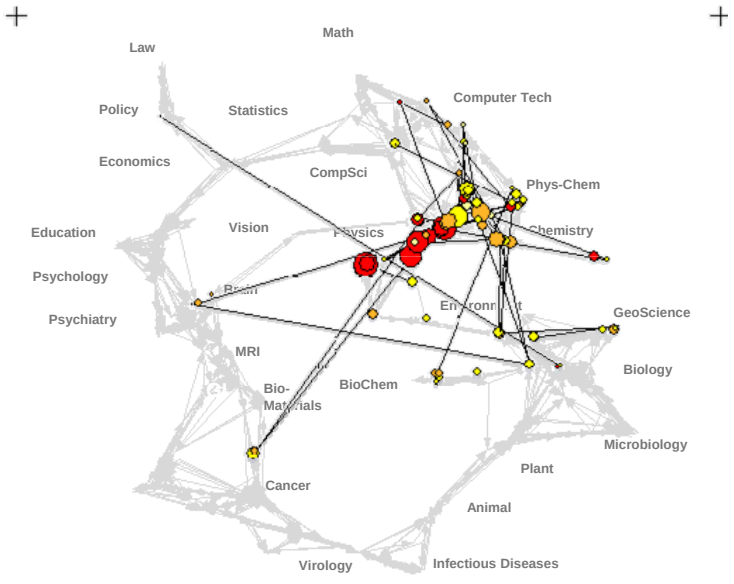
- Uses combined SCI/SSCI from 2002
  - 1.07M papers, 24.5M references, 7,300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts were reaggregated at the journal cluster level to calculate the
  - (x,y)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



Science map applications: Identifying core competency

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

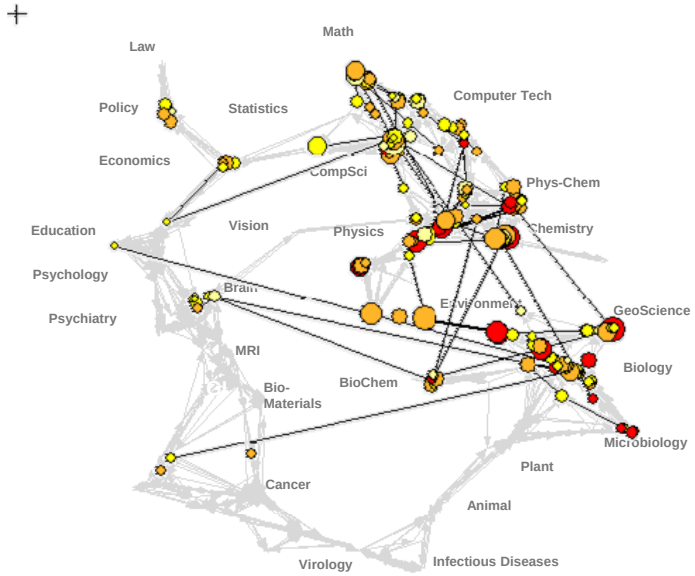
Funding patterns of the US Department of Energy (DOE)



Science map applications: Identifying core competency

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

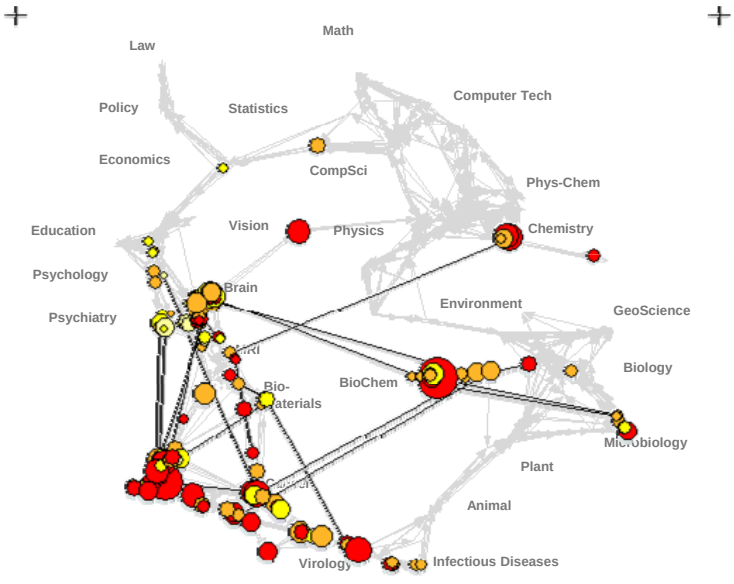
Funding Patterns of the National Science Foundation (NSF)



Science map applications: Identifying core competency

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

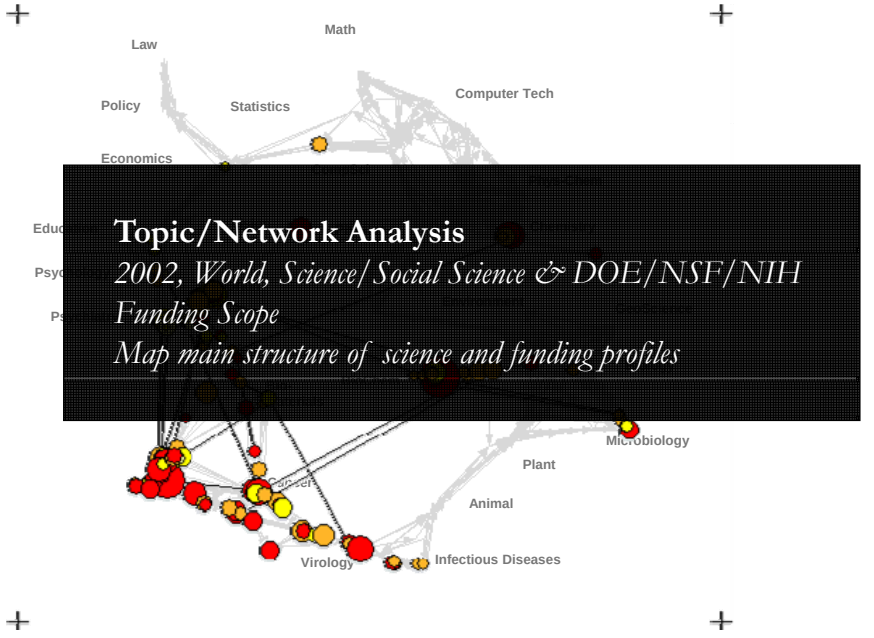
Funding Patterns of the National Institutes of Health (NIH)



Science map applications: Identifying core competency

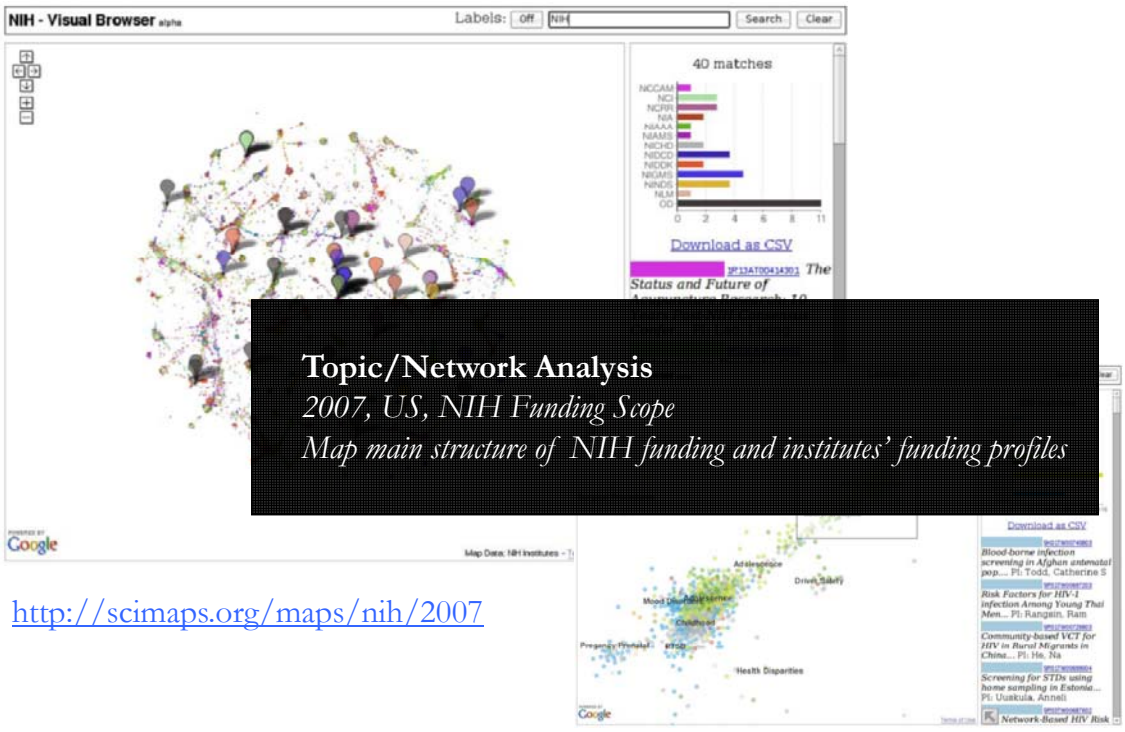
Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

Funding Patterns of the National Institutes of Health (NIH)



Interactive Science Map of NIH Funding

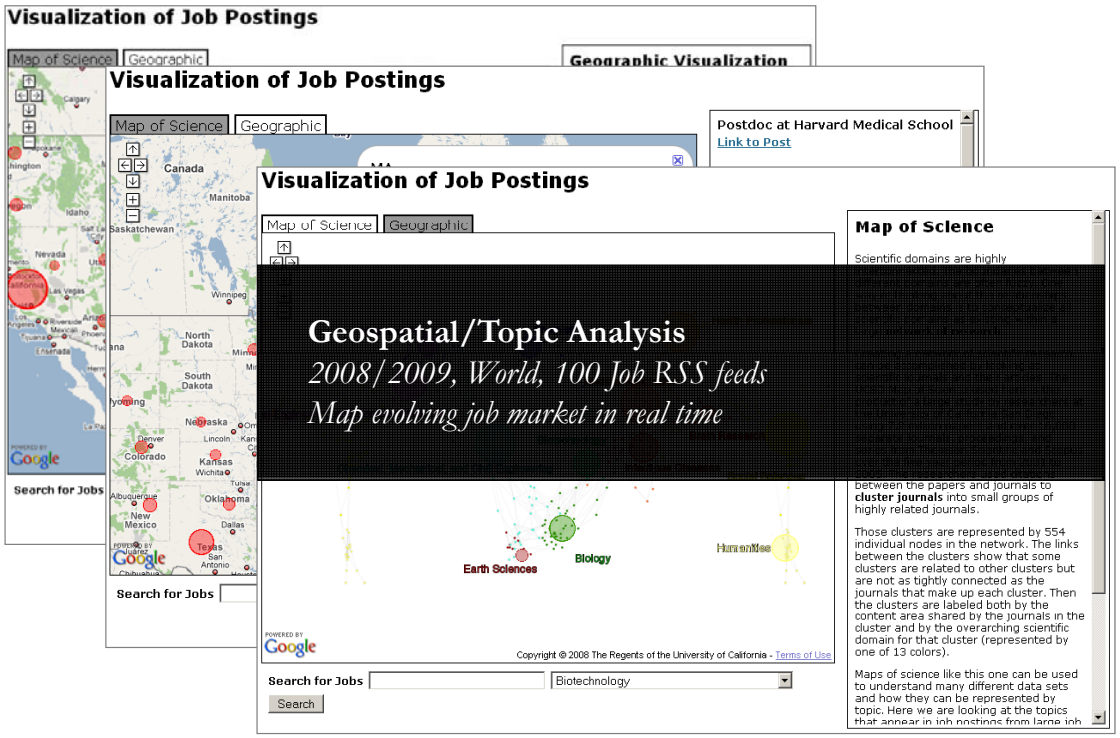
Herr II, Bruce W., Talley, Edmund M, Burns, Gully APC, Newman, David & La Rowe, Gavin. (2009).



http://scimaps.org/maps/nih/2007

Interactive World and Science Map of S&T Jobs

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



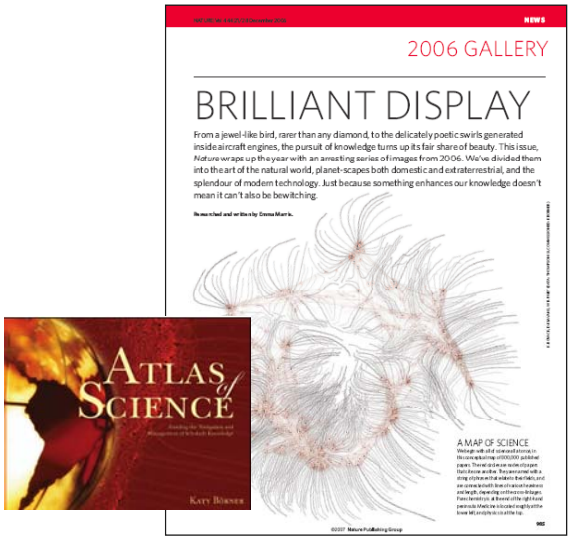
Computational Scientometrics: Studying Science by Scientific Means

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/

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



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.

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

Vol 464|25 March 2010

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

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**.

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^2 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

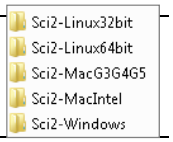
Similarities:

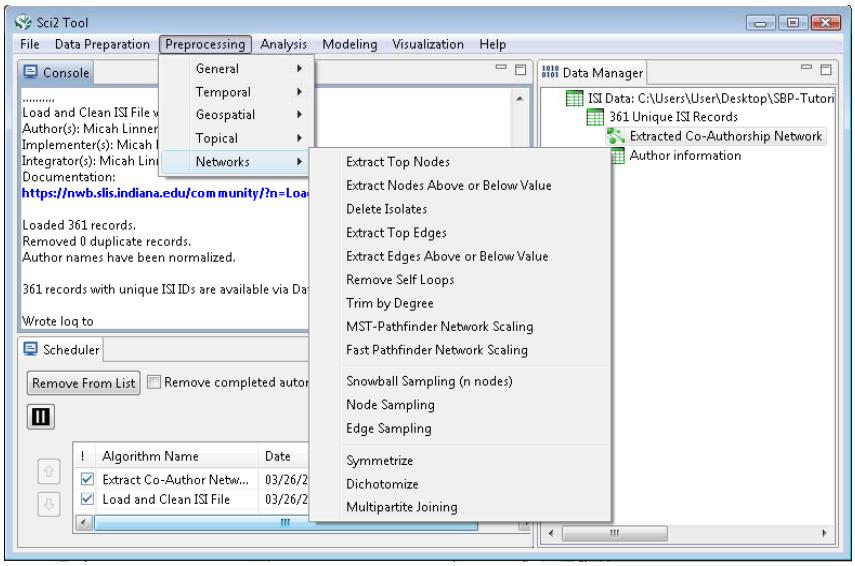
- Both use OSGi/CIShell and are easy to extend/customize.
- Same general interface, look and feel.
- Sci² uses many NWB plugins.

Differences:

- Different target communities, branding, tutorials.
- Both come with different OSGi/CIShell plugin sets, sample datasets, and menu structures.
- Sci² has database support which makes it more scalable.
- Sci² has improved GUESS functionality.
- Sci² has more standard workflows and visualizations.
- Sci² (alpha 0.3) is less mature than NWB (v 1.0).

Sci² Tool for Science of Science Research and Practice (See Demo DVD)





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.

INDIANA UNIVERSITY SCHOOL OF LIBRARY AND INFORMATION SCIENCE

James S. McDonnell Foundation

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.

BROADENING
- By citation
- By terms

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 Decomp (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, **Topics**
- Pathfinder networks (PFNet)
- Self-organizing maps (SOM) includes SOM, ET-maps, etc.

CLUSTER ANALYSIS

SCALAR
- Triangulation
- Force-directed placement (FDP)

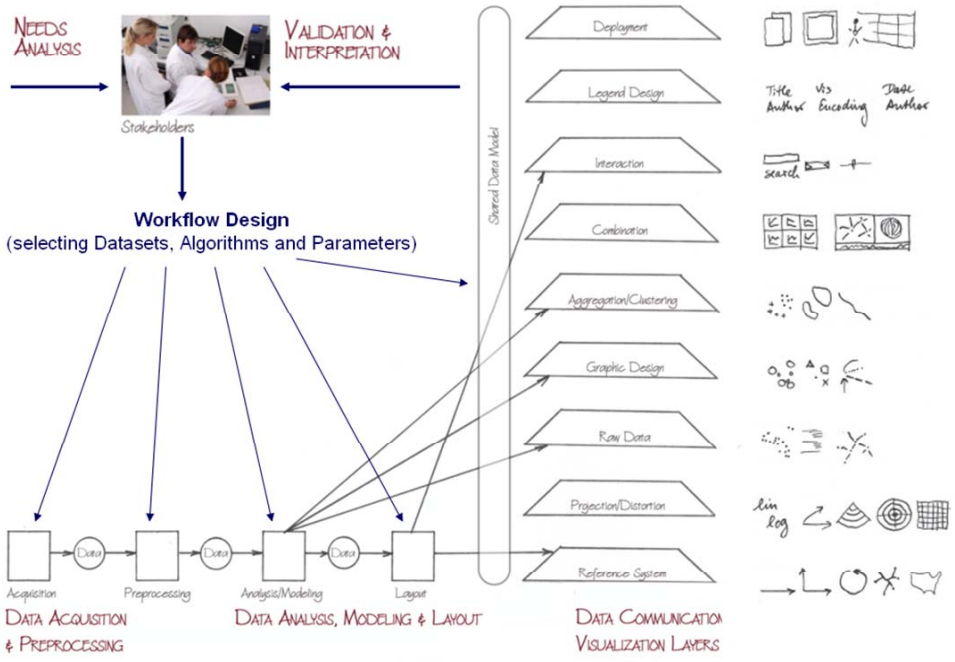
INTERACTION
- Browse
- Pan
- Zoom
- Filter
- Query
- Detail on demand

ANALYSIS

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.

Needs-Driven Workflow Design

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



Börner, Katy (2010) Atlas of Science. MIT Press.

Sci² 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-Coreness
- ---
- HITS

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
- ---
- HITS

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
- ---

Sci² Tool: Algorithms cont.

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

- ---
- Extract K-Core
- Annotate K-Coreness
- ---
- HITS
- PageRank

Weighted & Directed
- HITS
- 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

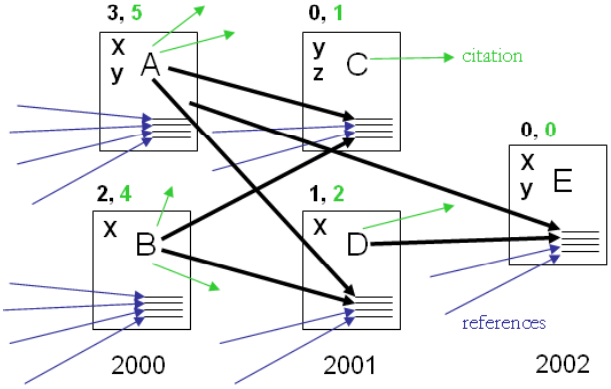
Extract Co-Occurrence Network
Extract Word Co-Occurrence Network
Extract Co-Author Network
Extract Reference Co-Occurrence (Bibliographic Coupling) Network
---
Extract Document Co-Citation Network

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

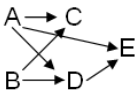
General Network extraction

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.



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

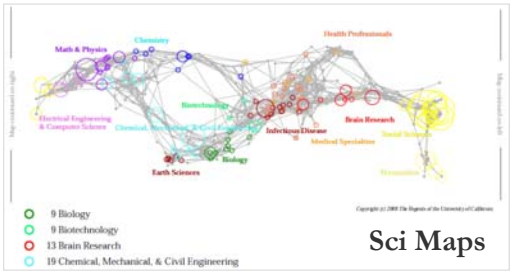
Document Co-Citation (DCA) Network
A and B are co-cited by C and D
A and D are co-cited by E

Reference Co-Occurrence (Bibliographic Coupling) Network
C and D are bibliographically coupled as they both cite/reference A and B.

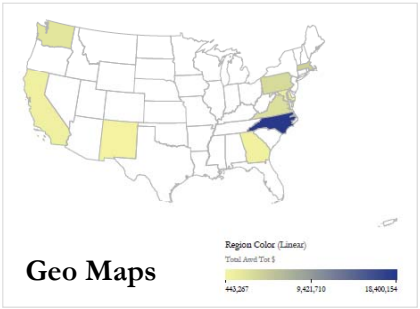
Local citation counts (within this dataset) are given in **black** and global citation counts (ISI times cited) are given in green above each paper.

Sci² Tool

Plugins that render into Postscript files:



Sci Maps



Geo Maps

Horizontal Time Graphs

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

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^2 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

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

Sci² 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)

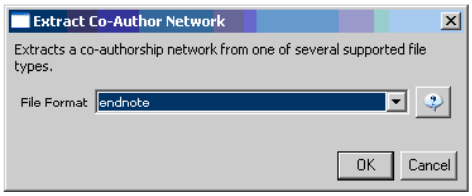
- Geo USPTO (SDB Data) *(section 5.3.1)*

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

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

Many researchers/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:



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:

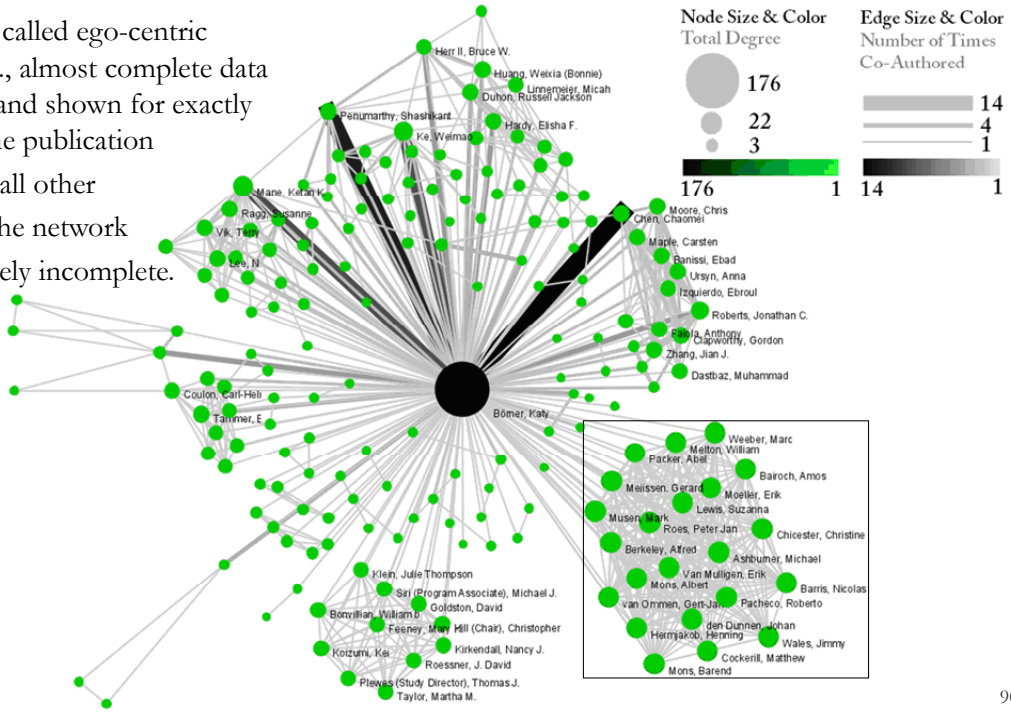
1. Resize Linear > Nodes > totaldegree > From: 5 To: 30 > Do Resize Linear (Note: total degree is the number of papers)
2. Resize Linear > Edges > weight From: 1 To: 10 > Do Resize Linear (Note: weight is the number of co-authored papers)
3. Colorize > Nodes > totaldegree From : To: > Do Colorize
4. 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
```


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

Co-Author Network

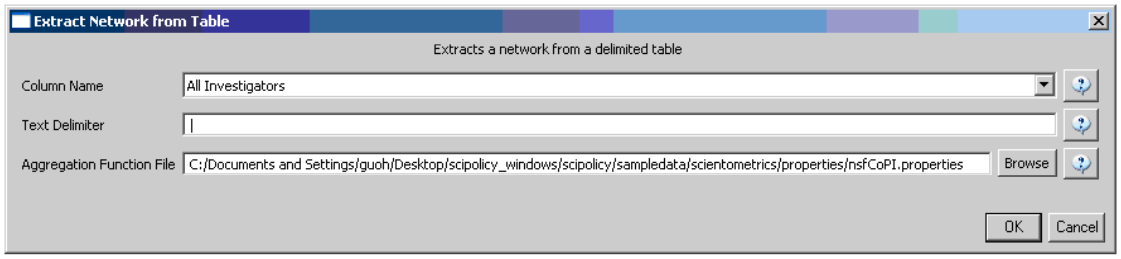
This is a so called ego-centric network, i.e., almost complete data is available and shown for exactly one ego. The publication records for all other authors in the network are most likely incomplete.



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:



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

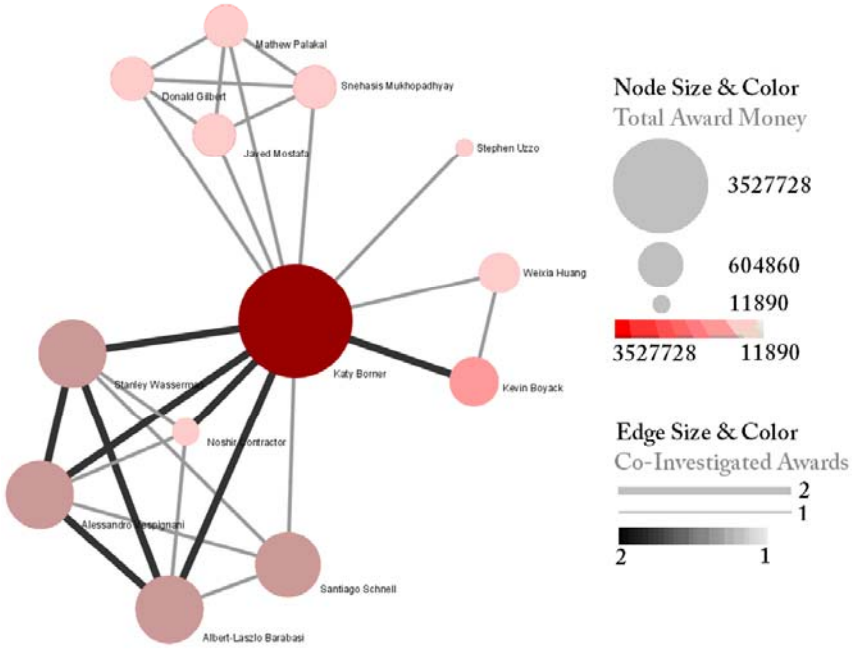
Load the default Co-PI visualization theme via '*File > Run Script …*' and load *'*yoursci2directory*/scripts/GUESS/co-PI-nw.py*'. Alternatively, use the "Graph Modifier" to customize the visualization. The resulting network in Figure 5.2 was modified using the following workflow:

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
```

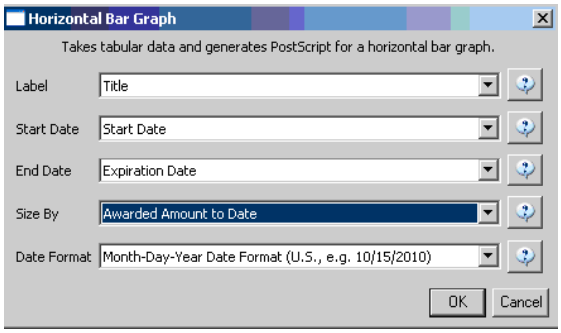
Co-PI Network

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.



Award Durations and Totals

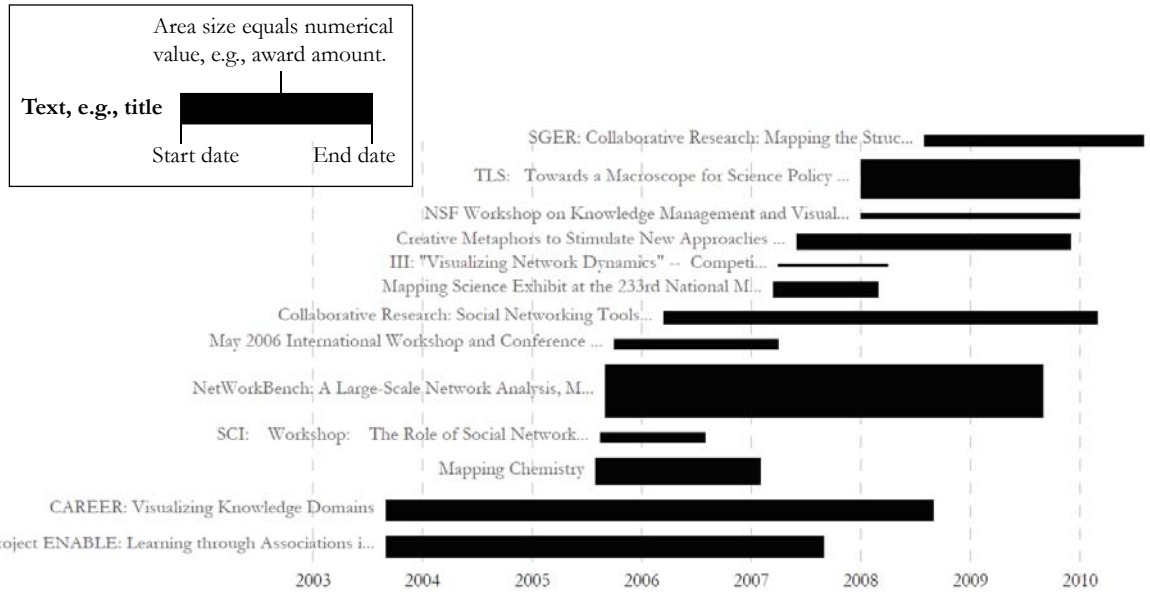
For a summary of the grants themselves, with a visual representation of their award amount, select the NSF csv file in the Data Manager and run '*Visualization* > *Temporal* > *Horizontal Bar Graph*', entering the following parameters:



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

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

Award Durations and Totals



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

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

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

To find all publications by an author, search for the last name and the first initial followed by an asterisk in the author field.

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

Comparison of Counts

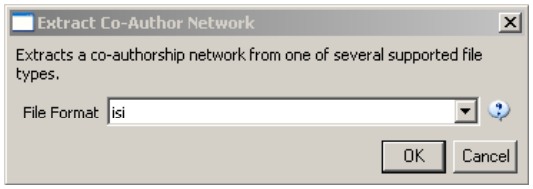
No books and other non-WoS publications are covered.

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

Extract Co-Author Network

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

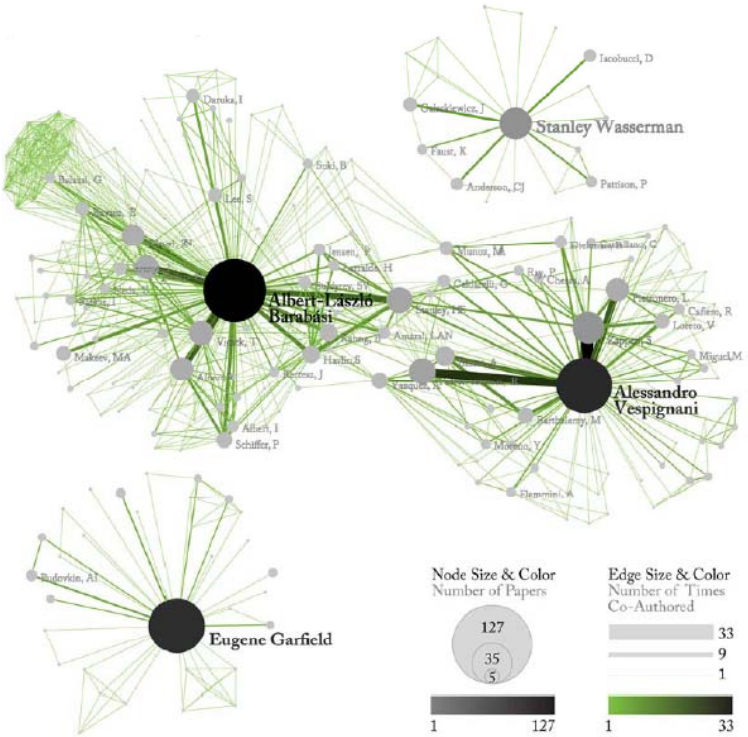
To extract the co-author network, select the *'361 Unique ISI Records'* table and run *'Scientometrics > Extract Co-Author Network'* using isi file format:



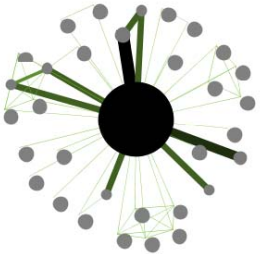
The result is an undirected network of co-authors in the Data Manager. It has 247 nodes and 891 edges.

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*.

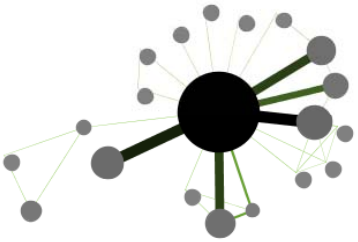
Joint Co-Author Network of all Four NetsSci Researchers



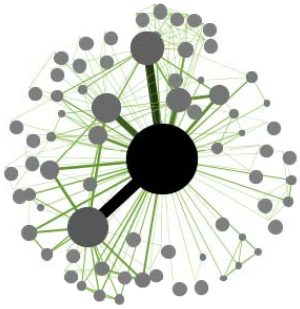
Individual Co-Author Networks (Read/map 4 files separately)



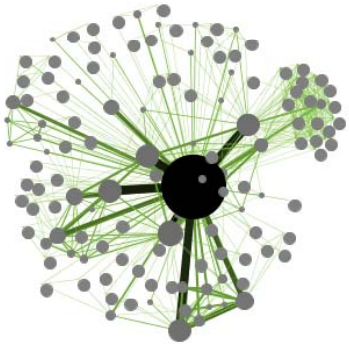
Eugene Garfield



Stanley Wasserman



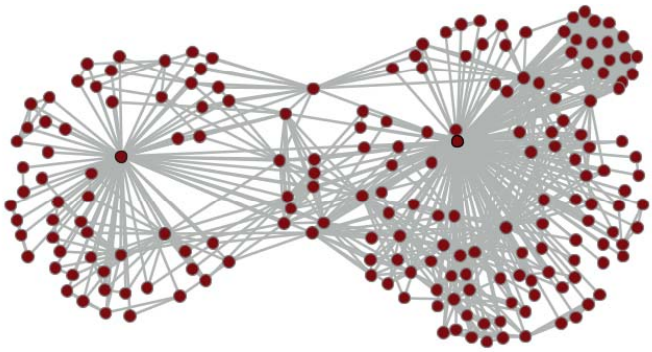
Alessandro Vespignani



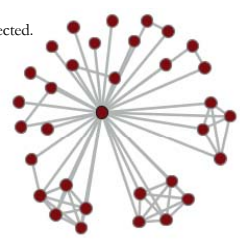
Albert-László Barabási

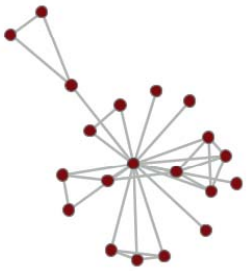
Network Visualization: Node Layout

Network Science researchers



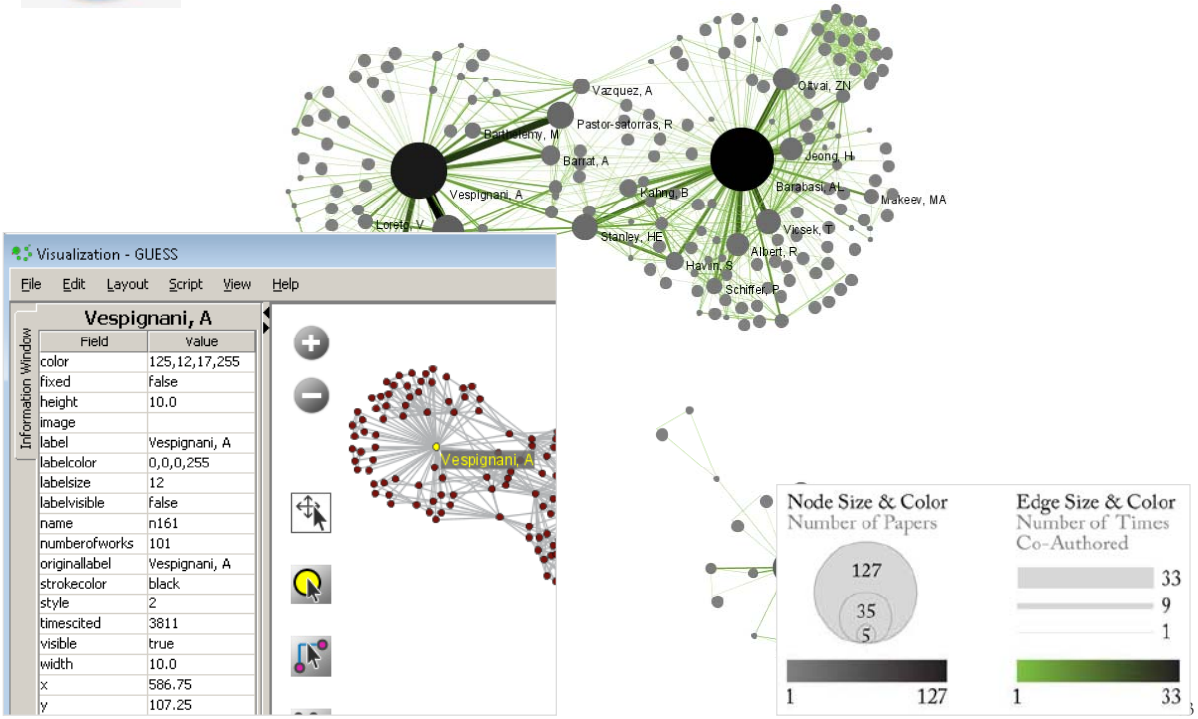
Load and Clean ISI File was selected.
Loaded 361 records.
Removed 0 duplicate records.
Author names have been normalized.
361 records with unique ISI IDs are available via Data Manager.
.....
Extract Co-Author Network was selected.
Input Parameters:
File Format: isi
.....
Network Analysis Toolkit (NAT)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





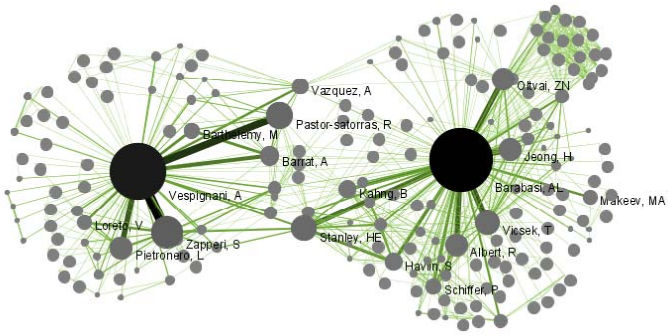
Network Visualization:
Color/Size Coding by Data Attribute Values

Network Science researchers



Network Visualization:
Giant Component

Network Science researchers

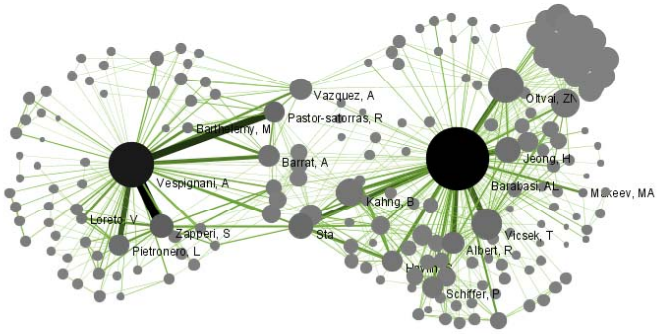


.....
Weak Component Clustering was selected.
Implementer(s): Russell Duhon
Integrator(s): Russell Duhon

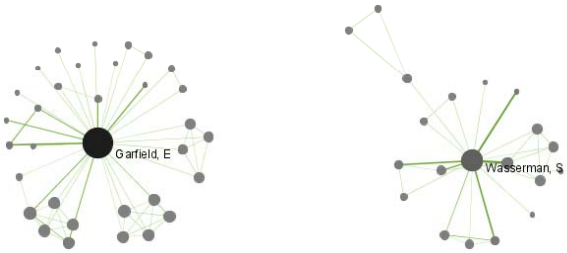
Input Parameters:
Number of top clusters: 10
3 clusters found, generating graphs for the top 3 clusters.
.....

Network Visualization:
Color/Size Coding by Degree

Network Science researchers

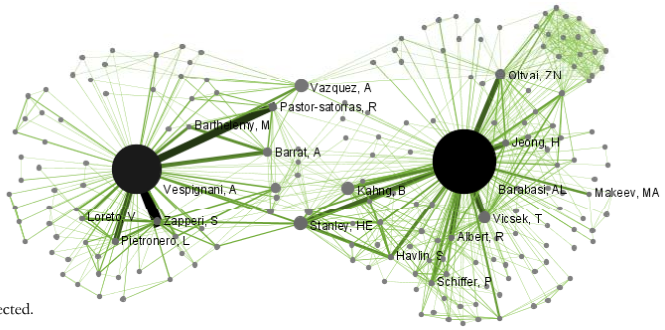


.....
Node Degree was selected.
Documentation:
https://nwb.slis.indiana.edu/community/?n=AnalyzeData.NodeDegree
.....



Network Visualization:
Color/Size Coding by Betweeness Centrality

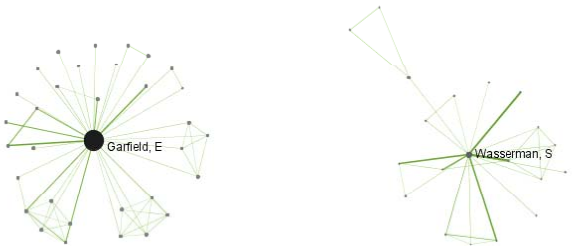
Network Science researchers



.....
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.

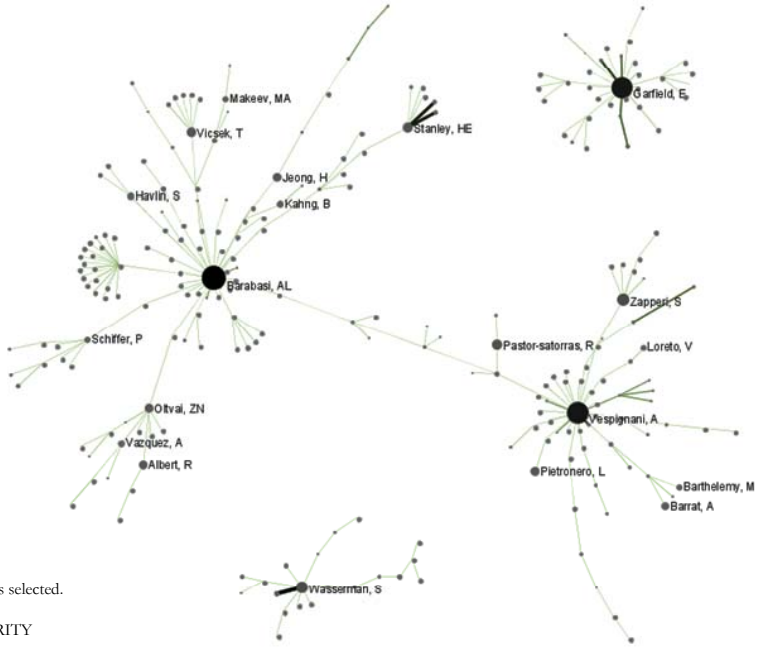
Input Parameters:
Number of bins: 10

umber of bins: 10
.....



Network Visualization:
Reduced Network After Pathfinder Network Scaling

Network Science researchers

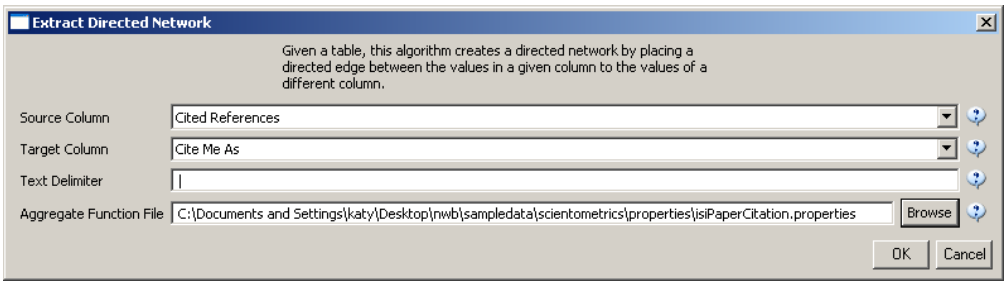


.....
MST-Pathfinder Network Scaling was selected.
Input Parameters:
Weight Attribute measures: SIMILARITY
Edge Weight Attribute: weight
.....

Paper-Citation Network Layout

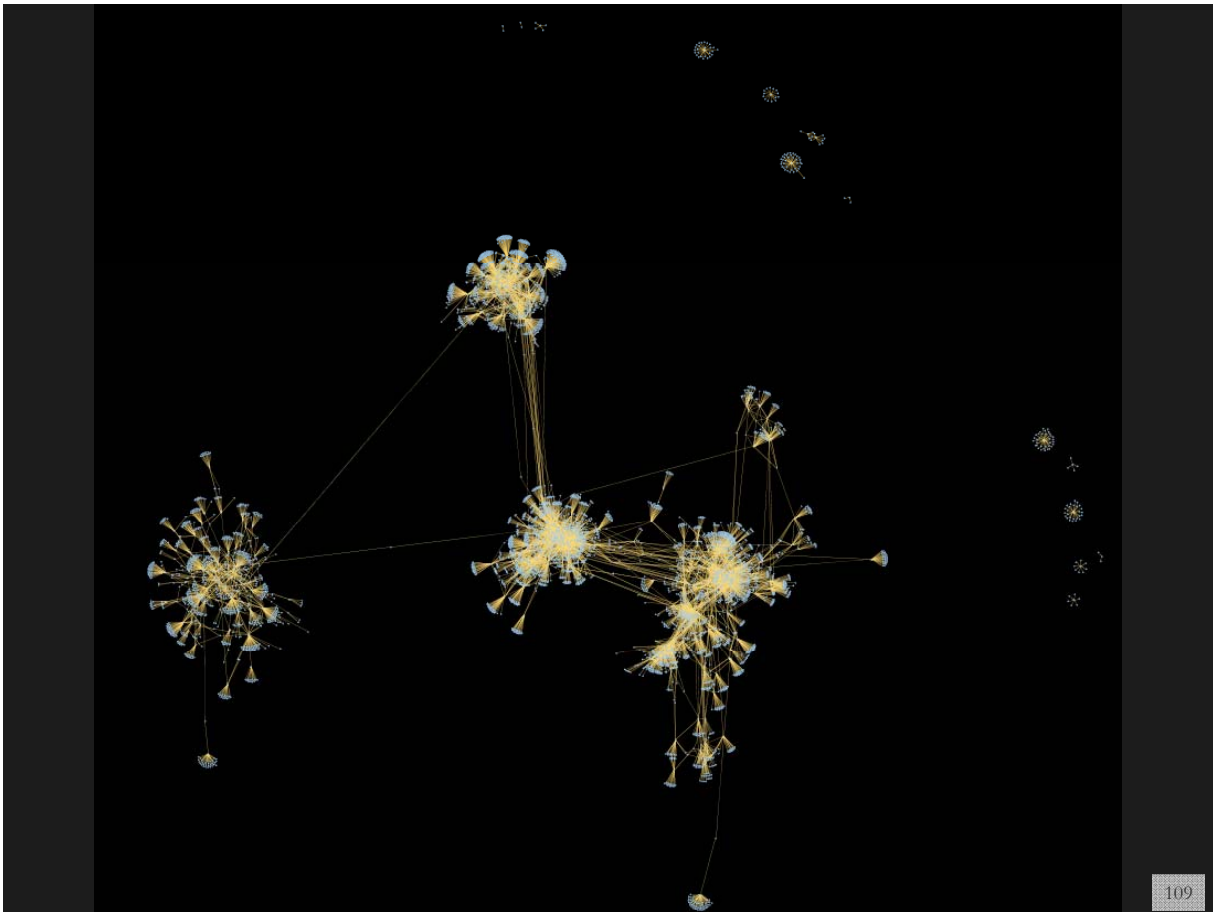
Load *'*yournwbdirectory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi'* using *'File > Load and Clean ISI File'*.

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



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'*.

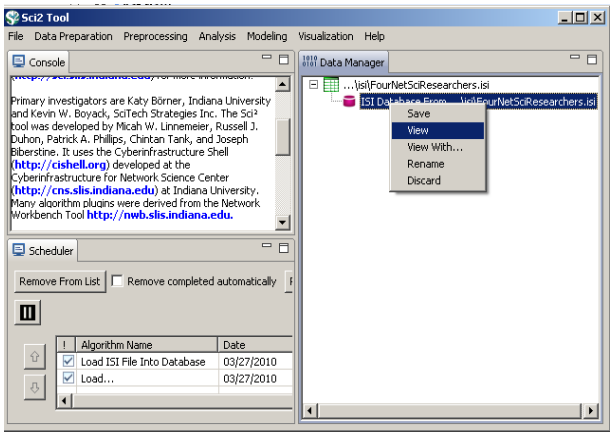


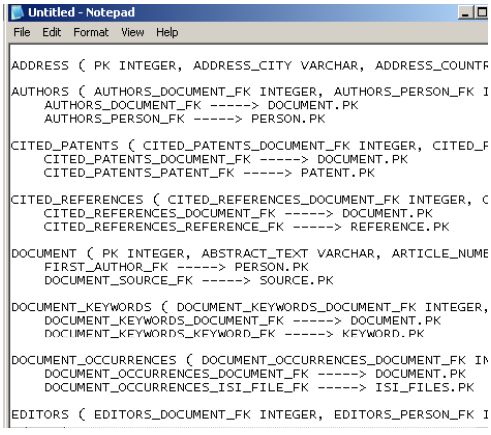
Studying Four Major NetSci Researchers (ISI Data) using Database *(section 5.1.5)*

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".

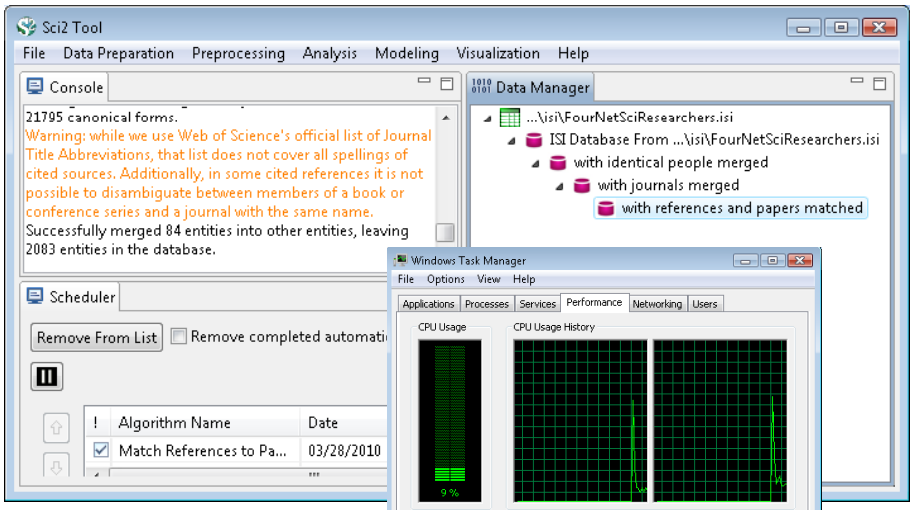




Studying Four Major NetSci Researchers (ISI Data) using Database *(section 5.1.5)*

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.



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_(	LOCAL_CI	ADDITIONA	FAMILY_NAME	FIRST_INI	FULL_NAM	MIDDLE_I	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, QK	1	41	0		Beg	Q		K		
19	Pudovkin, AI	5	32	6		Pudovkin	A		I		

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

Using Database Support – Records over time

Aggregate data by year by running '*Data Preparation* > *Database* > *ISI* > *Extract Longitudinal Summary*.' Result is a table which lists metrics for every year mentioned in the dataset. The longitudinal study table contains the volume of documents and references published per year, as well as the total amount of references made, the amount of distinct references, distinct authors, distinct sources, and distinct keywords per year.

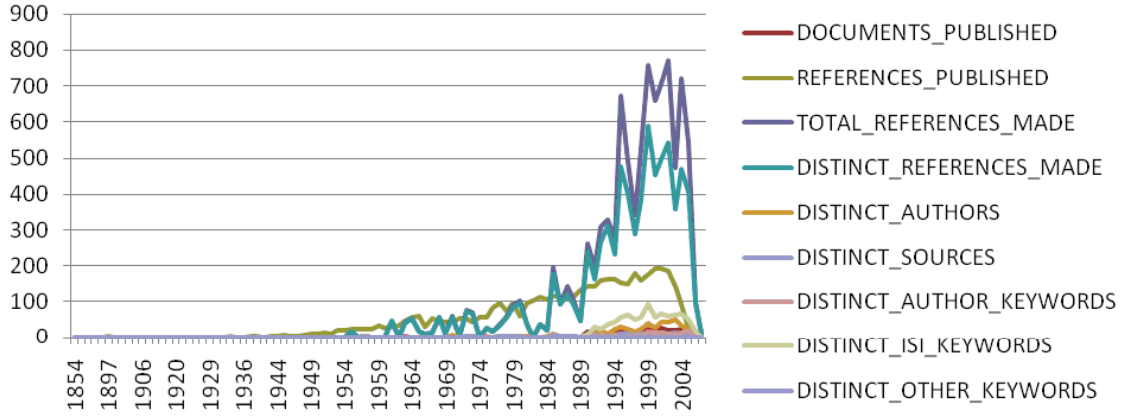
F1 DISTINCT_AUTHORS

	A	B	C	D	E	F	G	H	I	J	K	L
1	YR	DOCUMENTS_	REFERENCES_I	TOTAL_REFERI	DISTINCT_REF	DISTINCT_AU	DISTINCT_	DISTINCT_	DISTINCT_	DISTINCT_	OTHER_KEYWOR	
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		

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

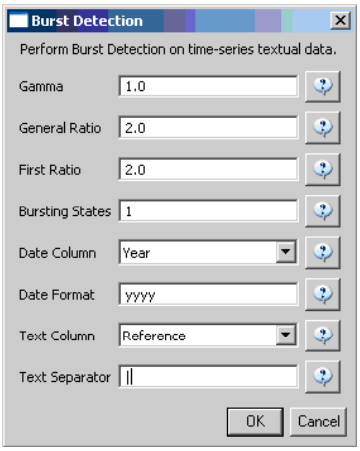
Using Database Support – Records over time

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.

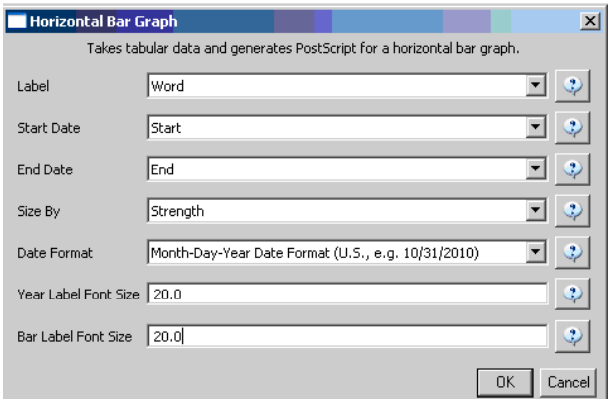


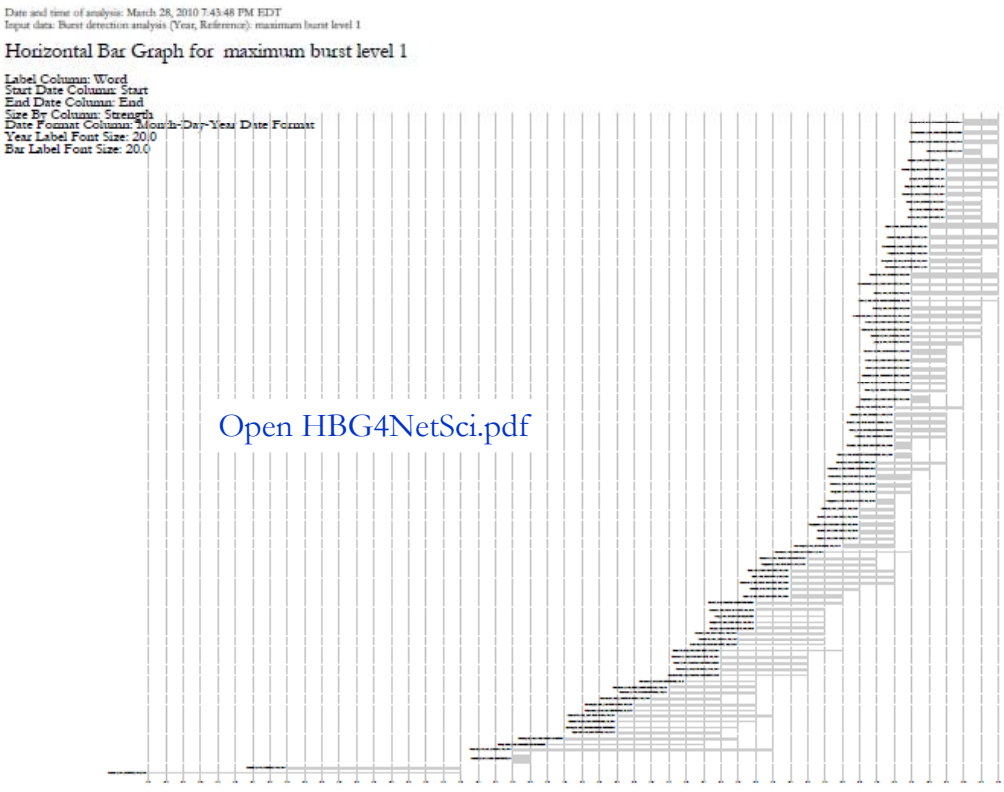
Using Database Support – Burst Analysis for References

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.



Watch those red warnings!

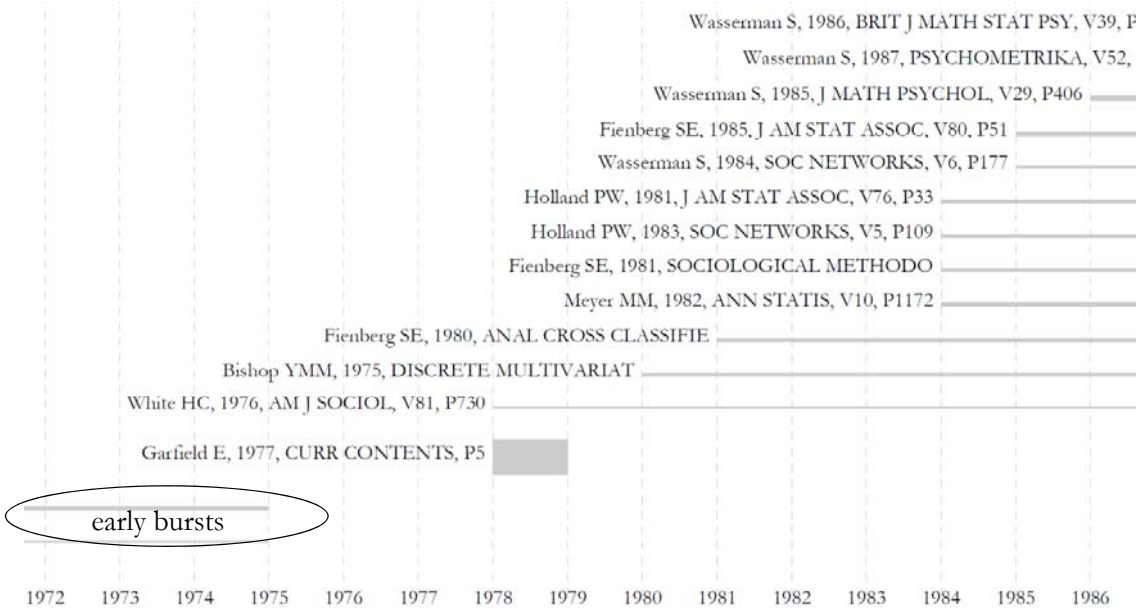




Cyberinfrastructure for Network Science Center (2009) Horizontal Bar Graph. Indiana University, http://sci.slis.indiana.edu

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

Using Database Support – Burst Analysis Result

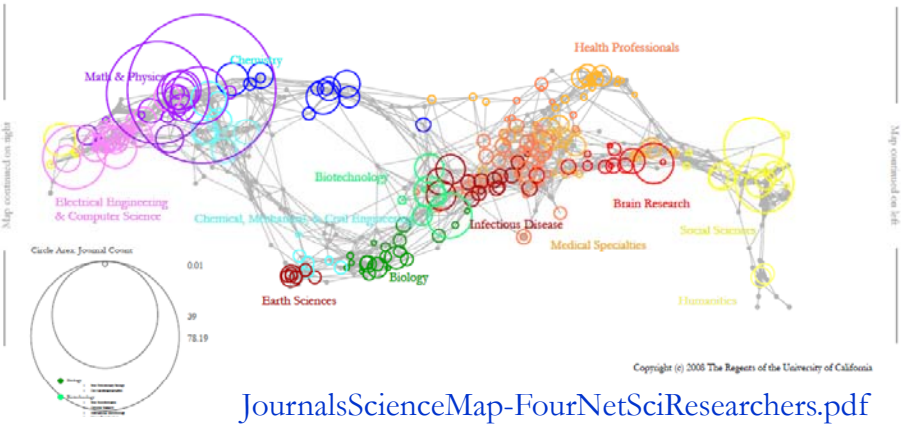


Topic Mapping: UCSD Science Map

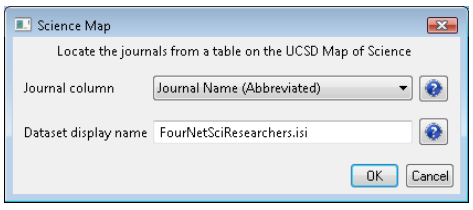
Science Map via Journals for FourNetSciResearchers.isi

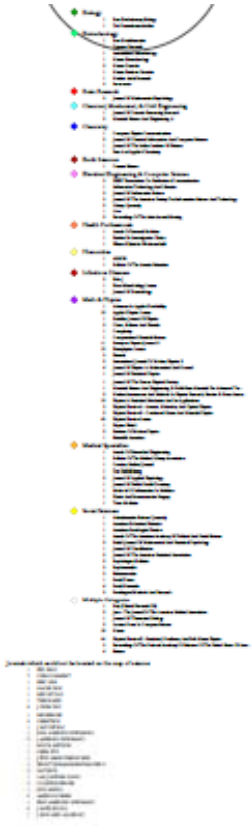
314 journal references matched out of 361 found.

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





Mapping CTSA Centers (NIH RePORTER Data)

(section 5.2.3)

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

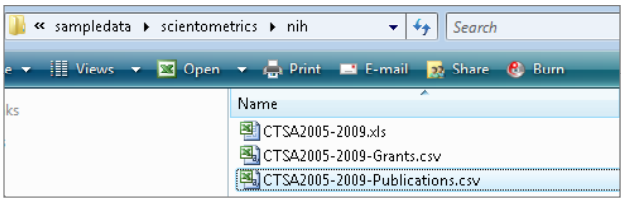
A study of all NIH Clinical and Translational Science Awards (CTSA) awards and resulting publications from 2005-2009, requires advanced data acquisition and manipulation to prepare the required data. Data comes from the union of NIH RePORTER downloads (see Section 4.2.2.2 NIH RePORTER) and NIH ExPORTER 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*.

Mapping CTSA Centers (NIH RePORTER Data)

(section 5.2.3)

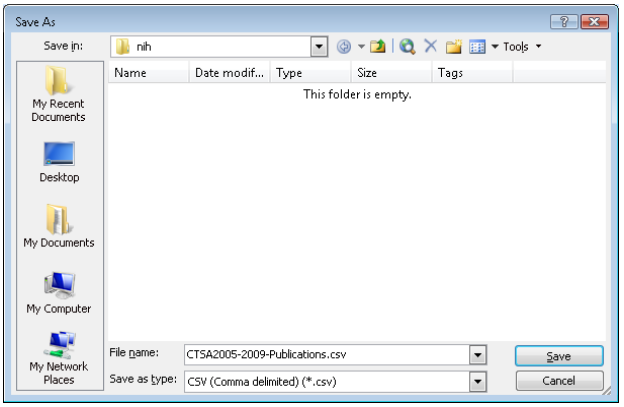
Data Preparation

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



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: |

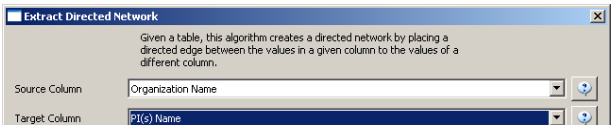
Target Column: PI(s) Name

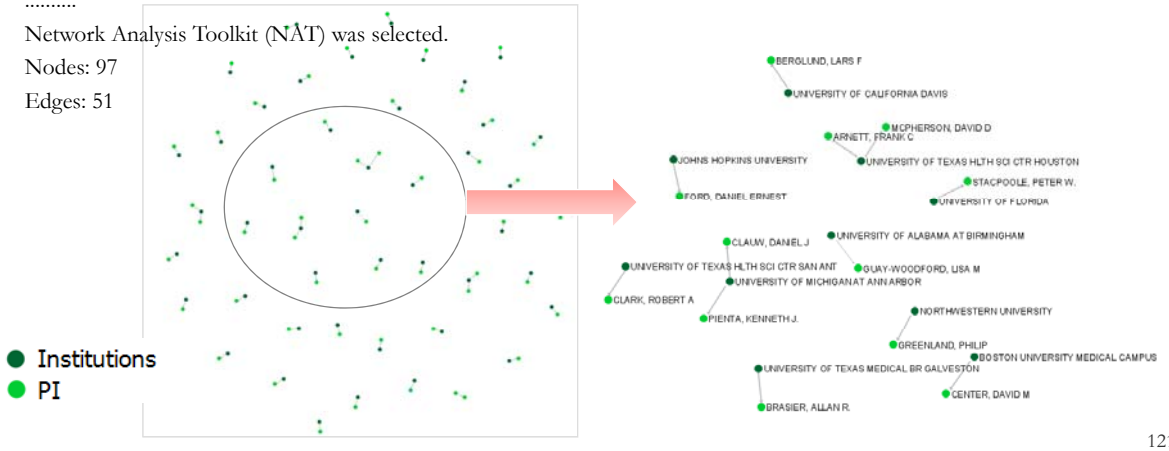
.....

Network Analysis Toolkit (NAT) was selected.

Nodes: 97

Edges: 51



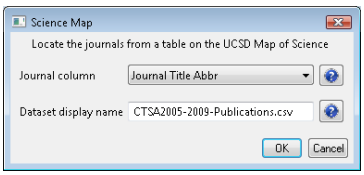


Mapping CTSA Centers (NIH RePORTER Data)

(section 5.2.3)

Topic Coverage of Publications

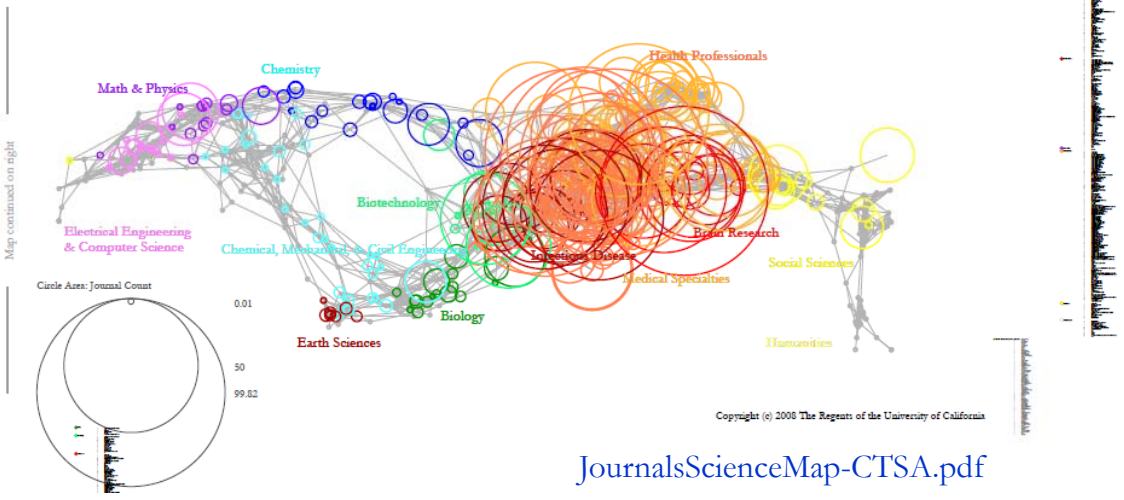
Visualization > Topical > Science Map via Journals'



Science Map via Journals for CTSA2005-2009-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 Scienc

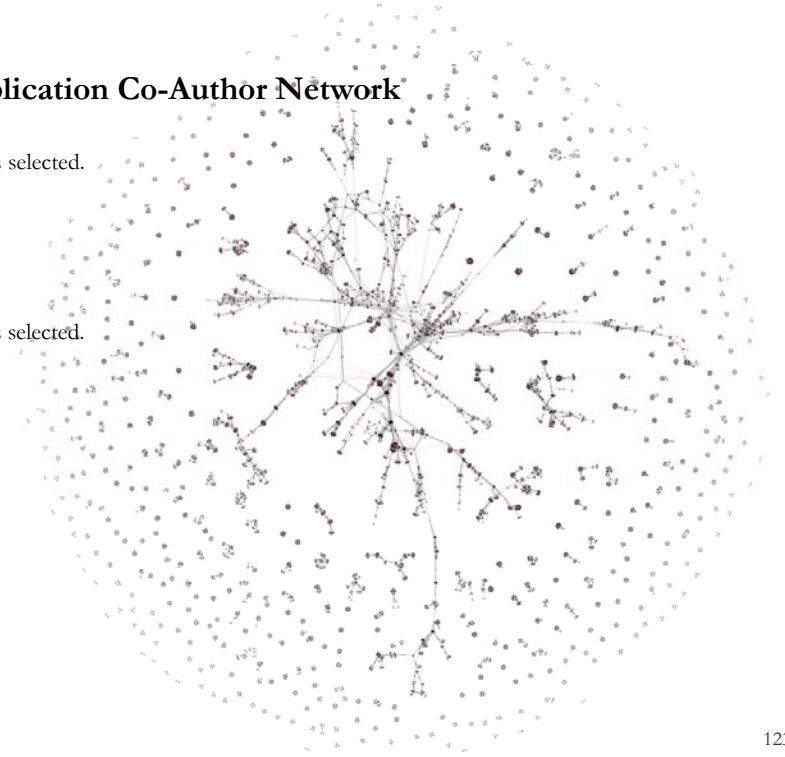


JournalsScienceMap-CTSA.pdf

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



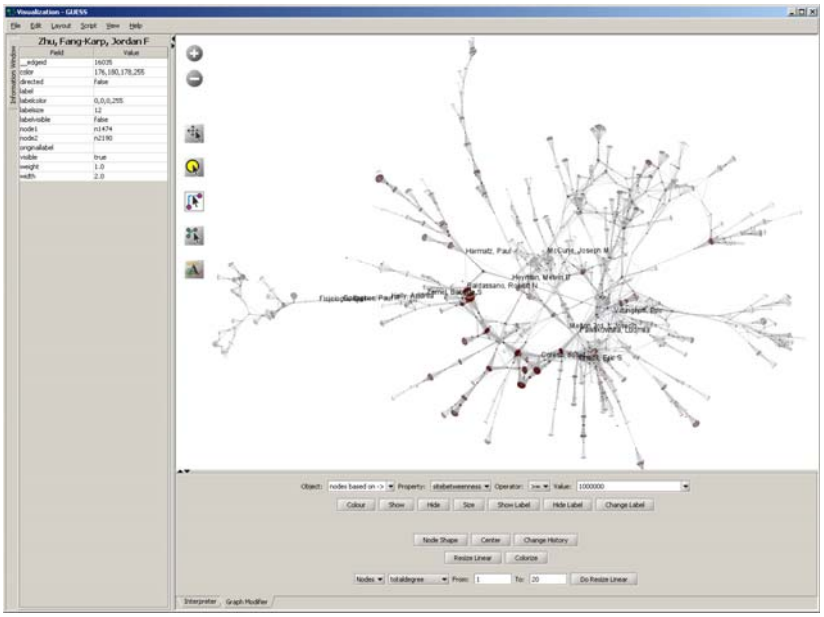
Mapping CTSA Centers (NIH RePORTER Data)
(section 5.2.3)

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



Biomedical Funding Profile of NSF (NSF Data)

(section 5.2.4)

MedicalAndHealth.nsf	
Time frame:	2003-2010
Region(s):	Miscellaneous
Topical Area(s):	Biomedical
Analysis Type(s):	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).

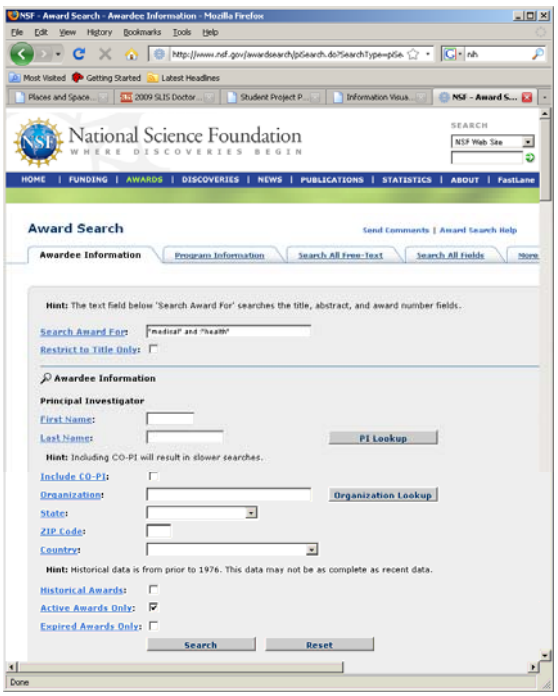
Biomedical Funding Profile of NSF (NSF Data)

(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

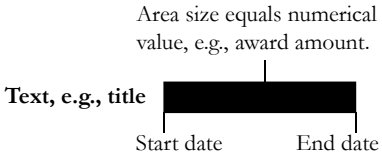
Retrieved on Oct 18, 2009



Biomedical Funding Profile

(section 5.2.4)

Horizontal Bargraph

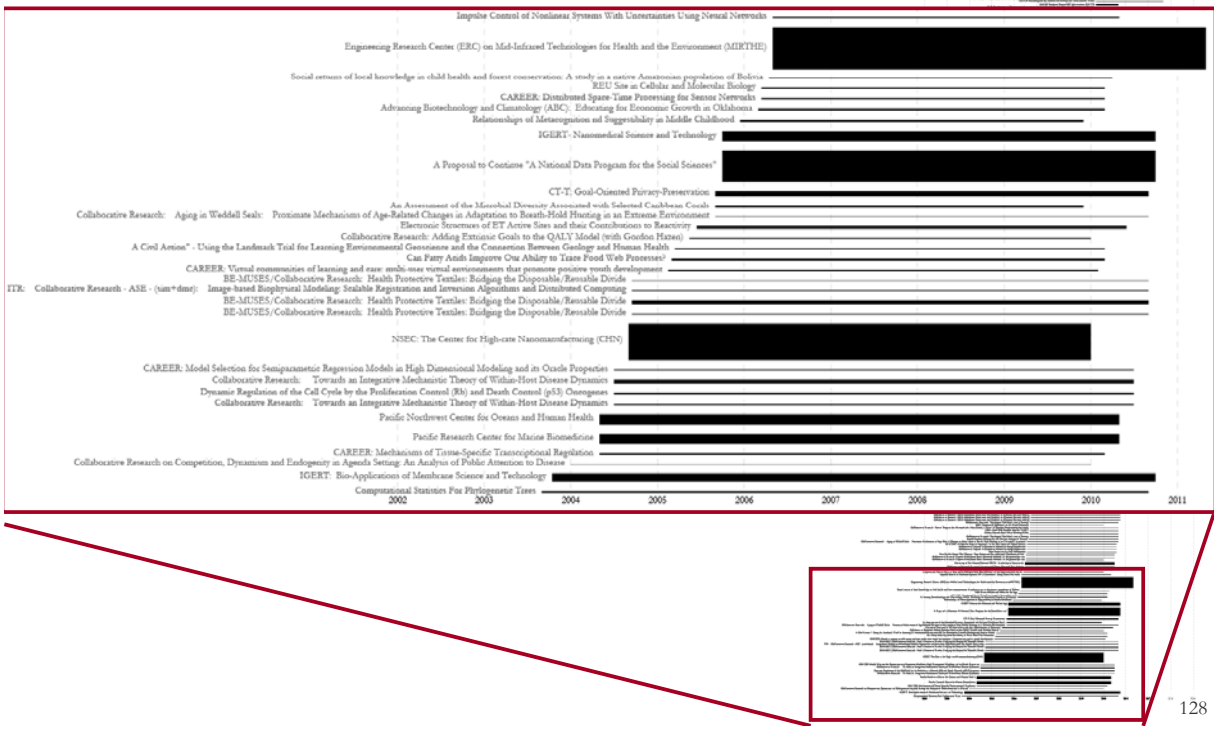


Top-10 grants with highest $Awarded to Date:

Title	NSF Org.	Program(s)	PI	State	Organization	$ Awarded to Date
University of NewMexico/Harvard PREM: Leadership in Bi	DMR	PREM|MATERIAL	Lopez, Gabriel	NM	University of New Mexico	2,037,500
TC: Large: Trustworthy Information Systems for Healthcare	CNS	TRUSTWORTHY C	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 Techr	DGE	HUMAN RESOURI	Fried, Joel	OH	University of Cincinnati Main Campu	3,644,410
Pacific Research Center for Marine Biomedicine	OCE	CHEMICAL OCEA	Laws, Edward	HI	University of Hawaii	3,816,943
Pacific Northwest Center for Oceans and Human Health	OCE	CHEMICAL OCEA	Faustman, Elaine	WA	University of Washington	4,026,968
A Proposal to Continue 'A National Data Program for the S	SES	|SCIENCE & ENG	Smith, Tom	IL	National Opinion Research Center	5,835,140
A Proposal to Continue "A National Data Program for the S	SES	|||SCIENCE & EN(	Davis, James	IL	National Opinion Research Center	10,053,668
NSEC: The Center for High-rate Nanomanufacturing (CHN)	EEC	Studies of Policy S	Busnaina, Ahmed	MA	Northeastern University	13,047,758
Engineering Research Center (ERC) on Mid-Infrared Techn	EEC	COLLABORATIVE	Gmachl, Claire	NJ	Princeton University	13,681,994

Biomedical Funding Profile

(section 5.2.4)



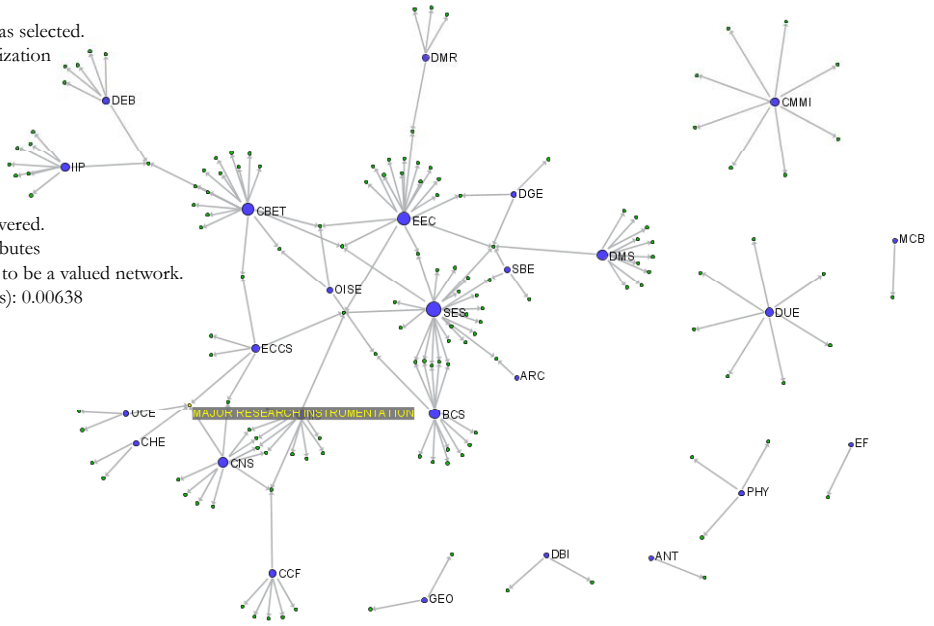
Biomedical Funding Profile of NSF (NSF Data)

(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



Mapping the Field of RNAi Research (SDB Data)

(section 5.2.7)

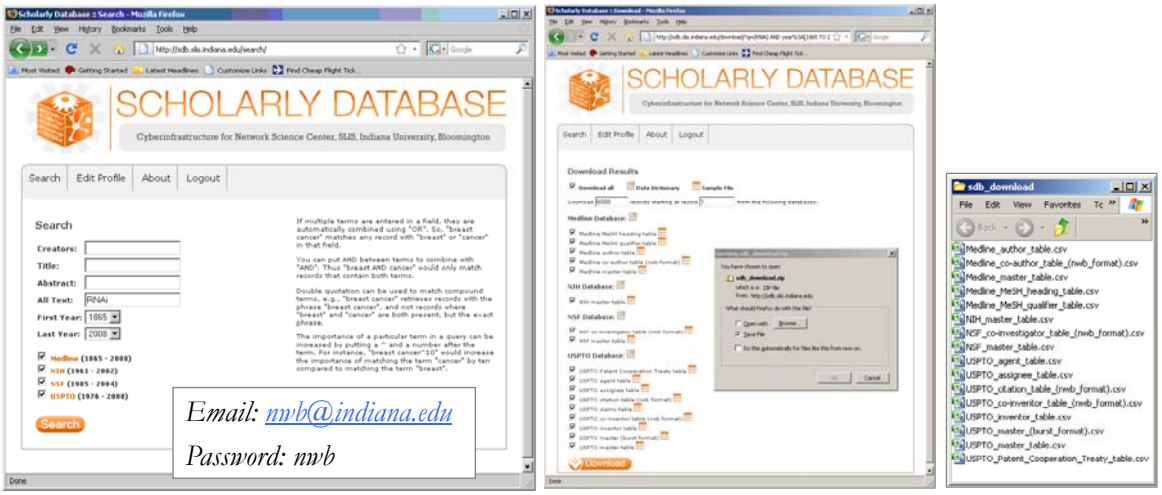
RNAi	
Time frame:	1865-2008
Region(s):	Miscellaneous
Topical Area(s):	RNAi
Analysis Type(s):	Co-Author Network, Patent-Citation Network, Burst Detection

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.

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



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)

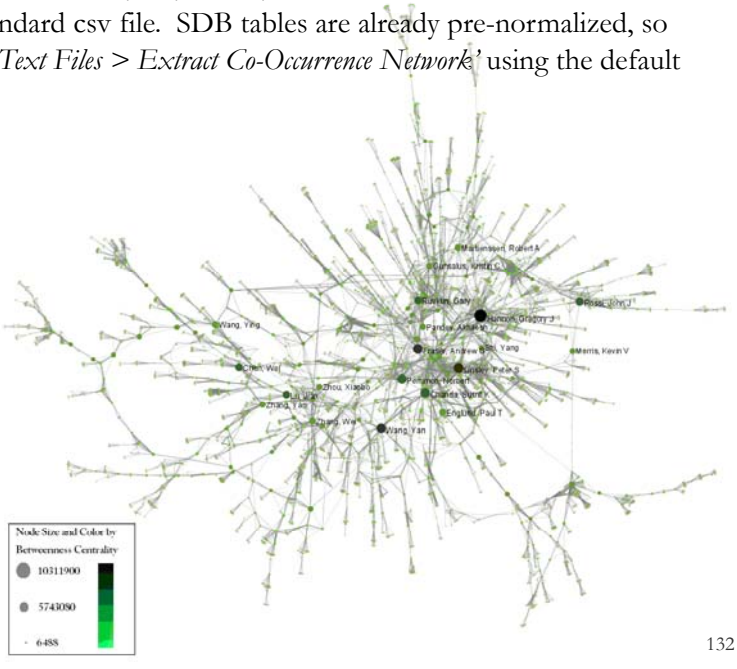
Co-Author Network

Load '**yoursci2directory*/sampledata/scientometrics/sdb/RNAi/Medline_co-author_table_(nwb_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.*'

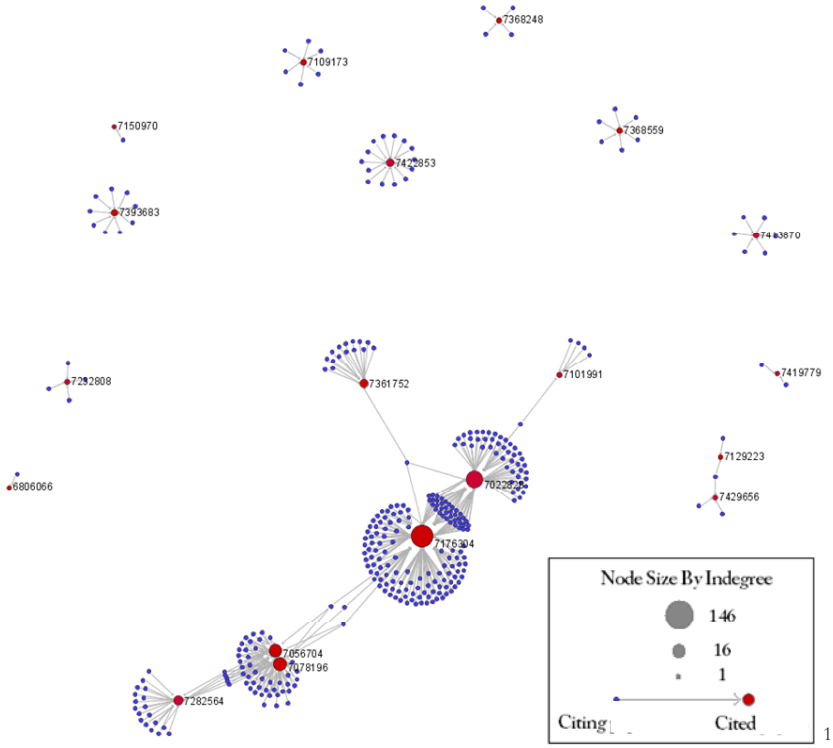
Visualize with *GUESS* using '*Layout* > *GEM*'.
Use a custom python script to color and size the network.



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.



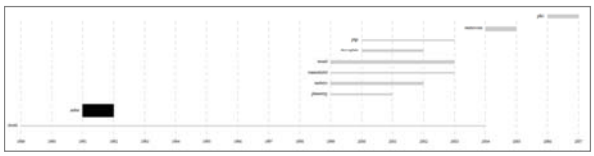
Mapping the Field of RNAi Research (SDB Data)
(section 5.2.7)

Topic Bursts

- Load '**yoursci2directory*/sampledat/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*.'

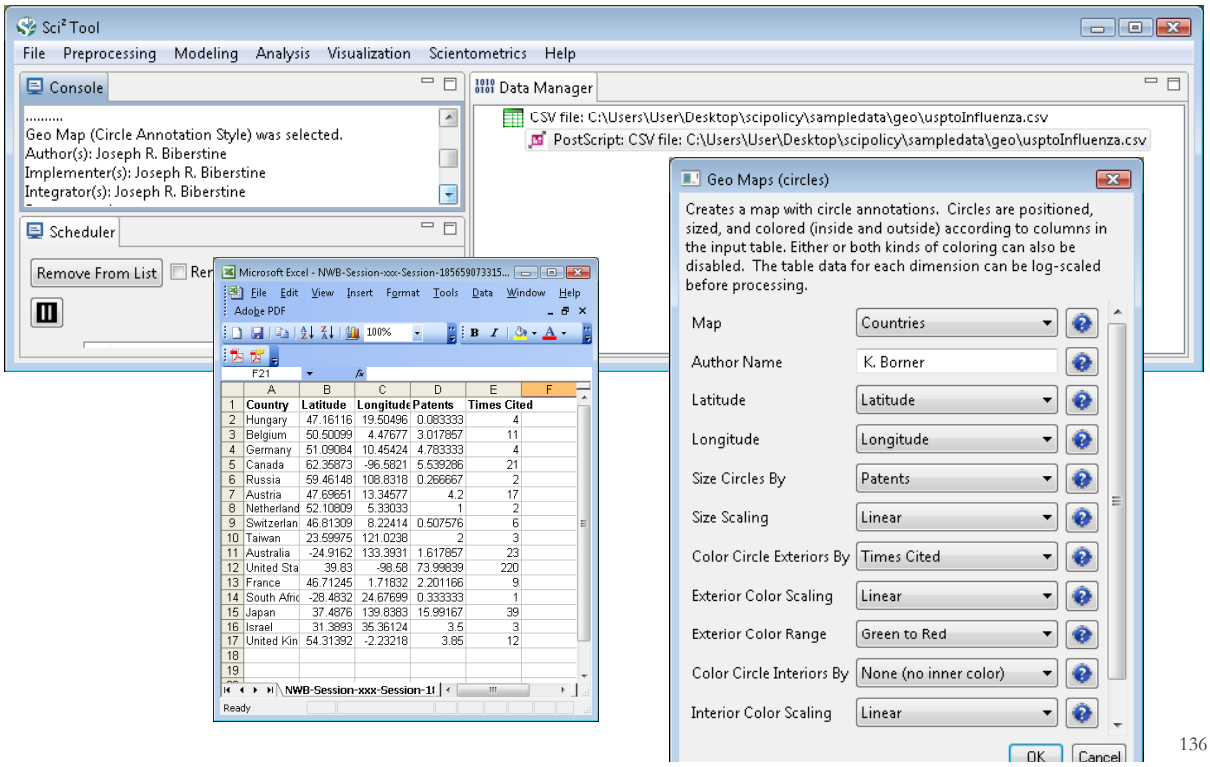


Geo USPTO (SDB Data)
(section 5.3.1)

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

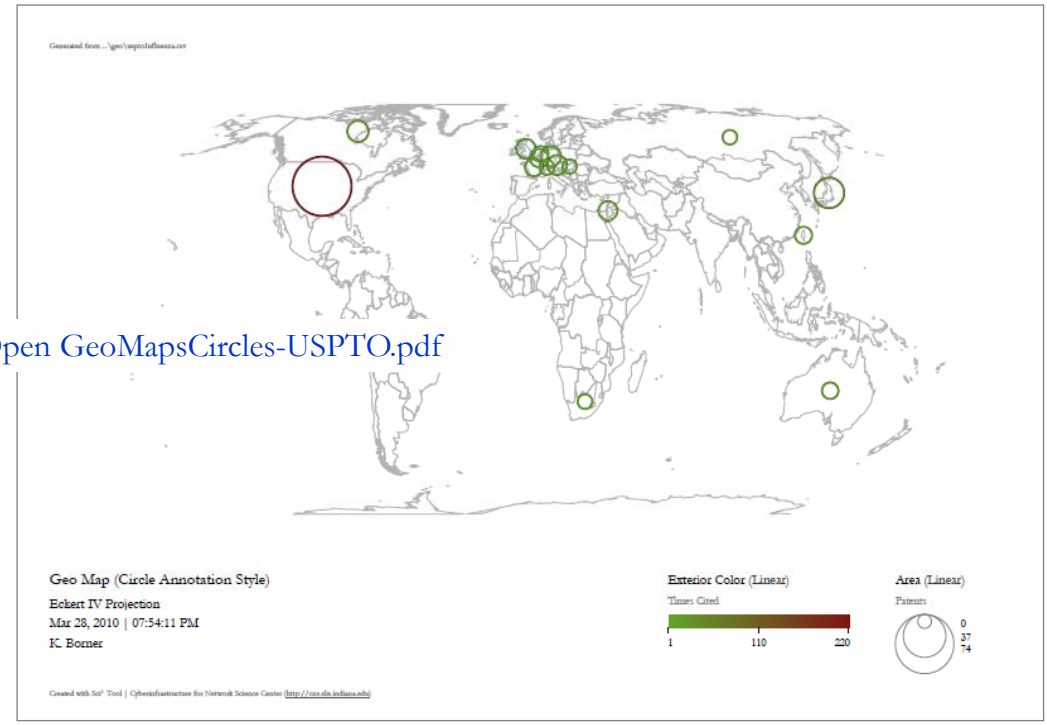
The file '*usptoInfluenza.csv*' was generated with an SDB search for patents containing the term "Influenza", and was heavily modified to produce a simple geographic table. Load it using '*File* > *Load* > *sampledata* > *geo* > *usptoInfluenza.csv*' and then select 'Standard csv format'. See the data 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).

Geo USPTO (SDB Data)
(section 5.3.1)



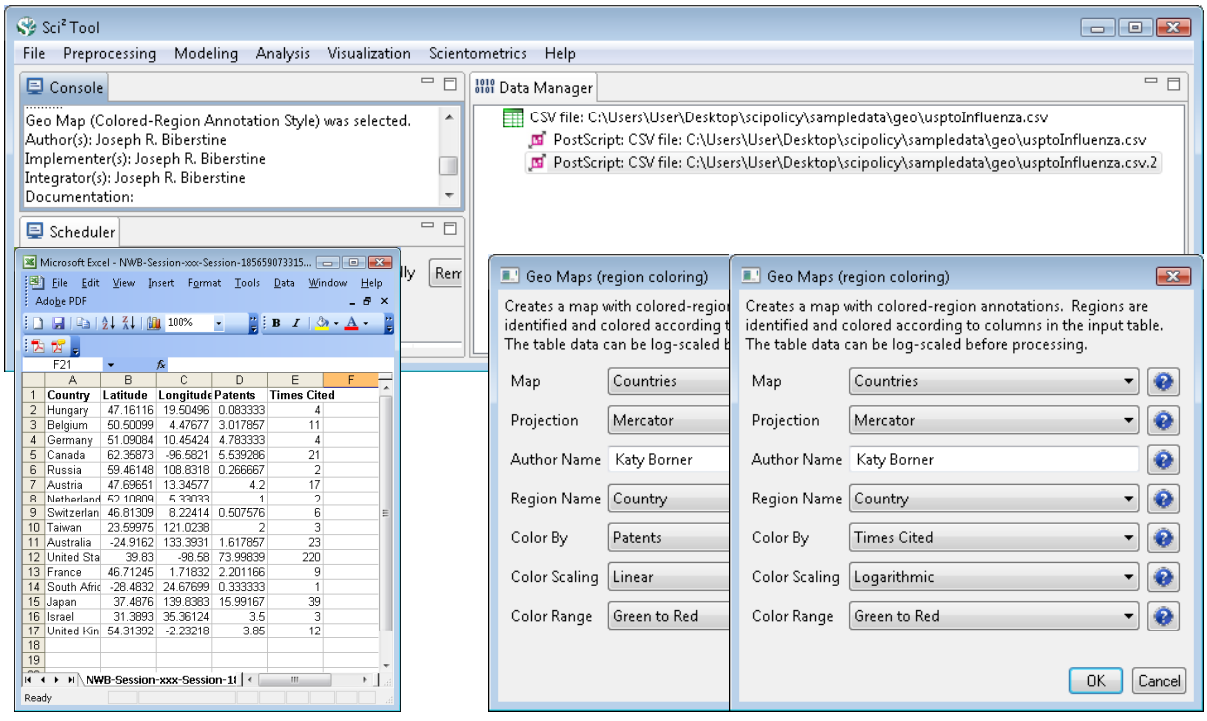
Geo USPTO (SDB Data) – Circle Coding

(section 5.3.1)



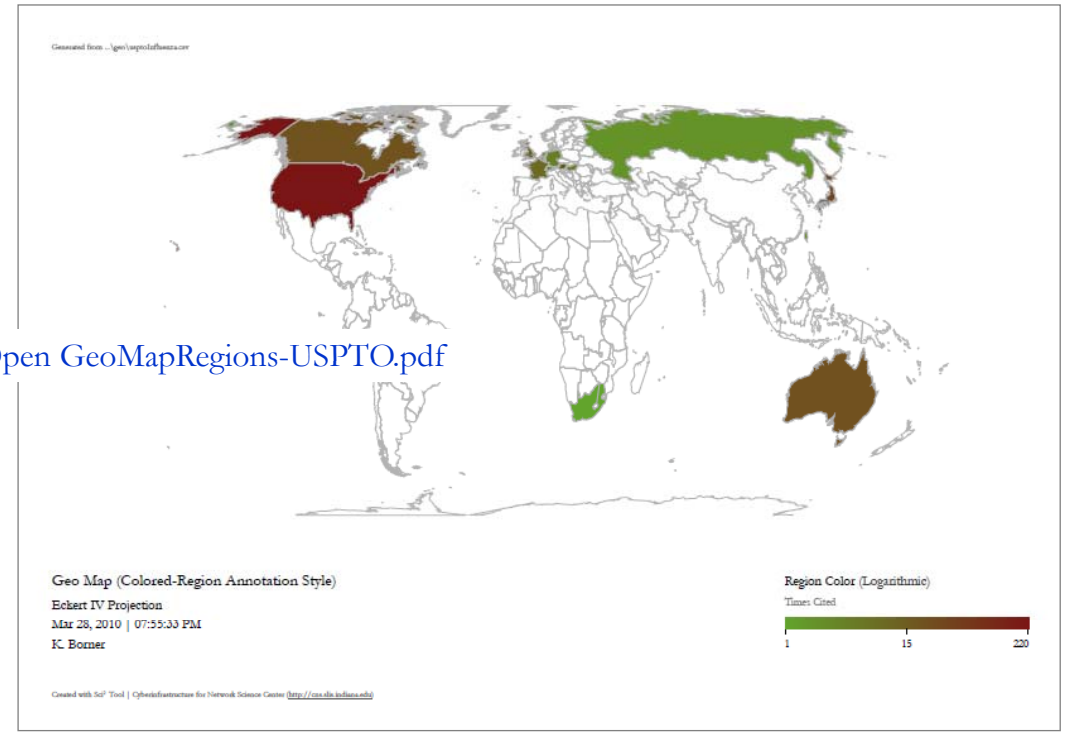
Open GeoMapsCircles-USPTO.pdf

Geospatial Maps – Region Coding

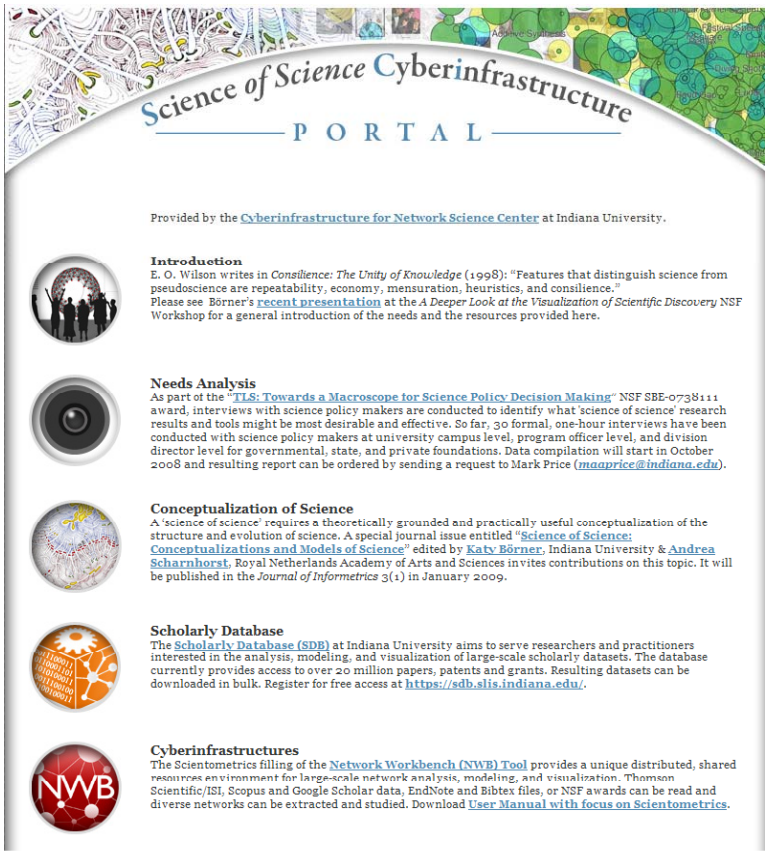


Geo USPTO (SDB Data) – Region Coding

(section 5.3.1)



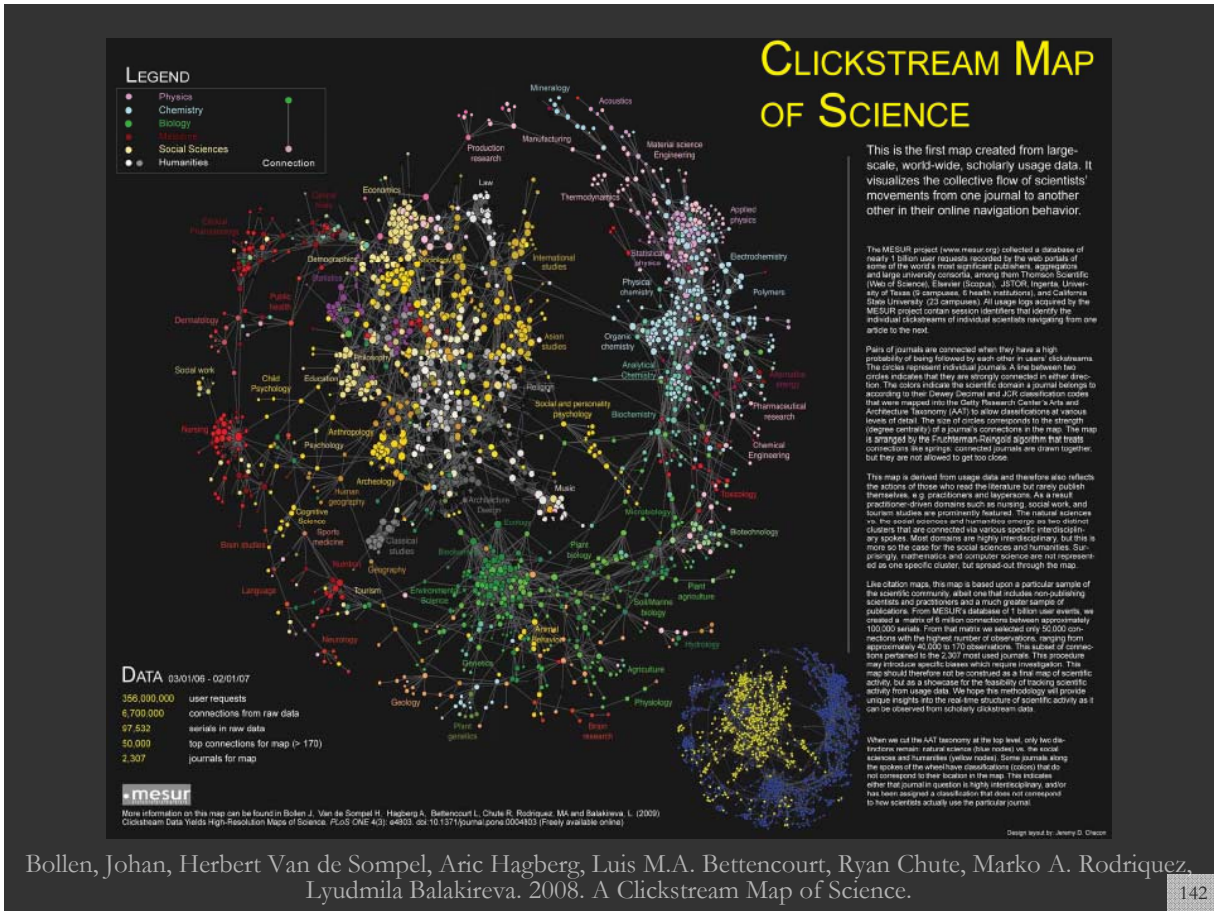
Open GeoMapRegions-USPTO.pdf



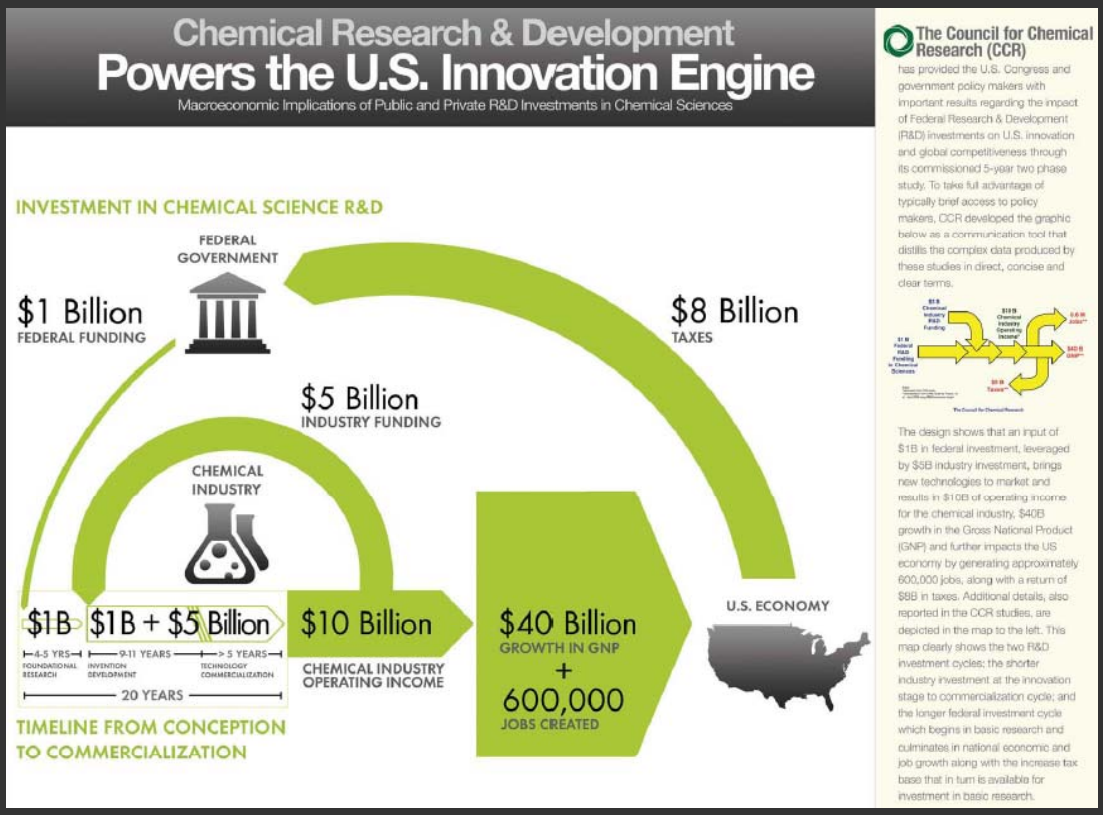
http://sci.slis.indiana.edu

Overview

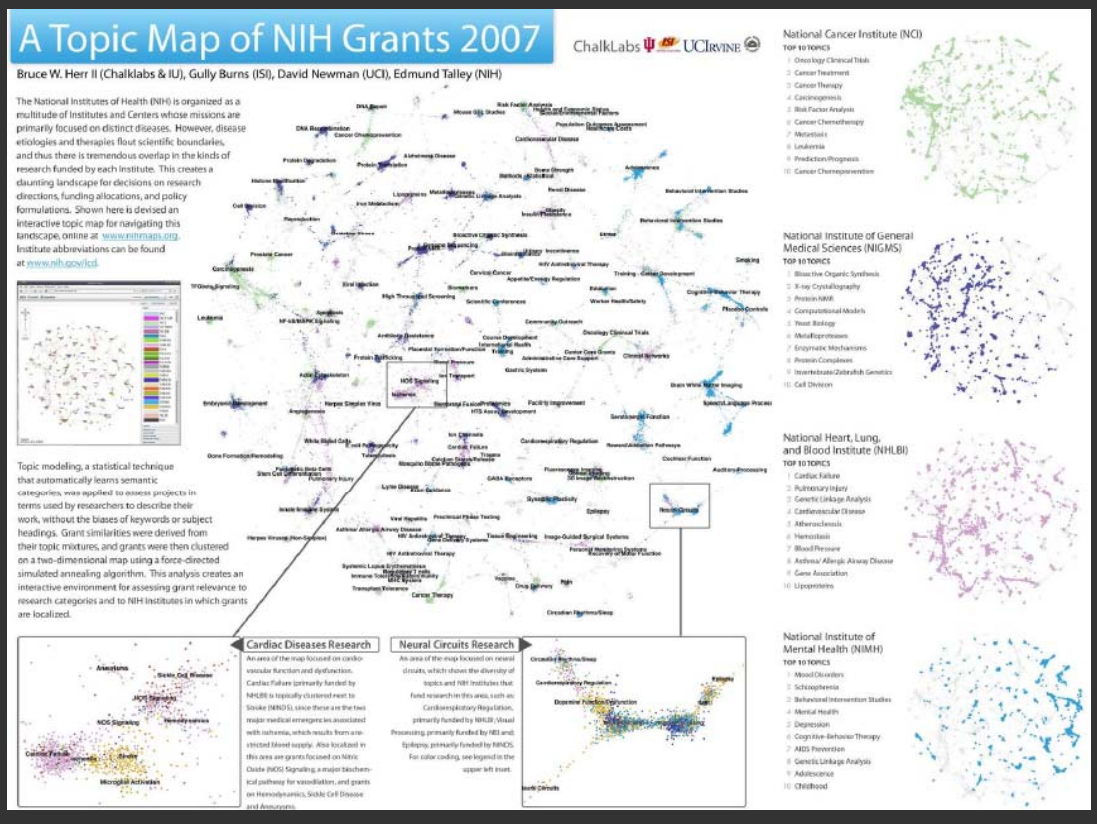
- **Introduction to the Sci² Tool**
- Demo and hands-on data analysis and visualization by participants
- Outlook
- Overview of validation approaches for science studies



Bollen, Johan, Herbert Van de Sompel, Aric Hagberg, Luis M.A. Bettencourt, Ryan Chute, Marko A. Rodriquez, Lyudmila Balakireva. 2008. A Clickstream Map of Science.



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



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 23rd, 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.

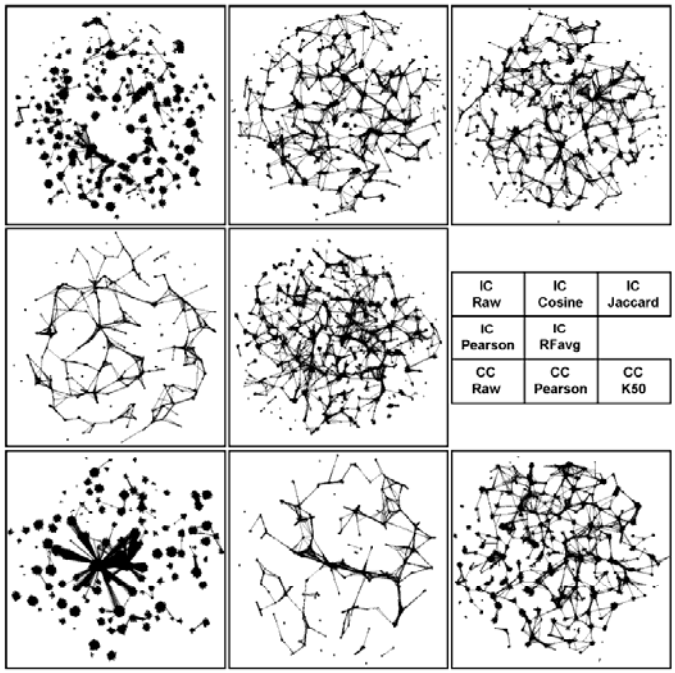


Table 1. Summary of validation results for maps based on eight similarity measures.

Measure	Local accuracy @ 95% coverage[1]	Scalability[1]	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.

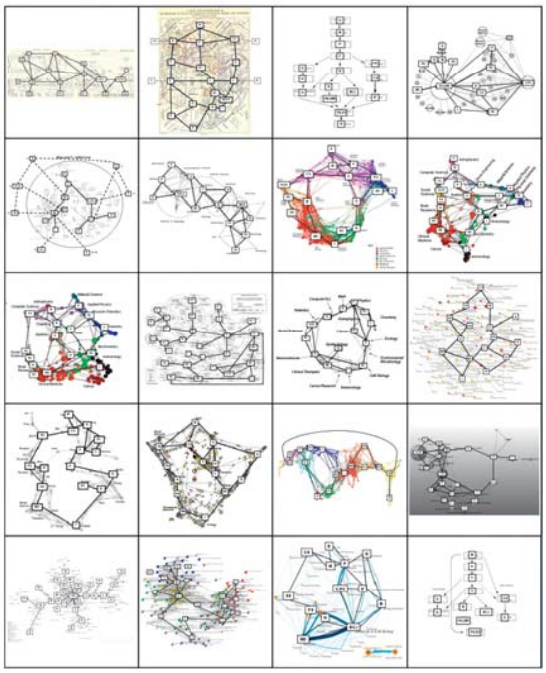


Table 1: Characteristics of twenty comprehensive maps of science. Abbreviations SC, SS, AH, and PR refer to Thomson Scientific's Science, Social Science, Arts & Humanities, and Proceedings Citation databases, respectively.

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) [a]	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
(Bassecoulard & 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) [b]	UCSD	Journals	16,235 jnl	554	SC/SS/AH + Scopus, 2001-05	Non-centric
(Rosvall & Bergstrom, 2008) [c]	Rosvall	Journals	6,116 jnl	87	SC+SS, 2004	Non-centric
(Moya-Anegón et al., 2004)	Scimago-I	Journal categories	25 categ	25	SC+SS+AH, 2000 Spanish papers	Non-centric
(Moya-Anegón et al., 2007) [d]	Scimago-II	Journal categories	219 categ	219	SC+SS+AH, 2002	Centric
(Leydesdorff & Rafols, 2008) [e]	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

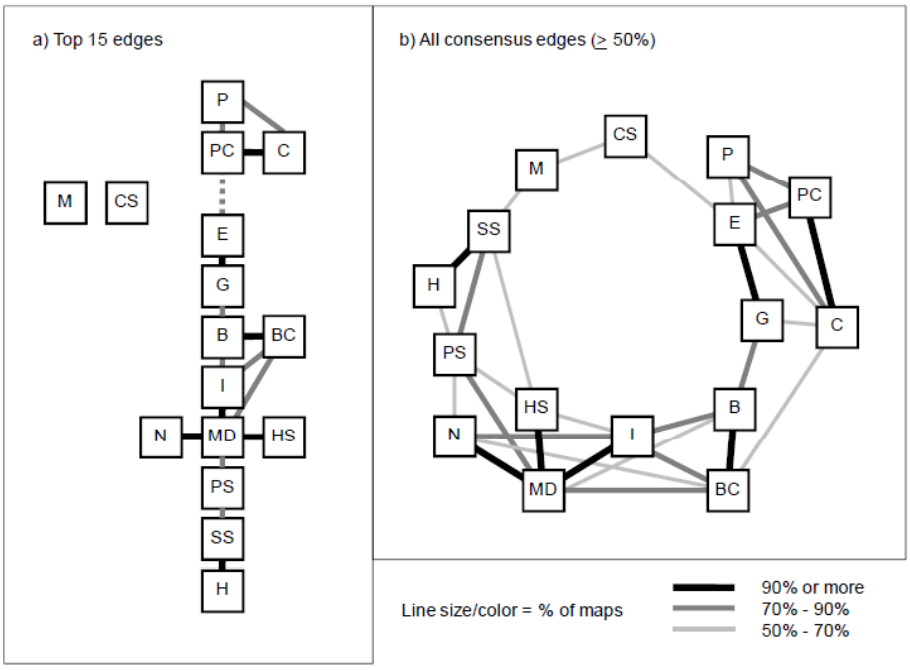


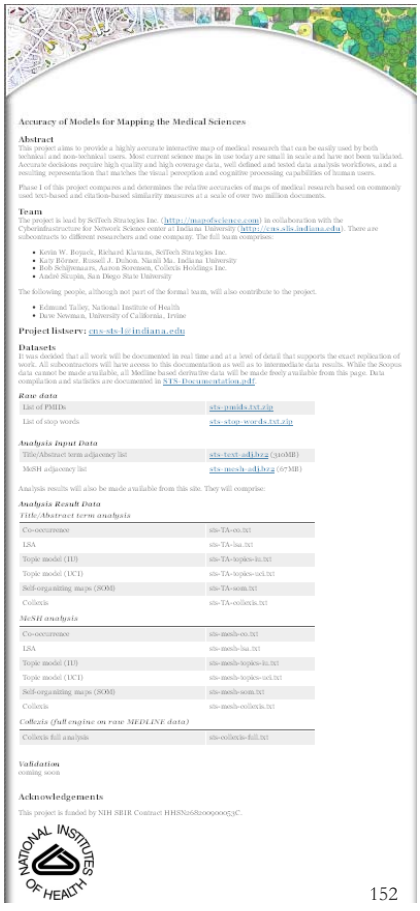
Figure 5: Two-dimensional consensus maps of science from twelve of the twenty input maps, excluding the eight input maps from Klavans, Boyack, and Börner.

Validating Science Maps

Accuracy of Models for Mapping the Medical Sciences
Kevin W. Boyack, Richard Klavans, SciTech Strategies Inc. Katy Börner, Russell J. Duhon, 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.


Accuracy of Models for Mapping the Medical Sciences

Abstract

Team

Project listserv: cns-sts-l@indiana.edu

Datasets

Raw data

Analysis Input Data

Analysis Result Data

Validation

Acknowledgements

Overview

- Macroscopes and the Changing Scientific Landscape 15 mins

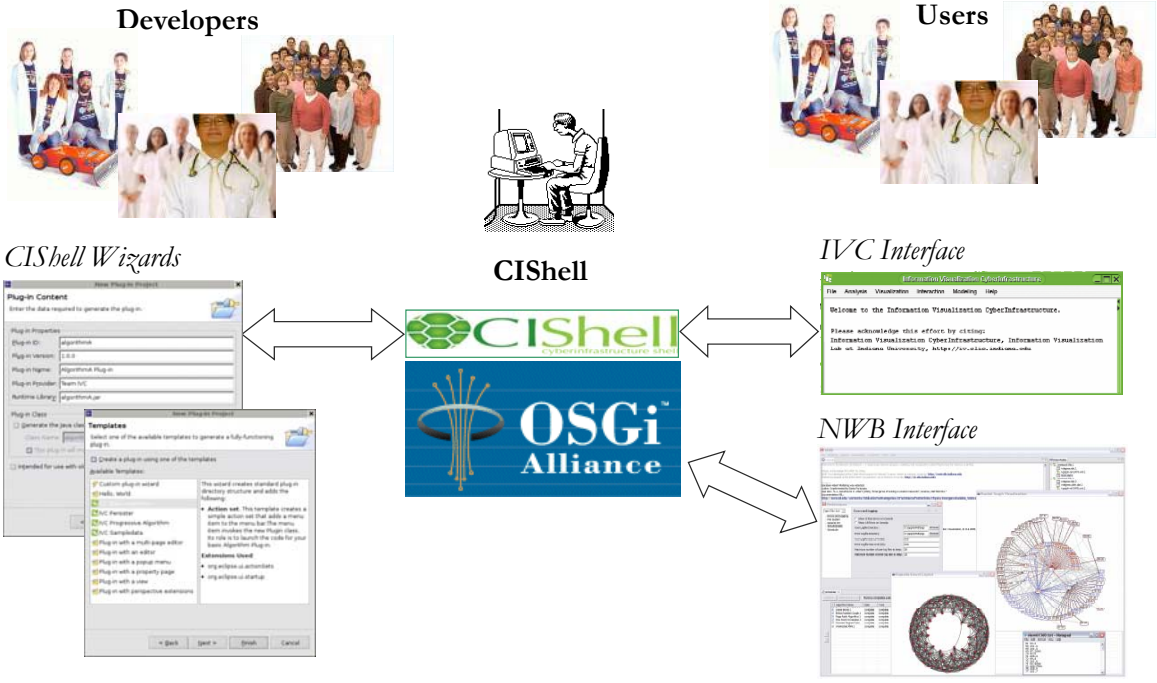
- Introduction to network science with sample maps and insights 15 mins
- Introduction to the Network Workbench Tool 15 mins
- Demo and hands-on data analysis and visualization by participants 60 mins

- Introduction to science studies with sample maps and insights 15 mins
- Introduction to the Sci^2 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

CIShell – Serving Non-CS Algorithm Developers & Users



CIShell – Builds on OSGi Industry Standard

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.

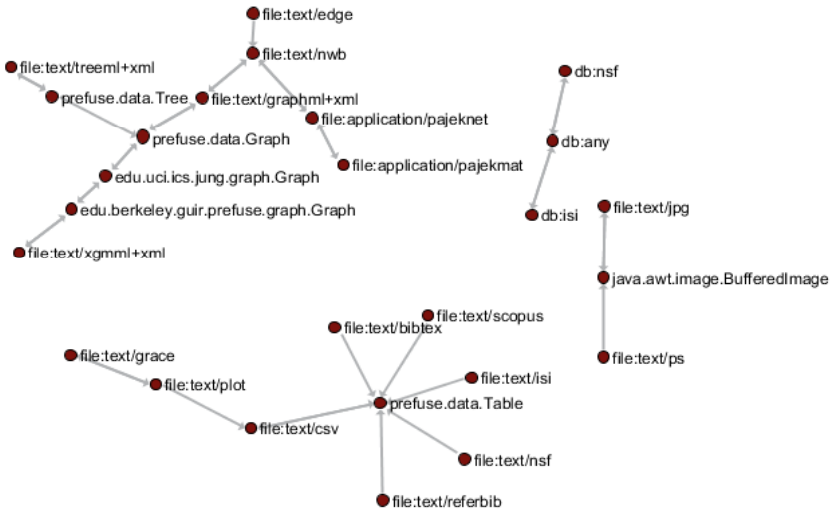
Advantages of Using OSGi

- Any CIShell algorithm is a service that can be used in any OSGi-framework based system.
- Using OSGi, running CIShells/tools can 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.

CIShell – Converter Graph

- No central data format.
- Sci^2 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).



CIShell – Add new Plugins, e.g., Cytoscape

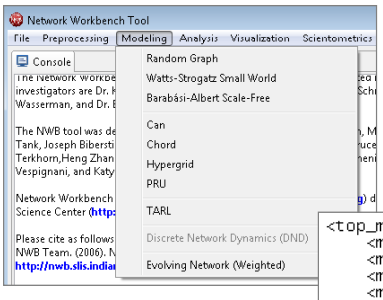
- Not all code can be shared freely (yet).
- To make the UCSD Science Map and Cytsocape tool available via the Sci^2 menu, simply add

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.

CIShell – Customize Menu

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




```
<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.TarlAlgorithm"/>
    <menu type="break"/>
    <menu pid="edu.iu.nwb.modeling.discretenetworkdynamics.DNDAlgorithm"/>
    <menu type="break"/>
    <menu pid="edu.iu.nwb.modeling.weighted.evolvingnetwork"/>
</top_menu>
```

CIShell – Integrate New Algorithms

Algorithm Developer's Guide

Overview

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

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

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

Table of Contents

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

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

Tools that Use OSGi and/or CIShell

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 (Sci² 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.

Overview

- Macroscopes and the Changing Scientific Landscape 15 mins

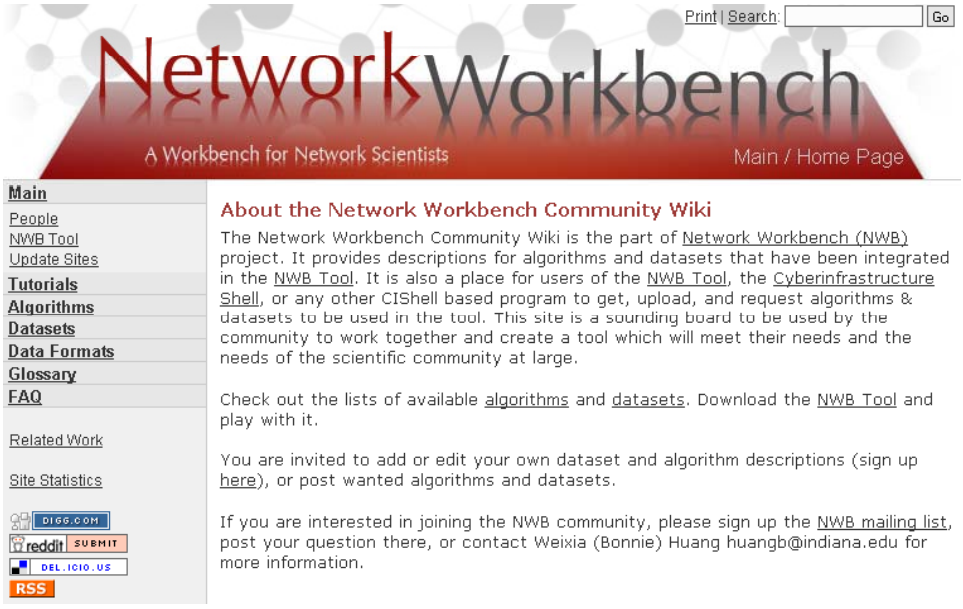
- Introduction to network science with sample maps and insights 15 mins
- Introduction to the Network Workbench Tool 15 mins
- Demo and hands-on data analysis and visualization by participants 60 mins

- Introduction to science studies with sample maps and insights 15 mins
- Introduction to the Sci² 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

Network Workbench Community Wiki



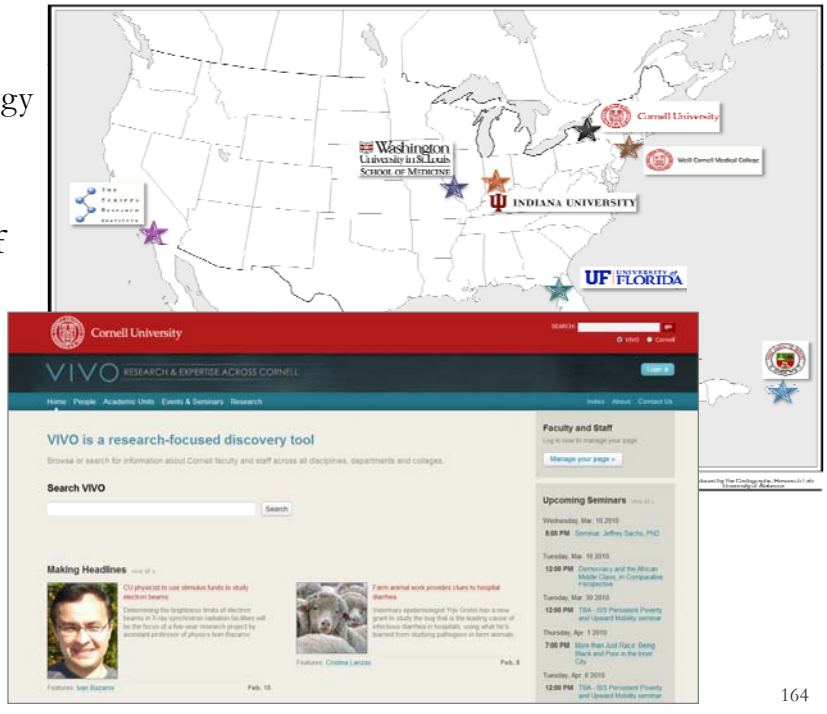
https://nwb.slis.indiana.edu/community

Epidemics Marketplace

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

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.

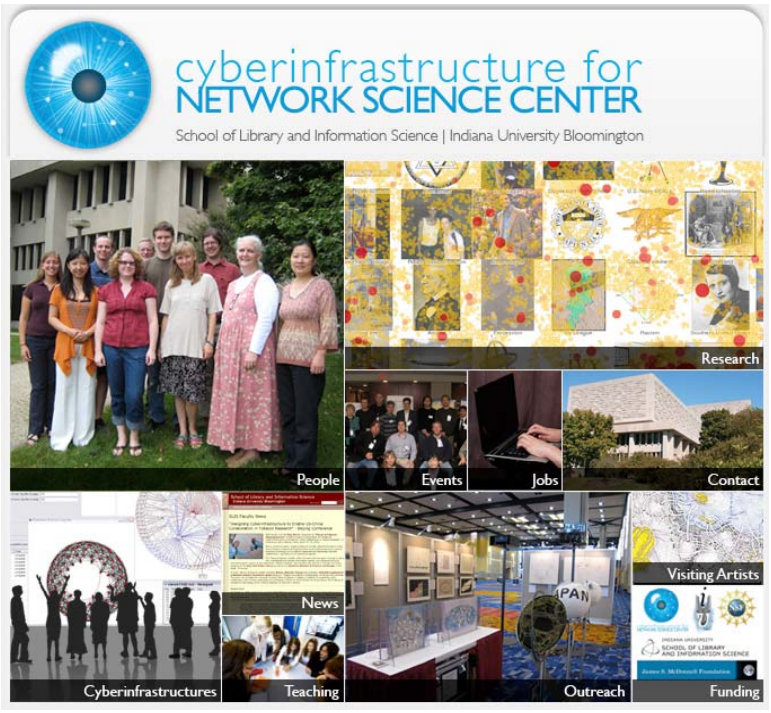


http://vivoweb.org

Question-Answer Session

All questions are welcome.

We appreciate your input to this line of tool development via the questionnaire.
Feel free to fax it back via (812) 855-6166.
Thank you.



All papers, maps, cyberinfrastructures, talks, press are linked from http://cns.slis.indiana.edu