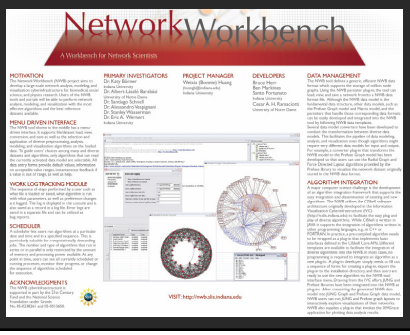
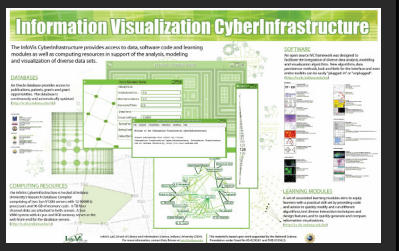
User Friendly Cyberinfrastructures

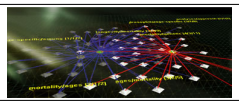
Dr. Katy Börner / Bruce Herr & Weixia (Bonnie) Huang
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
katy@indiana.edu



Workshop at the Cambridge Colloquium on Complexity and Social Networks
Harvard University, Boston, MA.
September 21, 2006

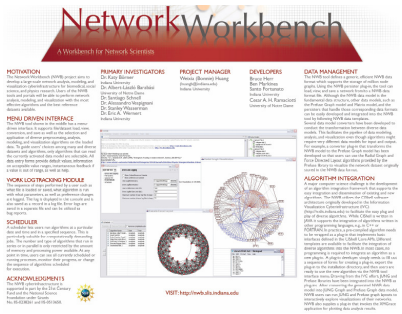


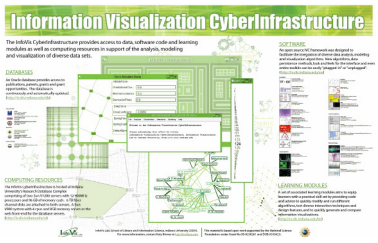




Questions

- Did you attended the Network Science Conference & Workshop 2006?
- What datasets do you use in your research? What is their size? Any NWs?
- What tools/software packages/scripting languages do you use? Any vis?
- Do you program? Develop new algorithms?
- Do you share your algorithms? How?





This Talk has Three Parts:

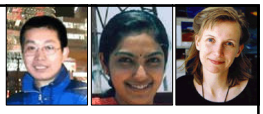
1. Why Develop CI in a InfoVis Lab / NetSci Center?
2. Cyberinfrastructure Shell (CIShell) Framework
3. Utilizing CIShell to Serve Different Communities
   - Information Visualization CI
   - Network Workbench

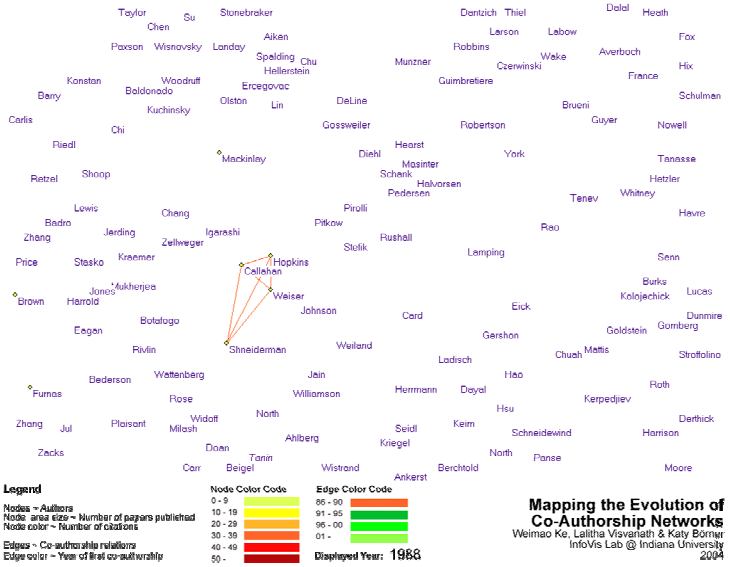
This Talk has Three Parts:

1. **Why Develop CI in a InfoVis Lab / NetSci Center?**
2. Cyberinfrastructure Shell (CIShell) Framework
3. Utilizing CIShell to Serve Different Communities
   - Information Visualization CI
   - Network Workbench

Mapping the Evolution of Co-Authorship Networks in Information Visualization, 1988 - 2004

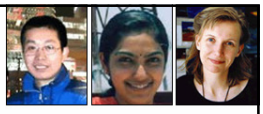
Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest.

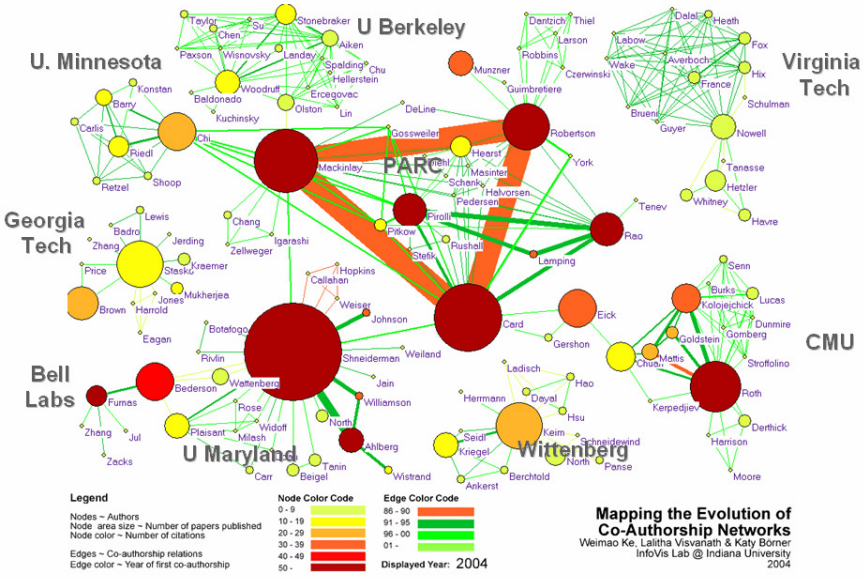




Mapping the Evolution of Co-Authorship Networks

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

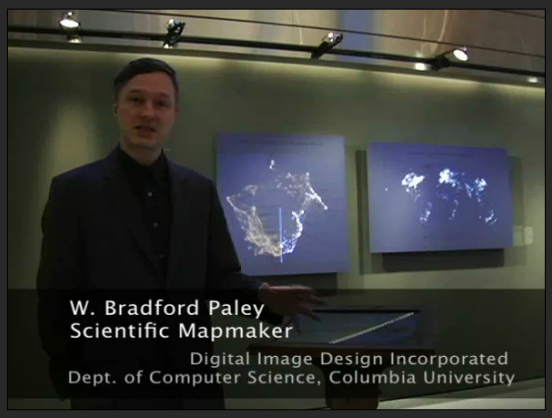


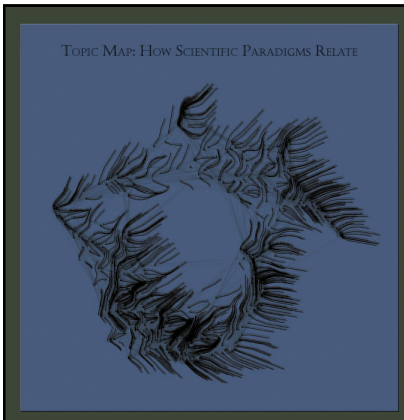


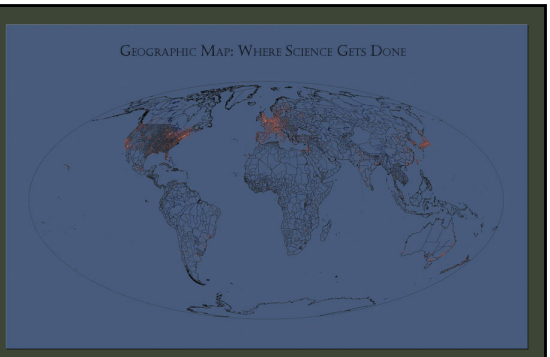
‘Places & Spaces: Mapping Science’ Exhibit

http://scimaps.org/exhibit

Illuminate Diagram Display







You may run your finger over each of these maps to control the lighting on the other: touching a place on the world map will light up topics studied in that place; touching a paradigm on the topic map will light up the places that study that topic.

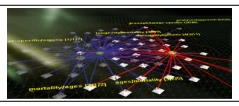
Nanotechnology

This overlay shows the distribution of nanotechnology within the paradigms of science. The majority of current work in nanotechnology takes places in physics, chemistry, and materials science, at the upper right portion of the map. However, an increasing amount of nanotechnology is being applied in the biological and medical sciences, at the lower right.

All Topics *Sweep through all 776 scientific paradigms*	**Nanotechnology** *Science on the tiny scale of molecules*	**Francis H. C. CRICK** *Co-discovered DNA's double helix*	**Albert EINSTEIN** *Revitalized physics with Relativity theories*	**Michael E. FISHER** *Models critical phase transitions of matter*	**Susan T. FISKE** *Connects perception and stereotypes*
Sustainability *The science behind our long-term hopes*	**Biology & Chemistry** *The interface between these two vital fields*	**Joshua LEDERBERG** *Pioneer in bacterial genetic mechanisms*	**Derek J. de Solla PRICE** *Known as the "Father of Scientometrics"*	**Richard N. ZARE** *Uses laser chemistry in molecular dynamics*	**About this display** *People & organizations that helped create it*

We sweep slowly through adjoining related topics, lighting up the places in the world that study each topic. You may select a subset of the topics that deal with these three interesting subjects by touching it.

A single person's spreading influence is shown as a series of four snapshots. First, we light only topics and places relating to that person's papers—papers that are still highly cited today. The second lights everything that cites that original work. Note that this first-generation impact extends to far more topics than did the original work. The third shapshot lights science that cites the second; and the fourth lights science that cites the third.



Example: Mapping Scientific Disciplines

First Analysis & Visualization

List of Documents

Term-Document Matrix Generator

Term-Document Matrix

Threshold Application

Highly frequent terms

Co-Occurrence Similarity Calculation

Similarity Matrix

Converter

Network in Pajek Format

Network Layout

Layout Optimization

Threshold Application

Reduced Network in Pajek format

Network Layout

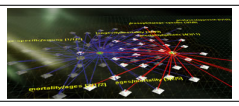
Pathfinder Network Scaling

Reduced Network in Pajek format

Network Layout

This Talk has Three Parts:

1. Why Develop CI in a InfoVis Lab / NetSci Center?
2. **Cyberinfrastructure Shell (CIShell) Framework**
3. Utilizing CIShell to Serve Different Communities
   - Information Visualization CI
   - Network Workbench



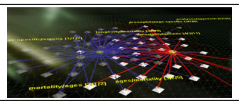
Why Do we Need Cyberinfrastructures?

Problem

- There are too many and too complex datasets that need to be correlated and understood to arrive at the best possible decisions.
- There are too many different data formats, different algorithms, different implementations of the same algorithm, different programming languages, different research purposes (modeling, analysis, visualization), different communities and practices.
- The analysis, modeling, and visualization of large datasets requires powerful computing infrastructures.
- Managing 1000+ of different data sets and 100+ of different algorithms requires a means to quickly select the best dataset(s)/algorithm(s).

Needed is a socio-technical cyberinfrastructure that supports

Easy access to datasets and algorithms, computer resources, their descriptions, and associated learning modules and access to expertise.



Cyberinfrastructure Shell (CIShell)

http://cishell.org

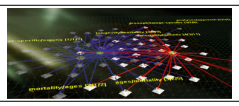
CIShell is an 'empty shell' that supports

- Easy integration of new datasets and algorithms by algorithm developers and
- Easy usage of algorithms by algorithm users.

Its plug-and-play architecture supports the integration and utilization of diverse

- Datasets, e.g., stored in files, databases, steaming data.
- Algorithms, e.g., data processing, analysis, modeling, visualization.
- Interfaces, e.g., remote services, scripting engines, peer-to-peer clients.
- Services, e.g., workflow support, scheduler.

Hence, it can be used for custom UI/Toolkit development.



CIShell – Technical Details

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

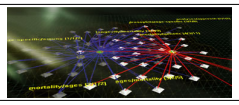
OSGi (http://www.osgi.org) is

- A standardized, component oriented, computing environment for networked services.
- Successfully used in the industry from high-end servers to embedded mobile devices since 7 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. Developed Tools/CI, e.g., IVC & NWB, provide a reference GUI for underlying services.

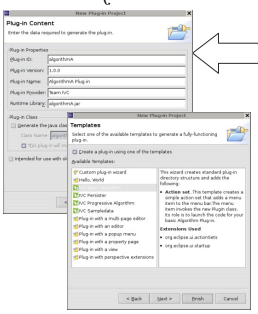


Serve Algorithms Developers & Users

Developers

Users

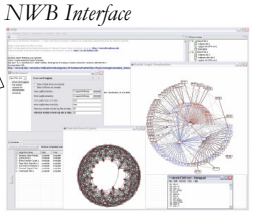
CIShell Wizards

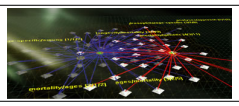


CIShell

IVC Interface

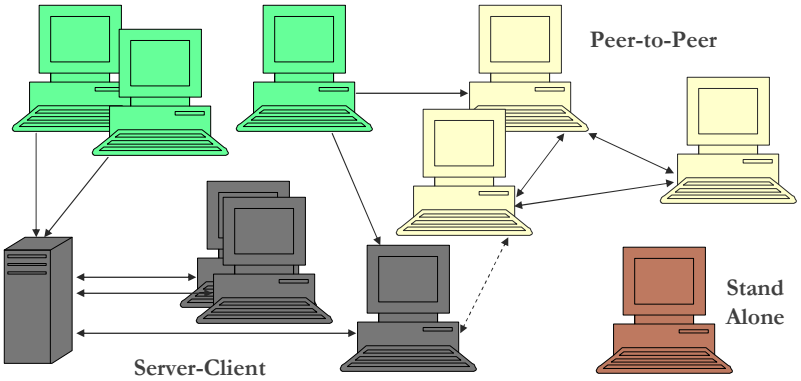
NWB Interface





Deployment

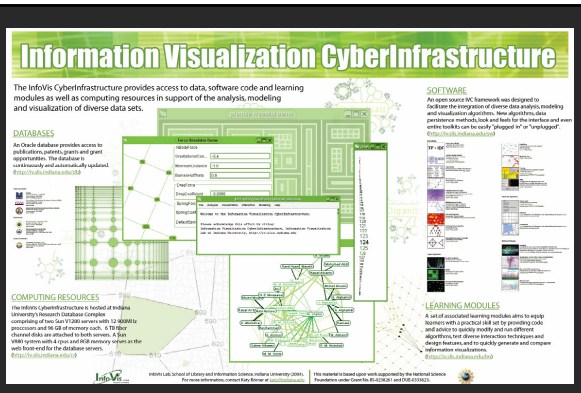
Data-Algorithm Repositories



CIShell applications can be deployed as distributed data and algorithm repositories, stand alone applications, peer-to-peer architectures, and server-client architectures.

This Talk has Three Parts:

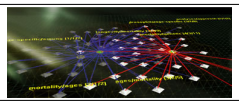
1. Why Develop CI in a InfoVis Lab / NetSci Center?
2. Cyberinfrastructure Shell (CIShell) Framework
3. **Utilizing CIShell to Serve Different Communities**
   - **Information Visualization CI**
   - Network Workbench



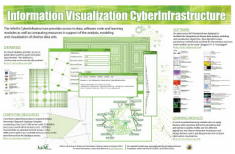
CAREER: Visualizing Knowledge Domains. NSF IIS-0238261 award (Katy Börner, $451,000) Sept. 03-Aug. 08.
http://iv.slis.indiana.edu/

NetworkWorkbench
A Workbench for Network Scientists

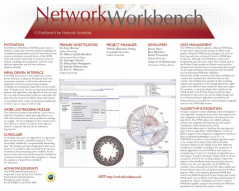
SEI: Network Workbench: A Large-Scale Network Analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research. NSF IIS-0513650 award (Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert (Senior Personnel), $1,120,926) Sept. 05 - Aug. 08.
http://nwb.slis.indiana.edu



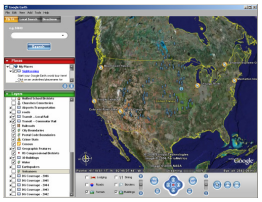
CIShell in Support of IV and NWB CIs & Tools



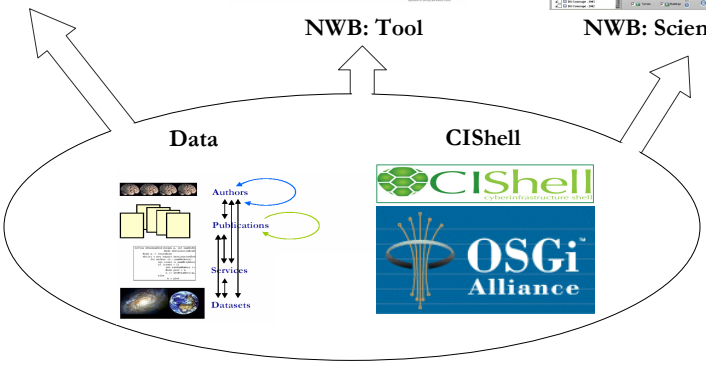
InfoVis CI

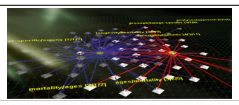


NWB: Tool



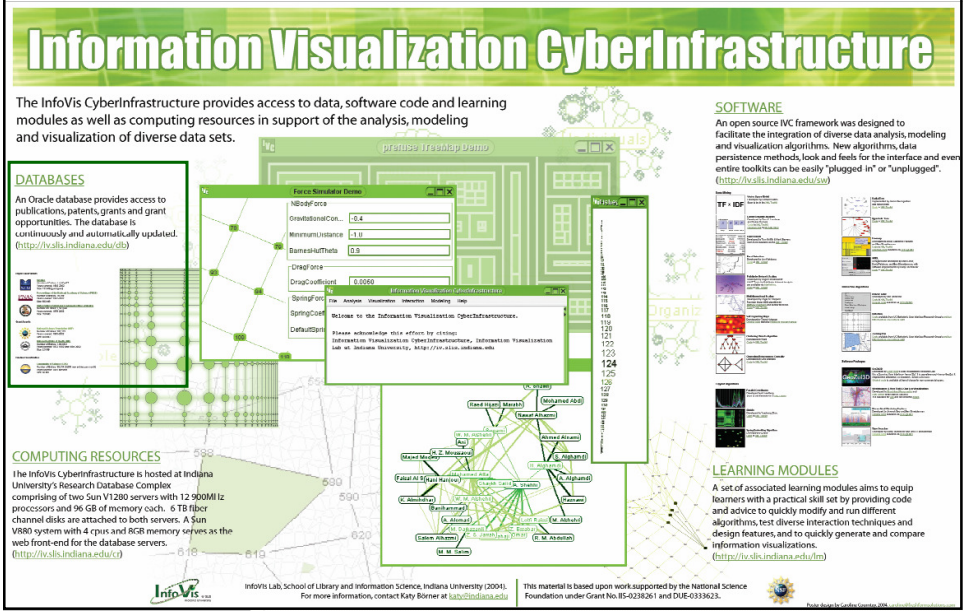
NWB: Science Map

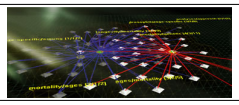




InfoVis Cyberinfrastructure

http://iv.slis.indiana.edu





IVC Database (http://iv.slis.indiana.edu/db)

Papers and Patents

Medline
Number of Entries: 11,693,477
Years covered: 1963-2002
Size: 135 MB (gunzipped)

Proceedings of the Natioanl Academy of Science (PNAS)
Number of Entries: 16,169
Years covered: 1997-2002
Size: 583 MB

United States Patent and Trademark Office (Patents)
Number of Entries: 2,582,647
Years covered: 1976-2003
Size: 350 MB

Grant Awards

National Science Foundation (NSF)
Number of Entries: 181,132
Years covered: 1985-2002
Size: 400 MB

National Institute of Health (NIH)
Number of Entries: 1,003,521
Years covered: 1972-1992 and 1994-2002
Size: 2.3 GB

Funding Opportunities

Community of Science (COS)
Number of Entries: 38,154 (5,000 new entries per month)
Years covered: 2001-present
Size: 60 MB

Information Visualization CyberInfrastructure

The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.

DATABASES

An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (http://iv.slis.indiana.edu/db)

SOFTWARE

An open source IVC framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and even entire toolkits can be easily "plugged in" or "unplugged". (http://iv.slis.indiana.edu/sw)

COMPUTING RESOURCES

The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex comprising of two Sun V1280 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V880 system with 4 cpus and 8GB memory serves as the web front-end for the database servers. (http://iv.slis.indiana.edu/cr)

LEARNING MODULES

A set of associated learning modules aims to equip learners with a practical skill set by providing code and advice to quickly modify and run different algorithms, test diverse interaction techniques and design features, and to quickly generate and compare information visualizations. (http://iv.slis.indiana.edu/lm)

InfoVis Lab, School of Library and Information Science, Indiana University (2004). For more information, contact Katy Börner at katy@indiana.edu

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0238261 and DUE-0333623.

Katy Börner, Workshop on User Friendly Cyberinfrastructures 21

Information Visualization CyberInfrastructure

The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.

DATABASES

An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (http://iv.slis.indiana.edu/db)

SOFTWARE

An open source IVC framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and even entire toolkits can be easily "plugged in" or "unplugged". (http://iv.slis.indiana.edu/sw)

COMPUTING RESOURCES

The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex comprising of two Sun V1280 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V880 system with 4 cpus and 8GB memory serves as the web front-end for the database servers. (http://iv.slis.indiana.edu/cr)

LEARNING MODULES

A set of associated learning modules aims to equip learners with a practical skill set by providing code and advice to quickly modify and run different algorithms, test diverse interaction techniques and design features, and to quickly generate and compare information visualizations. (http://iv.slis.indiana.edu/lm)

InfoVis Lab, School of Library and Information Science, Indiana University (2004). For more information, contact Katy Börner at katy@indiana.edu

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0238261 and DUE-0333623.

Katy Börner, Workshop on User Friendly Cyberinfrastructures 22

Information Visualization CyberInfrastructure

The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.

DATABASES

An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (http://iv.slis.indiana.edu/db)

SOFTWARE

An open source IVC framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and even entire toolkits can be easily "plugged in" or "unplugged". (http://iv.slis.indiana.edu/sw)

COMPUTING RESOURCES

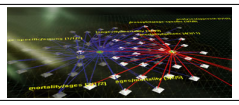
The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex comprising of two Sun V1280 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V880 system with 4 cpus and 8GB memory serves as the web front-end for the database servers. (http://iv.slis.indiana.edu/cr)

LEARNING MODULES

A set of associated learning modules aims to equip learners with a practical skill set by providing code and advice to quickly modify and run different algorithms, test diverse interaction techniques and design features, and to quickly generate and compare information visualizations. (http://iv.slis.indiana.edu/lm)

InfoVis Lab, School of Library and Information Science, Indiana University (2004). For more information, contact Katy Börner at katy@indiana.edu

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0238261 and DUE-0333623.



Time Series Analysis

Learning Module

http://iv.slis.indiana.edu/lm/lm-time-series.html

InfoVis CyberInfrastructure

A Data-Code-Compute Resource for Research and Education in Information Visualization

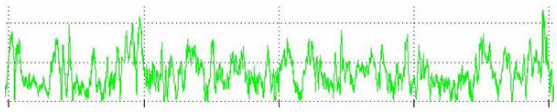
Home | Learning Modules | Software | Data Bases | Compute Resources | References

Learning Modules > Visualizing Time Series Data

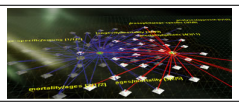
Description | Usage Hints | Learning Task | Discussion | References | Acknowledgments

Description

A time series is a sequence of events/observations which are ordered in one dimension, e.g., time. Frequently, successive observations depend on each other and it makes sence to display them in a (time) sorted fashion, e.g., as a scatter plot. Alternatively, one could be interested to know how many observations of a certain value have been made. Here one would sort the observations by value, count the number of observations for each value and derive a histogram. Time series data can be continuous, i.e., there is an observation at every instant of time see figure below, or discrete, i.e., observations exist for regularly or irregularly spaced intervals.



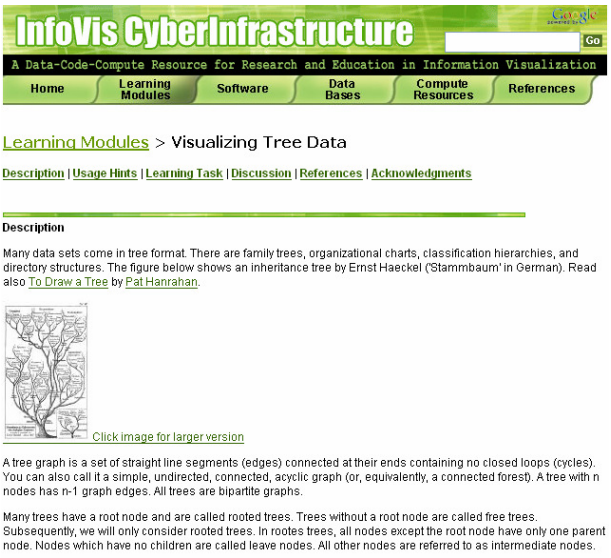
Time series are recorded, analized and used in diverse domains of science. Check out the Time Series Data Library maintained by Rob Hyndman and Muhammad Akram for numerous data sets from Agriculture, Chemistry, Crime, Demography, Ecology, Finance, Health, Hydrology, Industry, Labour market, Macro-Economics, Meteorology, Micro-Economics, Physics, Production, Sales, Simulated series, Sport, Transport & Tourism or Utilities.



Visualizing Tree Data

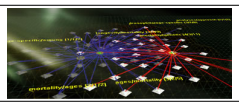
Learning Module

http://iv.slis.indiana.edu/lm/lm-trees.html



This Talk has Three Parts:

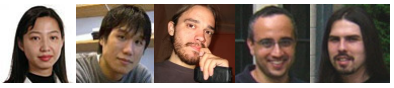
1. Why Develop CI in a InfoVis Lab / NetSci Center?
2. Cyberinfrastructure Shell (CIShell) Framework
3. **Utilizing CIShell to Serve Different Communities**
   - Information Visualization CI
   - **Network Workbench**



Network Workbench

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

Software Team: **Team Lead: Weixia (Bonnie) Huang**
Software Developers: Bruce Herr & Ben Markines
Algorithm Developers: Santo Fortunato & Cesar Hidalgo

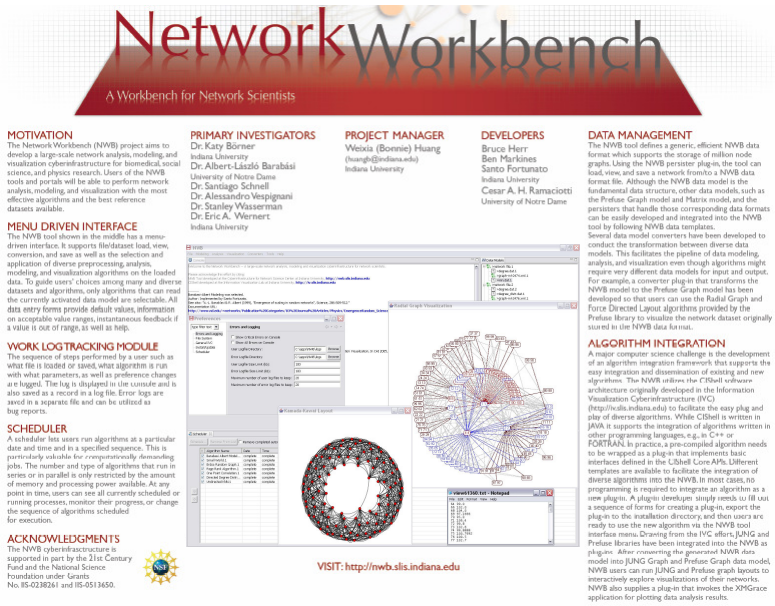


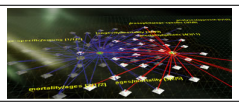
Goal: Develop a large-scale network analysis, modeling and visualization toolkit for biomedical, social science and physics research.

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

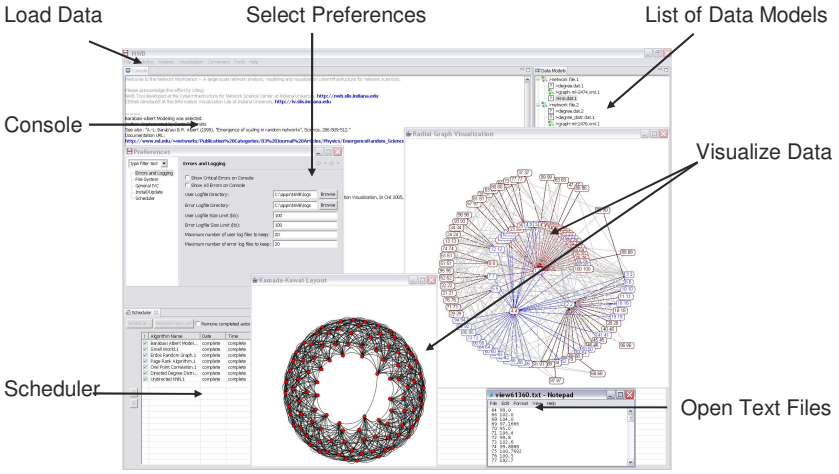
Duration: Sept. 2005 - Aug. 2008

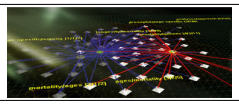
Website: http://nwb.slis.indiana.edu





NWB Tool: Interface Elements





List of Algorithms (partially implemented)

Modeling
Random Network Model
Random

Preferential Attachment Algorithms
Barabasi-Albert Model
Dorogovtsev-Mendes-Samukhin
Fitness
Vertices/edges deletion
Copying strategy
Finite vertex capacity
TARL

Rewiring algorithms
Rewiring based on degree distribution
Watts Strogatz Small World Model

Peer-to-Peer Models

Structured
CAN Model
Chord Model

Unstructured
PRU Model
Hypergrid Model

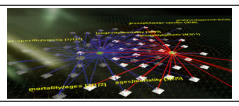
Measurement
Edge/Node level
node degree
BC value of nodes/edges
Max flow edge
Hub/Authority value for nodes
Distribution of node distances (Hop plot)

Local (directed and weighted versions)
Clustering Coefficient (Watts Strogatz)
Clustering Coefficient (Newman)
k-Core Count
Distributions (Plot and gamma, and R^2)
Degree Distributions (in, out, total) (Directed/Total Degree Distribution)
Degree Correlations (in-out, out-out, out-in, in-in, total-total)
Clustering Coefficient over k
Coherence for weighted graphs
Distribution of weights
Probability of degree distribution

Global
Density
Square of Adjacency Matrix
Giant Component
Strongly Connected Component
Betweenness Centrality
Diameter
Shortest Path = Geodesic Distance
Average Path Length

What other algorithms do you use?

Motif Identification
Page Rank
Closeness centrality
Reach centrality
Eigenvector centrality
Minimum Spanning Tree



List of Algorithms (partially implemented)

NetworkWorkbench
A Workbench for Network Scientists

Basic Processes on Networks
Search
k Random-Walk Search
Depth First Search
p-rand Breadth-First Search
P2P
CAN Search
Chord Search

Epidemics Spreading
SIR
SIS

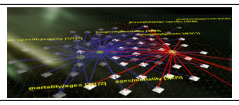
Graph Matching
Simple Match
Similarity Flooding
ABSURDIST

Clustering
Based on Attributes
Hierarchical Clustering
Single Link
Complete Link
Average Link
Ward's Algorithm

Based on Network Structure
Newman Girvan
Clauset-Newman-Moore
Newman
Cecconi-Parisi
Simulated annealing of modularity
Caldarelli
Weak Component Clustering
vanDongen (random walk)
Cfinder (Clique percolation method)
Reichardt, Bornholdt (q-potts model)

Visualization
Distribution
Scatterplot
Histogram
Geospatial
Circle layout
Grid-based
Dendrogram
Treemap
Hyperbolic tree
Radial Tree
Sparse Matrix Visualization
Kamada-Kawaii
Fruchterman-Rheingold
Orthogonal Layout
k-core visualization

What other algorithms do you use?

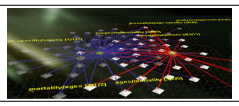


Download

- Download the IVC from http://sourceforge.net/projects/ivc
- Click 

- Download the ivc.1.0.4.full-installer.jar (with plugins).
- Install.
- Run.
- Enjoy.

Note:
If you have a Mac and the most recent JVM installed then you can download the empty IVC shell without plugins and get super old plugins from the update site at http://iv.slis.indiana.edu/ivc/update.
Or you wait for the new release that will be available on October 2nd, 2006.



References

- Bruce Herr, Weixia Huang, Shashikant Penumarthy, Katy Börner. Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. Submitted to William S. Bainbridge (Ed.) *Progress in Convergence*. Annals of the New York Academy of Sciences. http://ella.slis.indiana.edu/~katy/paper/06-cishell.pdf
- Börner, Katy, Sanyal, Soma and Vespignani, Alessandro. Network Science: A Theoretical and Practical Framework. (in press) In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Volume 41, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology. https://ivl.slis.indiana.edu/publications/borner-2006-netsci.pdf
- See also publications linked from Network Workbench web site at http://nwb.slis.indiana.edu/

For further information please contact:

Weixia (Bonnie) Huang, NWB Software Team Lead, Indiana University, huangb@indiana.edu

Bruce Herr, CIShell Architect, Indiana University, bherr@indiana.edu