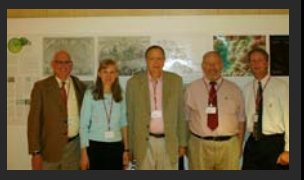
Scholarly Data, Network Science, and (Google) Maps – Modeling, Mapping, and Exploring the Dynamic Landscape of Science

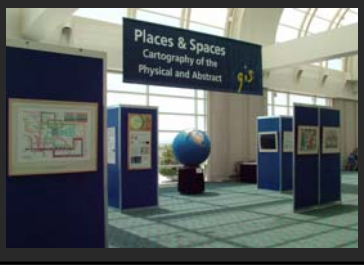
Dr. Katy Börner
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
katy@indiana.edu

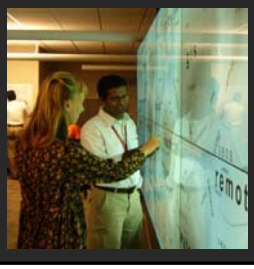
Talk at Borders in Bloomington, IN
2007.05.10



Places & Spaces: Mapping Science
a science exhibit that introduces people to maps of sciences, their makers and users.
Scimaps.org

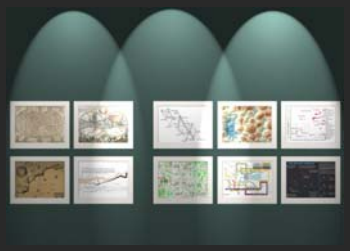
Exhibit Curators:
Dr. Katy Börner &
Julie Smith





The Power of Maps

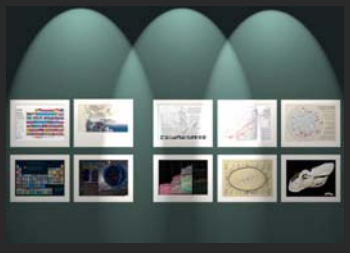
Four Early Maps of Our World
VERSUS
Six Early Maps of Science



(1st Iteration of Places & Spaces Exhibit - 2005)

The Power of Reference Systems

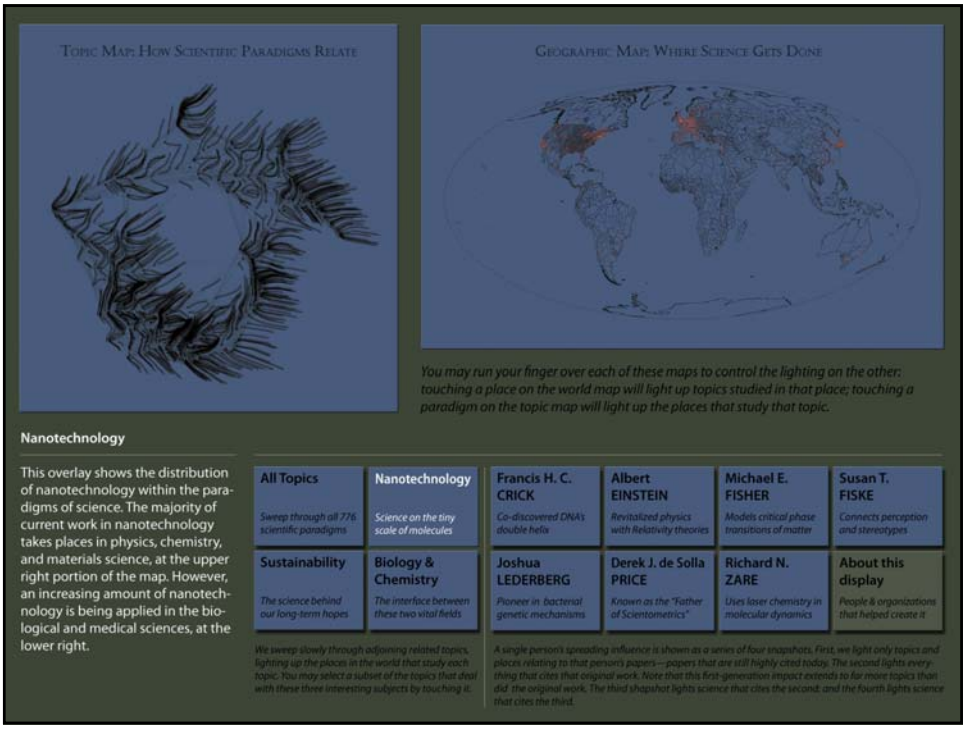
Four Existing Reference Systems
VERSUS
Six Potential Reference Systems of Science

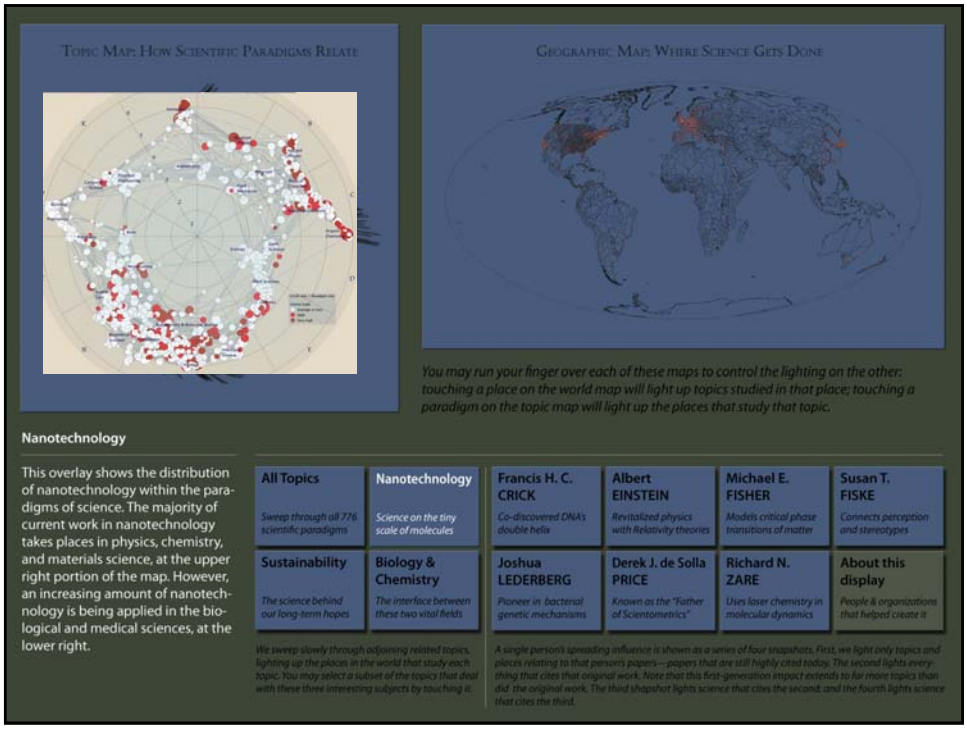


(2nd Iteration of Places & Spaces Exhibit - 2006)

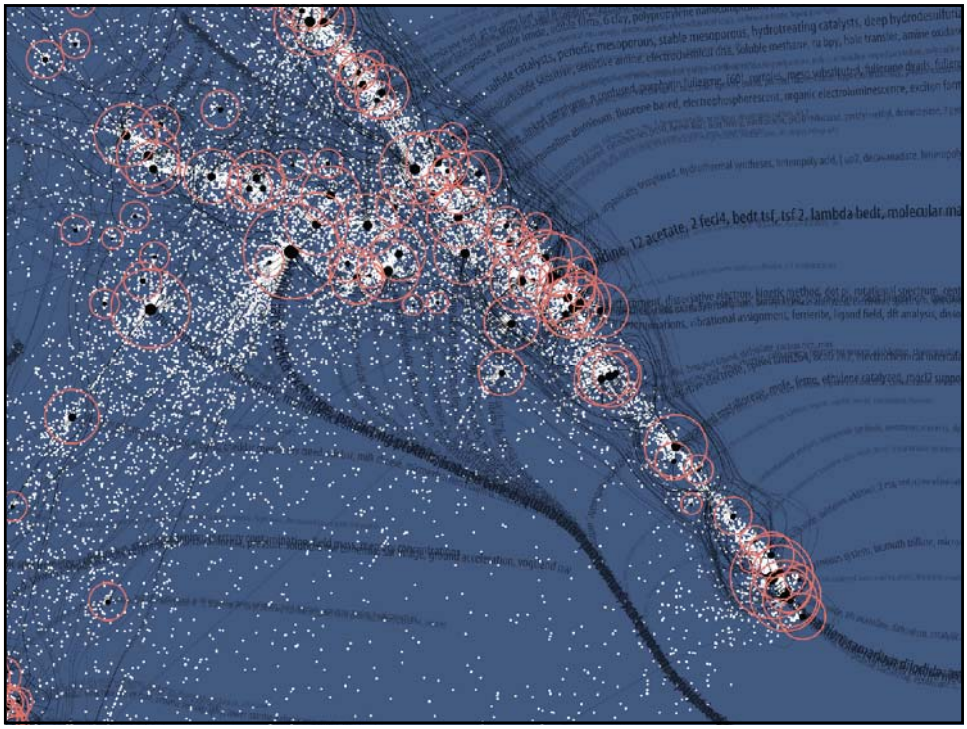
"Places & Spaces: Mapping Science"
on display at the New York Hall of Science
(5 mins from LaGuardia Airport)

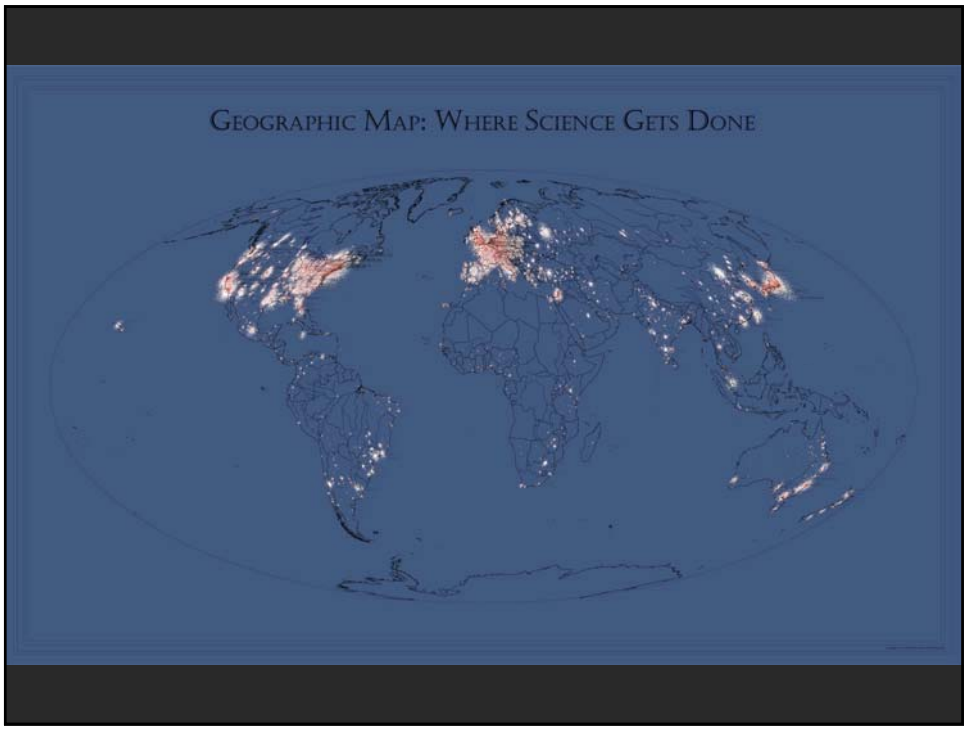
December 9th, 2006 – February 25th, 2007.

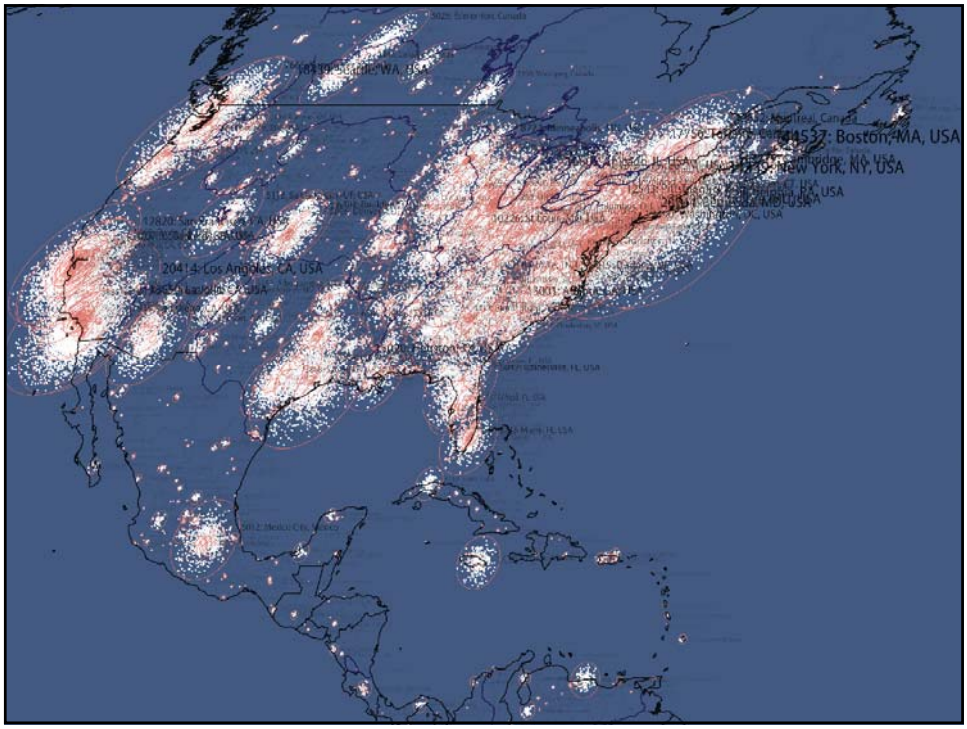


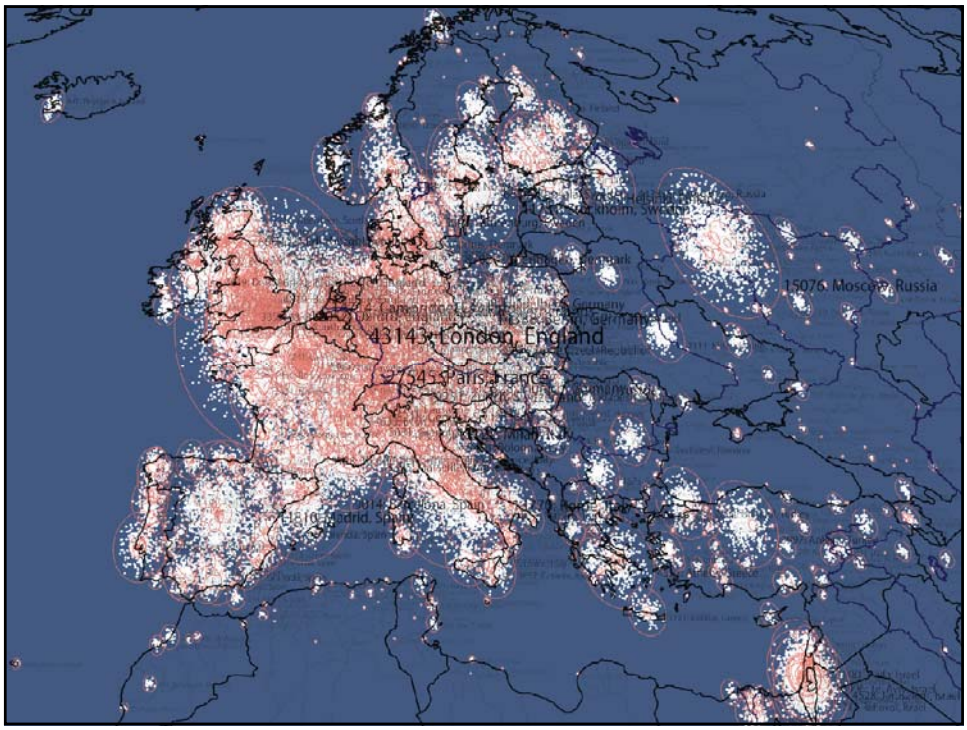

Topic Map: How Scientific Paradigms Relate
Geographic Map: Where Science Gets Done
You may run your finger over each of these maps to control the lighting on the other: touching a place on the world map will light up topics studied in that place; touching a paradigm on the topic map will light up the 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 snapshot lights science that cites the second; and the fourth lights science that cites the third.

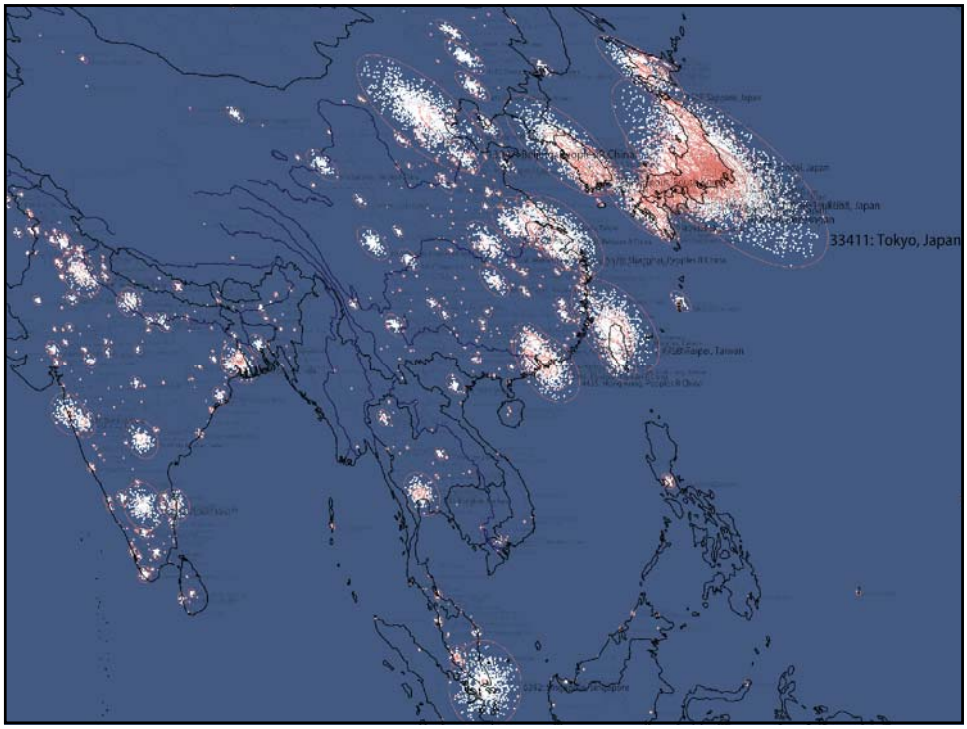
Topic Map: How Scientific Paradigms Relate




Geographic Map: Where Science Gets Done

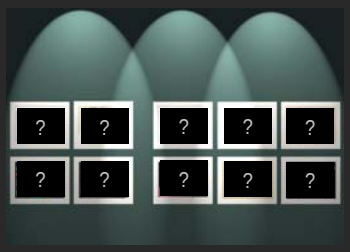

20414 Los Angeles, CA, USA
New York, NY, USA


15076 Moscow, Russia
43143 London, England
27543 Paris, France



The Power of Forecasts

Four Existing Forecasts
VERSUS
Six Potential Science 'Weather' Forecasts



(3rd Iteration of Places & Spaces Exhibit - 2007)

10 Iterations in 10 Years

part I:
2005 The power off maps
2006 The power off reference systems
2007 The power off forecasts
Here four relevant maps from any domain are compared with six powerful maps/reference systems/forecasts.

part II:
2008 Science maps for industry
2009 Science maps for science policy makers
2010 Science maps for researchers
2011 Science maps as visual interfaces to digital libraries
2012 Science maps for kids
2013 Science maps as daily science forecasts
Here four relevant maps from that specific domain, e.g., kids, are compared with six science maps for kids.

part III:
2014 How to lie with maps
Examples how to do a bad/misleading job.

Behind the Scene:

- **General Process**
- **People**
- **Scholarly Data Integration**
- **Network Science & (Google) Maps**

How to Make a Science Map

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

CLUSTER ANALYSIS

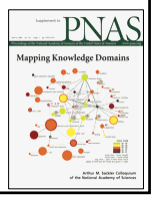
SCALAR
- Triangulation
- Force-directed placement (FDP)

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

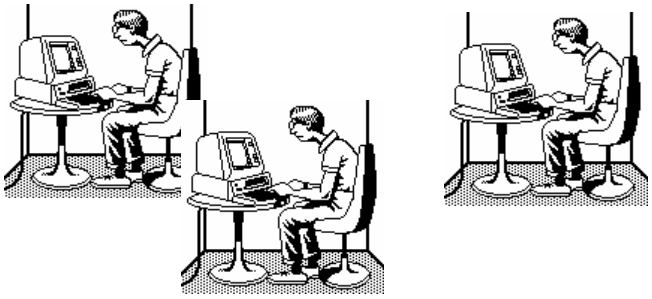
ANALYSIS

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).

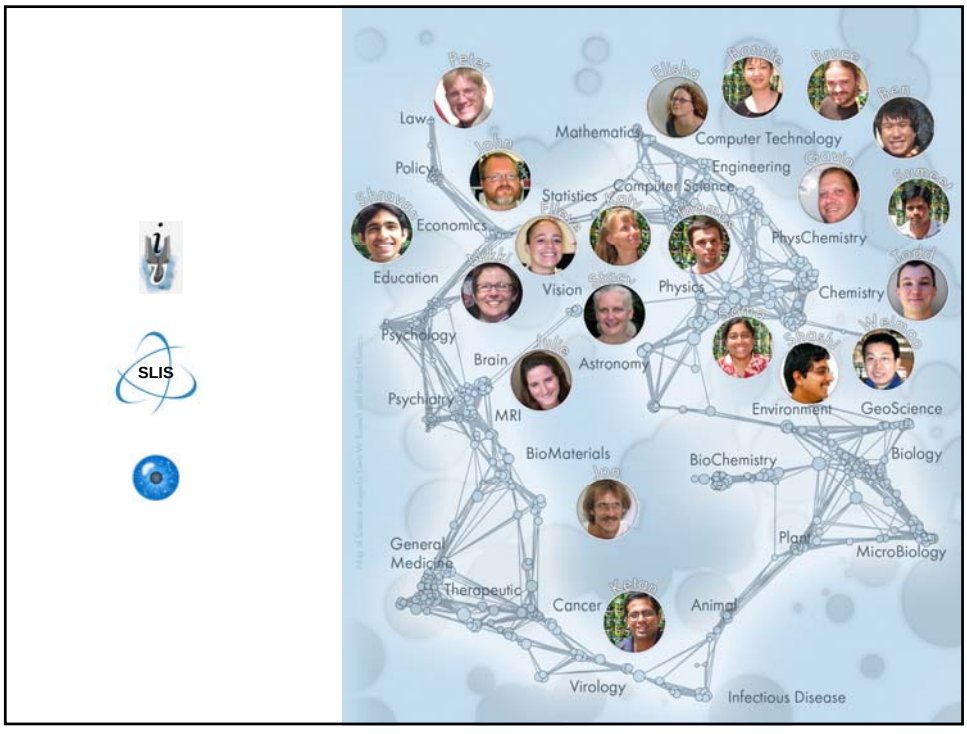
Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), Annual Review of Information Science & Technology, Volume 37, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 5, pp. 179-255.



The Cyberinfrastructure Developers



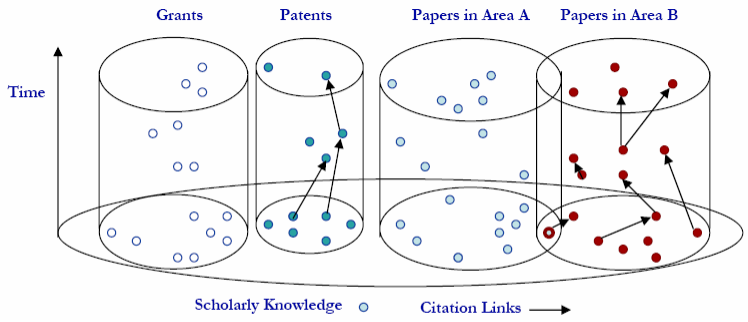
The Map Makers


SLIS
Peter
Elsbo
Ronnie
Bruce
Ben
Law
Mathematics
Computer Technology
Policy
Engineering
Statistics
Computer Science
Economics
PhysChemistry
Education
Vision
Physics
Chemistry
Psychology
Brain
Astronomy
Psychiatry
MRI
Environment
GeoScience
BioMaterials
BioChemistry
Biology
General Medicine
Plant
MicroBiology
Therapeutic
Cancer
Animal
Virology
Infectious Disease

Behind the Scene:

- People
- **Scholarly Data Integration**
- Network Science & (Google) Maps

Challenges - Interlink $ Input & Publication/Patent Citation Output

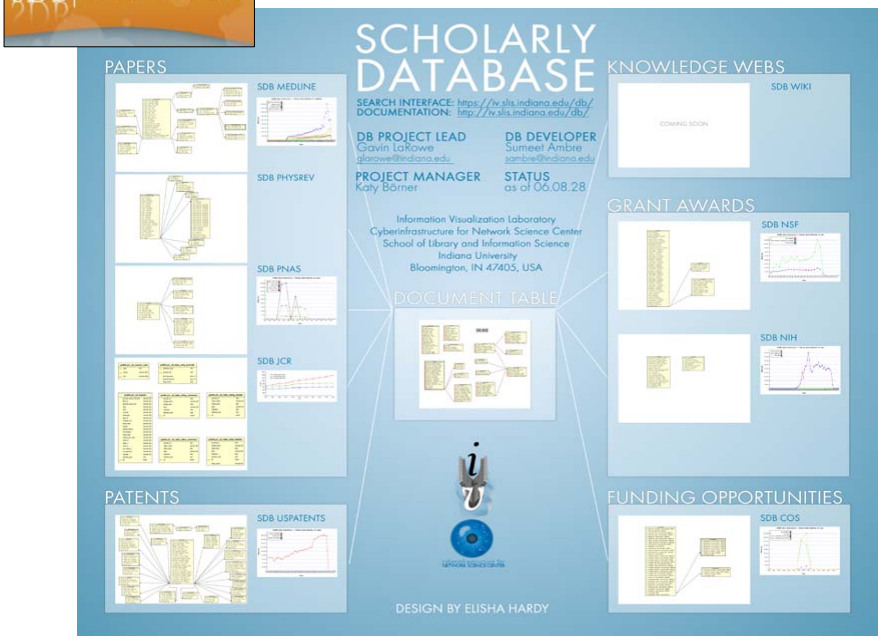


Need to interlink

- Grants and papers/patents.
- Grants/papers/patents and their PIs/authors/inventors, etc.

Use resulting networks to

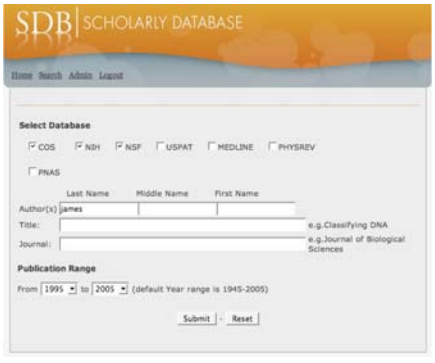
- Count #papers, #citations, etc.
- Determine strength of co-PI/author/inventor relations, etc.



Scholarly Database: Web Interface

Search across publications, patents, grants.

Download records and/or (evolving) co-author, paper-citation networks.

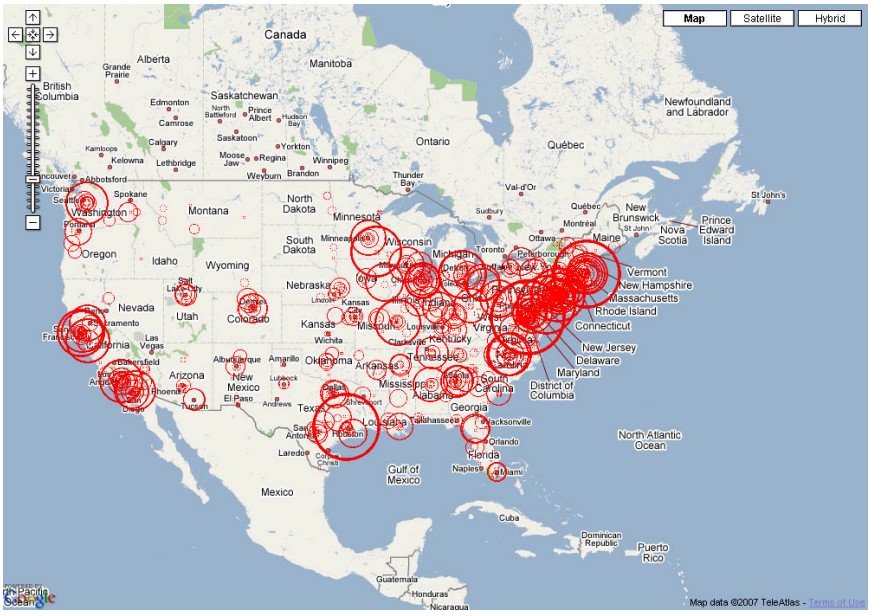


Scholarly Database: # Records & Years Covered

Datasets available via the Scholarly Database (* future feature)

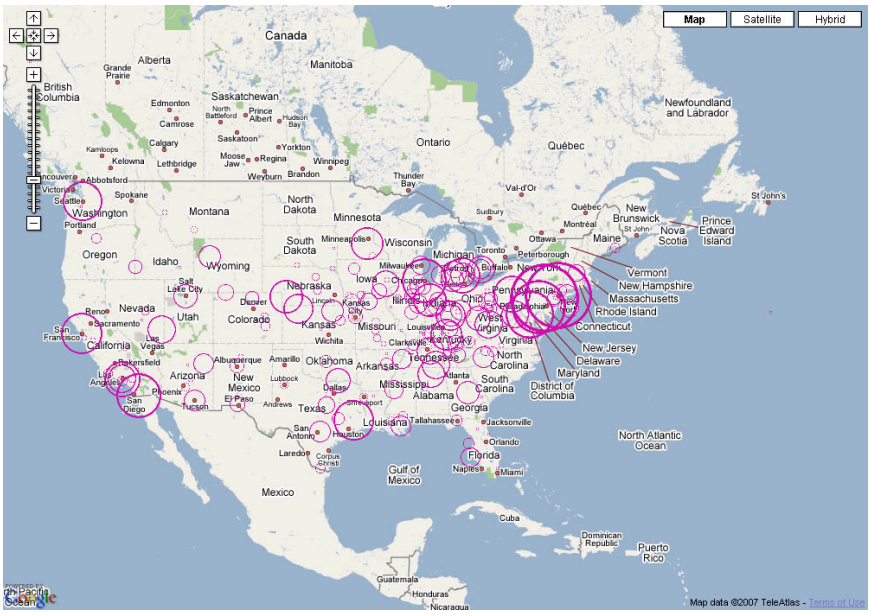
Dataset	# Records	Years Covered	Updated	Restricted Access
Medline	13,149,741	1965-2005	Yes	
PhysRev	398,005	1893-2006		Yes
PNAS	16,167	1997-2002		Yes
JCR	59,078	1974, 1979, 1984, 1989 1994-2004		Yes
USPTO	3,179,930	1976-2004	Yes*	
NSF	174,835	1985-2003	Yes*	
NIH	1,043,804	1972-2002	Yes*	
Total	**18,021,560**	**1893-2006**	**4**	**3**

Medline Publications



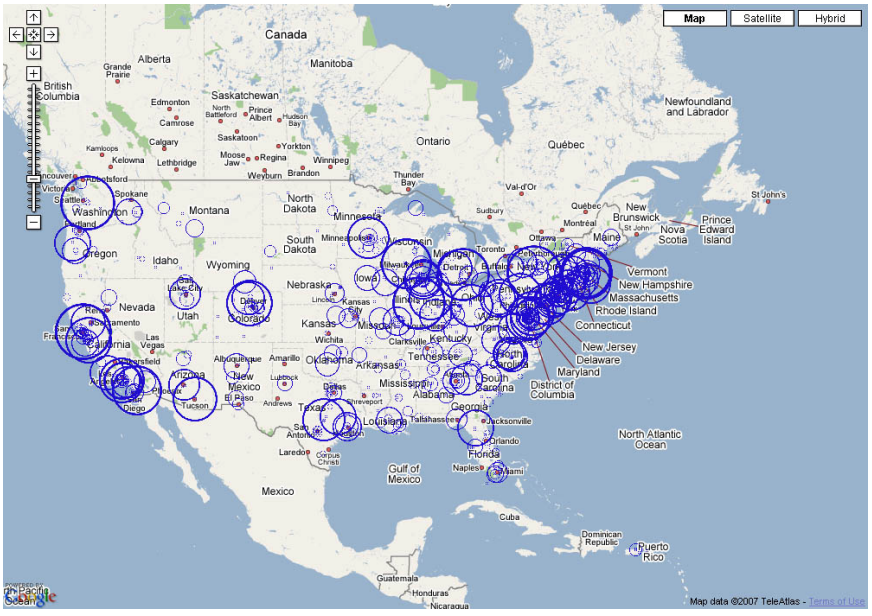
SDB | SCHOLARLY DATABASE

NIH



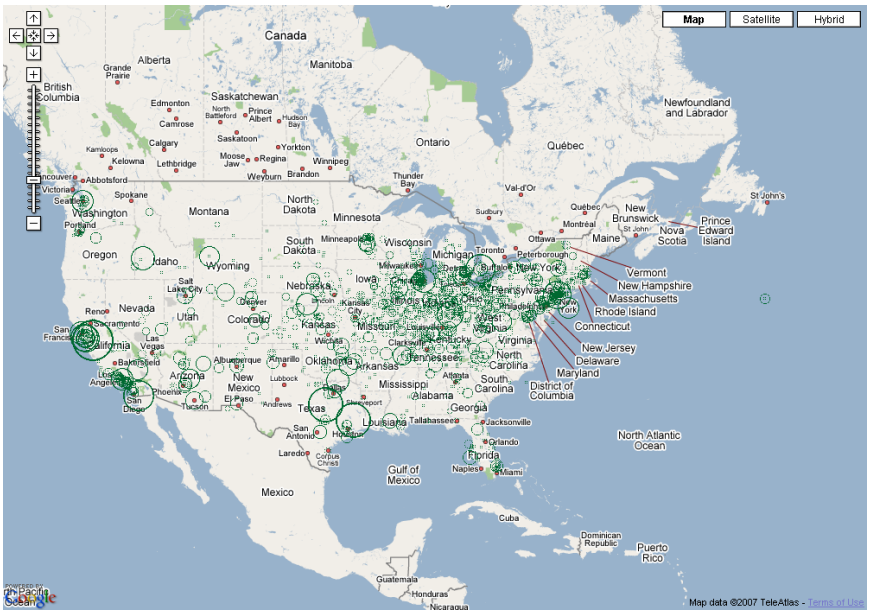
SDB | SCHOLARLY DATABASE

NSF Funding



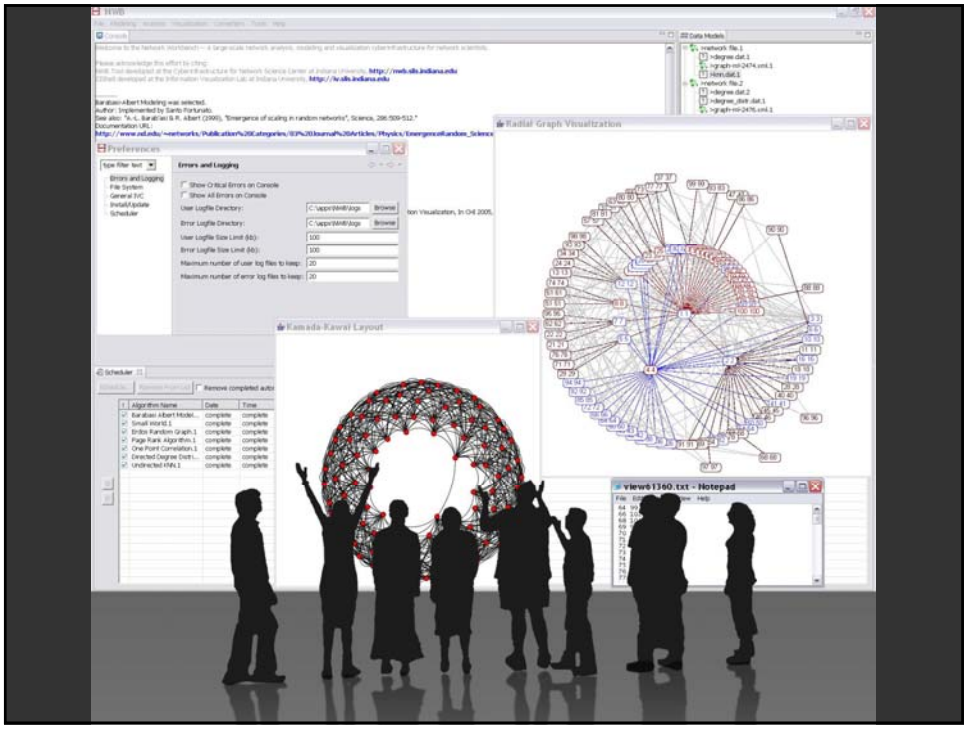
SDB | SCHOLARLY DATABASE

US Patents



Behind the Scene:

- People
- Scholarly Data Integration
- **Network Science & (Google) Maps**



Building Market Places not Cathedrals

- Requires the design & implementation of 'software glue' that can interlink datasets and algorithms written in different languages using different data formats.
- The smaller the glue or 'CI Shell', the more likely it can be maintained.
- Dataset and algorithm 'plugins' are provided by application holders/ community.
- Applications resemble custom 'fillings'.

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.

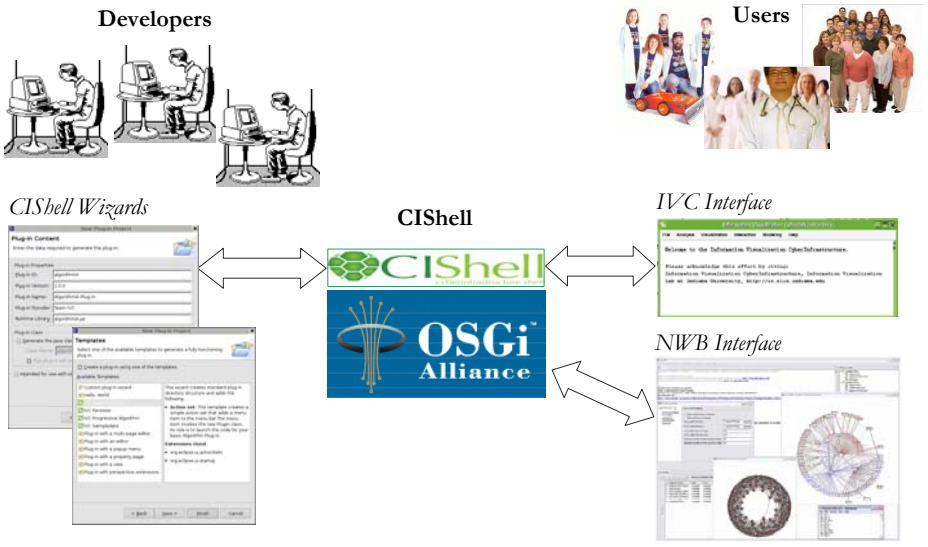
Its plug-and-play architecture supports the integration and utilization of diverse

- 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.

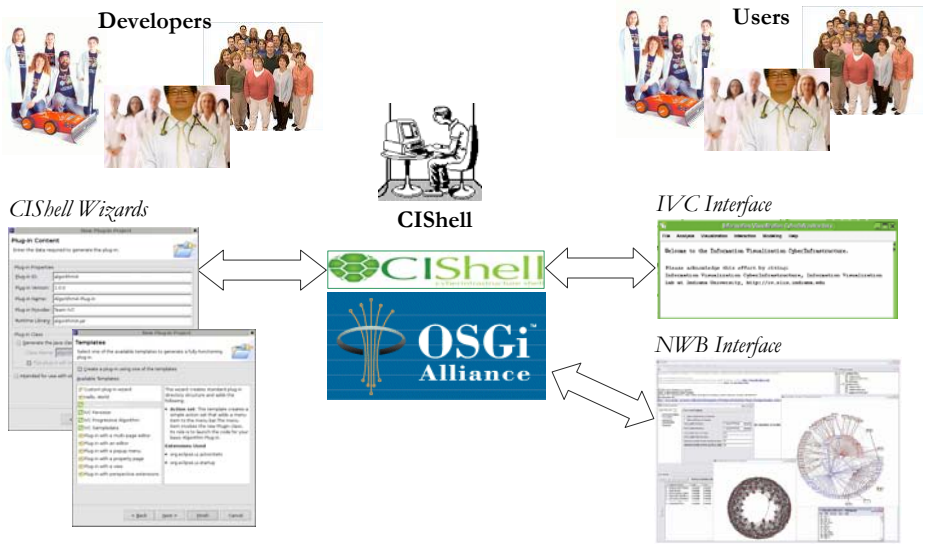
Hence, it can be used for custom UI/Toolkit development.

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 – Needs of Algorithm Developers & Users



CIShell – Needs of Algorithm Developers & Users



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.

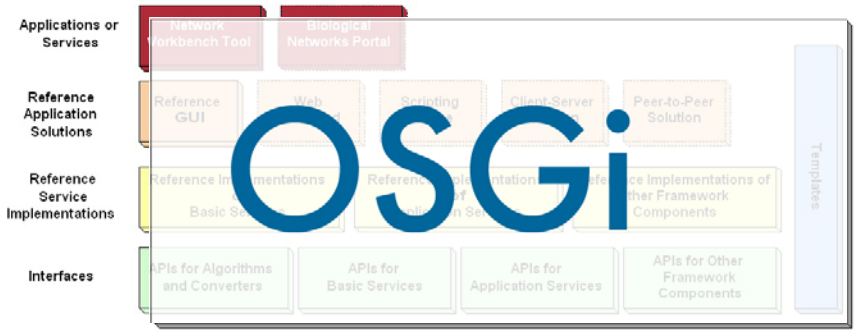
Advantages of Using OSGi

- 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.

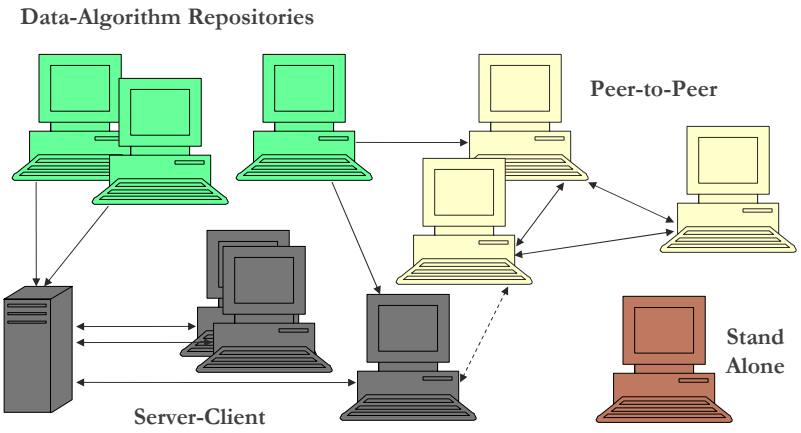
Ideally, CIShell becomes a standard for creating OSGi Services for algorithms. Developed Tools/CI, e.g., IVC & NWB, provide a reference GUI for underlying services.

CIShell – Technical Details

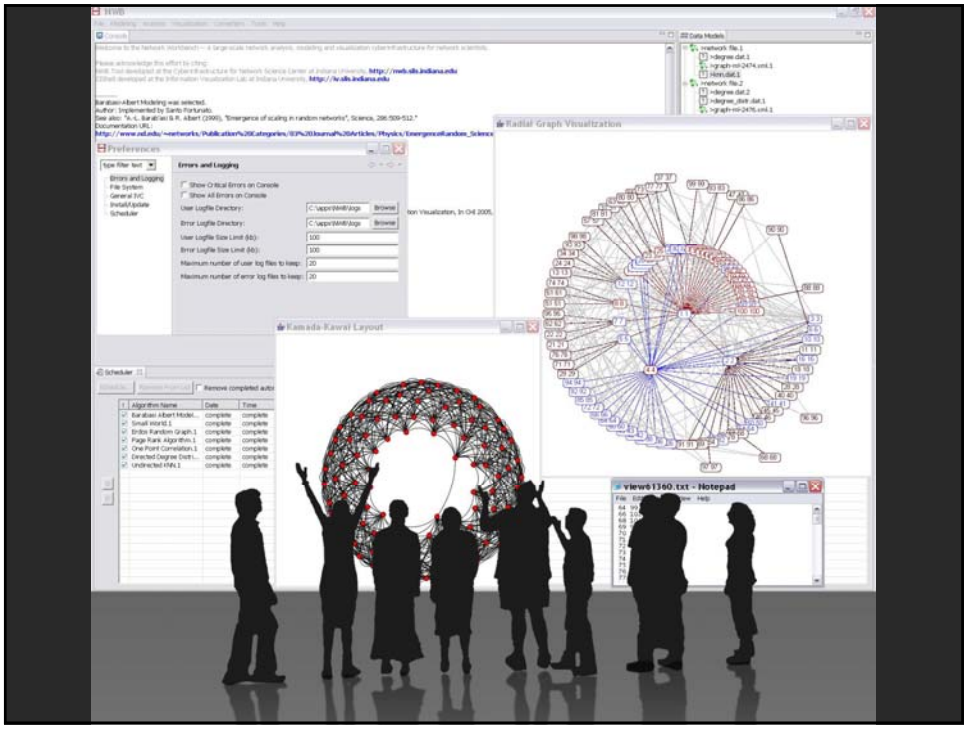
CIShell layer cake.



CIShell – Deployment



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



Network Workbench (NWB)

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

Software Team: Lead: Weixia (Bonnie) Huang

Developers: Bruce Herr, Ben Markines, Santo Fortunato, Cesar Hidalgo, Ramya Sabbineni, Vivek S. Thakre, & Russell Duhon

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 Advisory Board

- James Hendler (Semantic Web)
- Jason Leigh (CI)
- Neo Martinez (Biology)
- Michael Macy, Cornell University (Sociology)
- Ulrik Brandes (Graph Theory)
- Mark Gerstein, Yale University (Bioinformatics)
- Stephen North (AT&T)
- Tom Snijders, University of Groningen (Social Science)

NWB CI Deliverables

Glue:

- CIShell — Core programmer team lead by Bonnie Huang

Tools, Services & Portals:

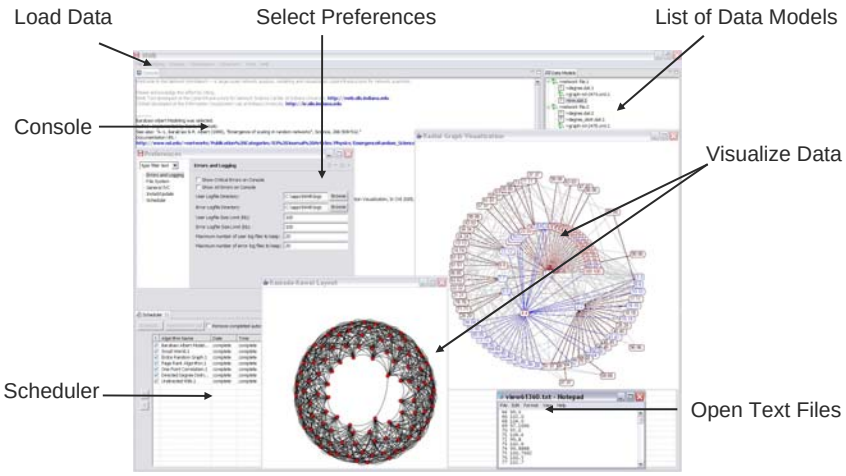
- NWB Tool — Lead by Alex Vespignani with input from other PIs
- SciMaps Service Online — Lead by Katy Borner
- Bio Portal — Lead by Laszlo Barabasi & Santiago Schnell

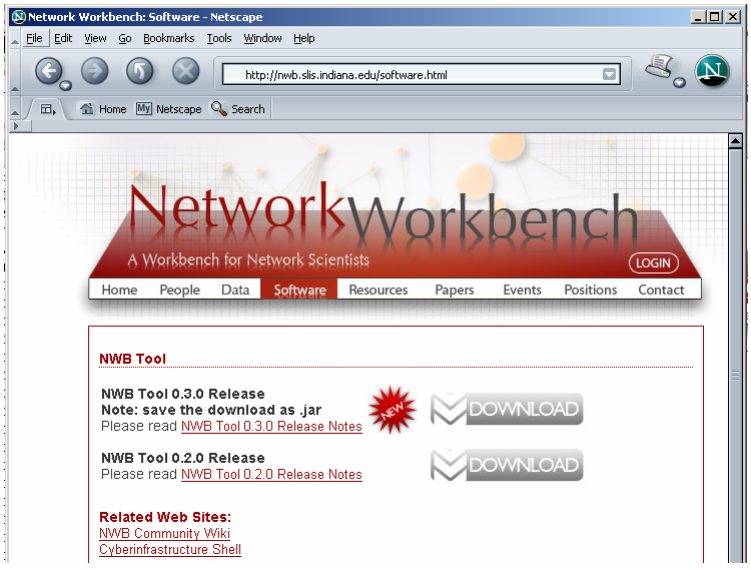
All three are prototypical instantiations of CIShell serving as reference implementations.

Documentation/Registry/Market Place:

- NWB Community Wiki — Lead by Katy Borner

NWB Tool: Interface Elements

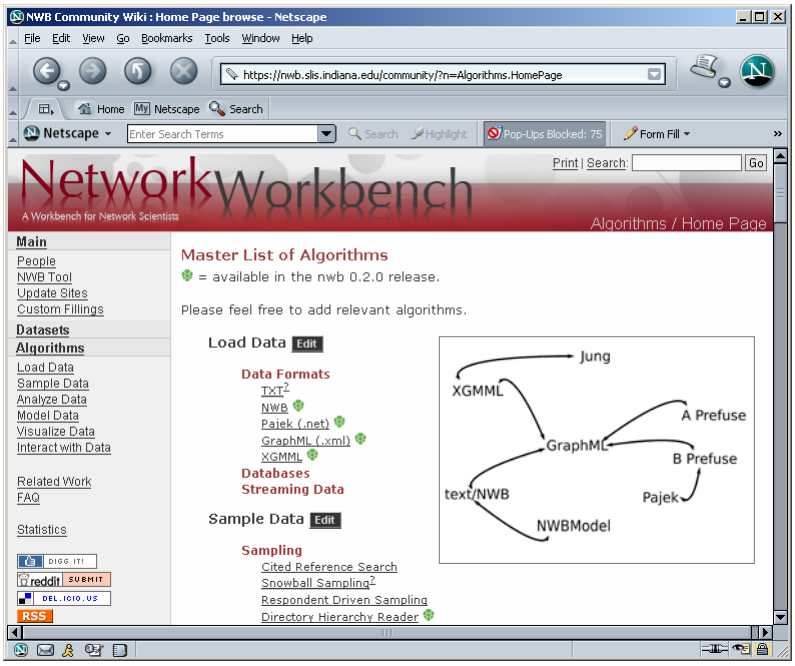




NWB Tool 0.2.0: List of Algorithms

Category	*Algorithm*	*Language*
Preprocessing	Directory Hierarchy Reader	JAVA
Modeling	Erdös-Rényi Random	FORTRAN
	Barabási-Albert Scale-Free	FORTRAN
	Watts-Strogatz Small World	FORTRAN
	Chord	JAVA
	CAN	JAVA
	Hypergrid	JAVA
	PRU	JAVA
Visualization	Tree Map	JAVA
	Tree Viz	JAVA
	Radial Tree / Graph	JAVA
	Kamada-Kawai	JAVA
	Force Directed	JAVA
	Spring	JAVA
	Fruchterman-Reingold	JAVA
	Circular	JAVA
	Parallel Coordinates (demo)	JAVA
Tool	XMGrace	

Analysis Algorithm	*Language*
Attack Tolerance	JAVA
Error Tolerance	JAVA
Betweenness Centrality	JAVA
Site Betweenness	FORTRAN
Average Shortest Path	FORTRAN
Connected Components	FORTRAN
Diameter	FORTRAN
Page Rank	FORTRAN
Shortest Path Distribution	FORTRAN
Watts-Strogatz Clustering Coefficient	FORTRAN
Watts-Strogatz Clustering Coefficient Versus Degree	FORTRAN
Directed k-Nearest Neighbor	FORTRAN
Undirected k-Nearest Neighbor	FORTRAN
Indegree Distribution	FORTRAN
Outdegree Distribution	FORTRAN
Node Indegree	FORTRAN
Node Outdegree	FORTRAN
One-point Degree Correlations	FORTRAN
Undirected Degree Distribution	FORTRAN
Node Degree	FORTRAN
k Random-Walk Search	JAVA
Random Breadth First Search	JAVA
CAN Search	JAVA

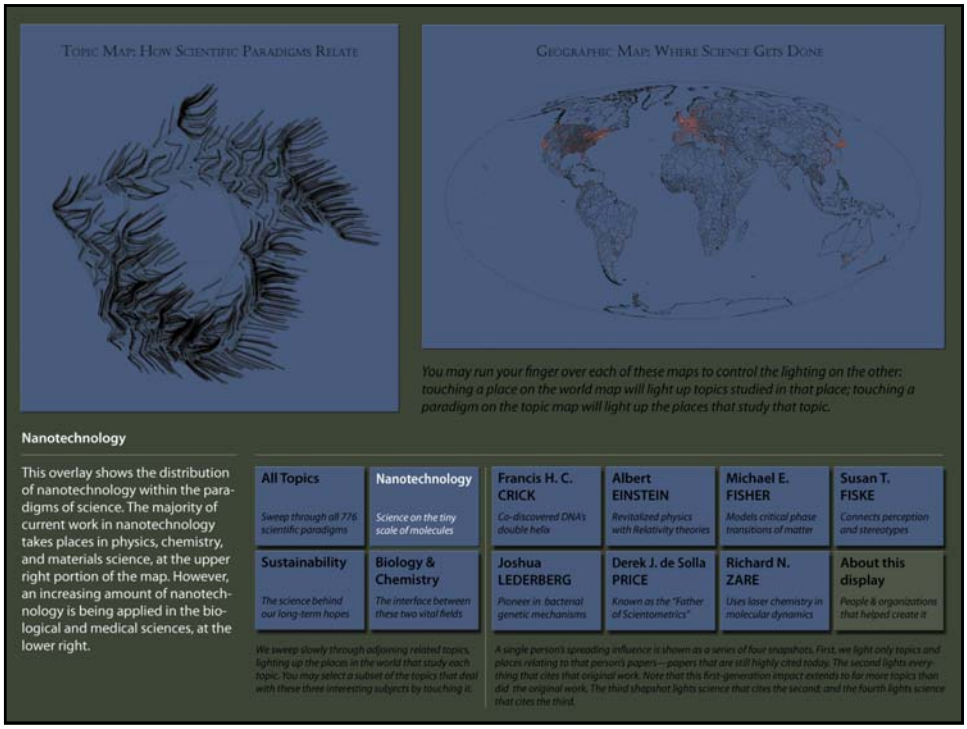


https://nwb.slis.indiana.edu/community

A Potential Future:
Science Maps in Action

KIDS first …

All maps of science are on sale
via http://scimaps.org



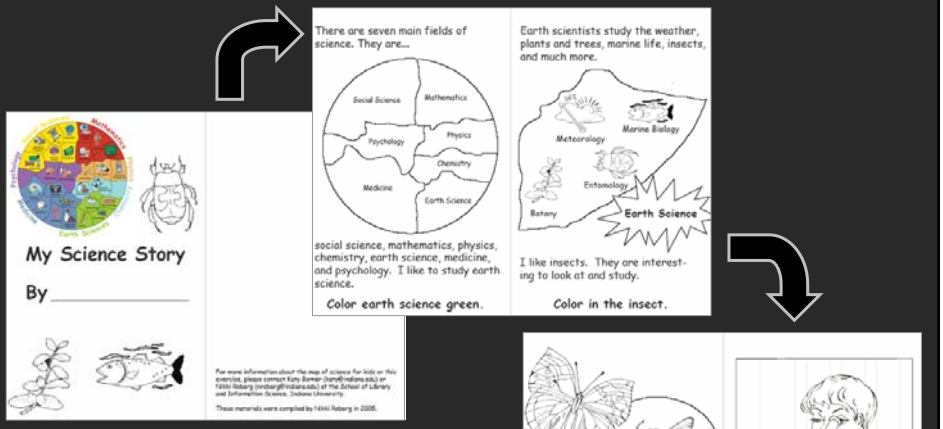
Science Puzzle Map for Kids by Fileve Palmer, Julie Smith, Elisha Hardy and Katy Börner, Indiana University, 2006. (Base map taken from Illuminated Diagram display by Kevin Boyack, Richard Klavans, and W. Bradford Paley.)

Inventors

Inventors & Inventions
ASTRONOMY
PHYSICS
ENGINEERING
COMPUTER SCIENCE
MATH
SOCIAL SCIENCE
PHYSICAL CHEMISTRY
BRAIN
MENTAL HEALTH
EARTH SCIENCES
ORGANIC THEORY
ECOLOGY
HEALTH CARE
BIOLOGY
THERAPEUTIC RESEARCH
INFECTIOUS DISEASE
CELL BIOLOGY
VIRUSES

Inventors
ASTRONOMY
PHYSICS
ENGINEERING
COMPUTER SCIENCE
MATH
SOCIAL SCIENCE
PHYSICAL CHEMISTRY
BRAIN
MENTAL HEALTH
EARTH SCIENCES
ORGANIC THEORY
ECOLOGY
HEALTH CARE
BIOLOGY
THERAPEUTIC RESEARCH
CELL BIOLOGY
INFECTIOUS DISEASE
VIRUSES

54
SOCCER



Activities:

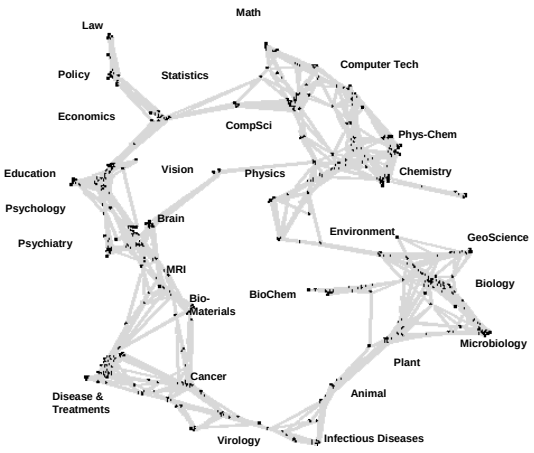
- Solve the puzzle.
- Navigate to 'Earth Science'.
- Identify major inventions.
- Place major inventors.
- Find your dream job on the map.
- Why is mathematics important?

… my SPONSORS next …

Latest 'Base Map' of Science

Kevin W. Boyack & Richard Klavans, unpublished work.

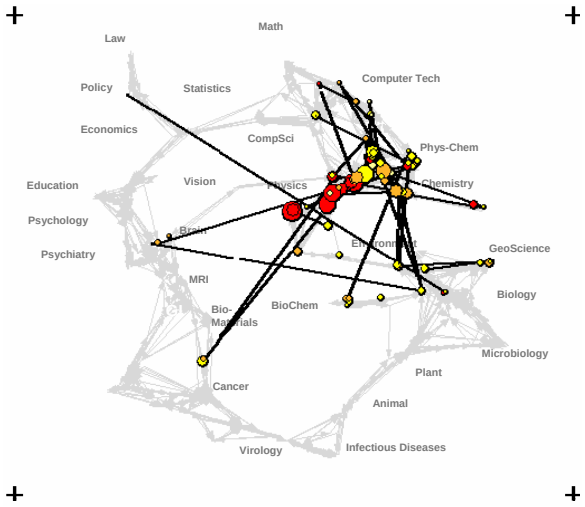
- Uses combined SCI/SSCI from 2002
  - 1.07M papers, 24.5M references, 7,300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts were reaggregated at the journal cluster level to calculate the
  - (x,y)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



Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

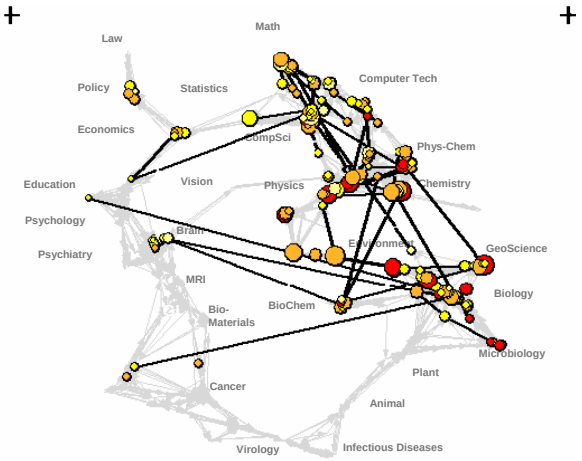
Funding patterns of the US Department of Energy (DOE)



Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

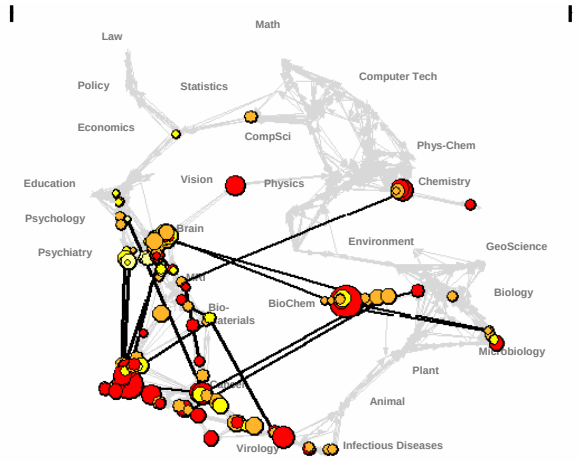
Funding Patterns of the National Science Foundation (NSF)



Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

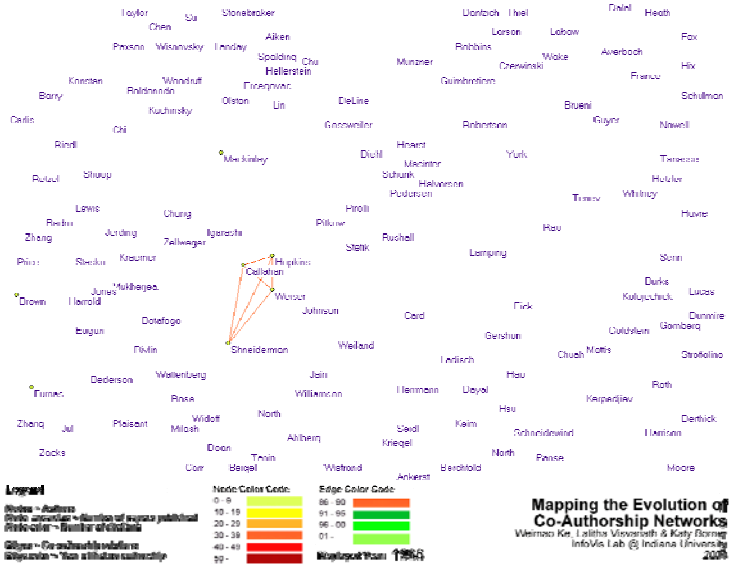
Funding Patterns of the National Institutes of Health (NIH)



… then SCIENTISTS …

Mapping the Evolution of Co-Authorship Networks

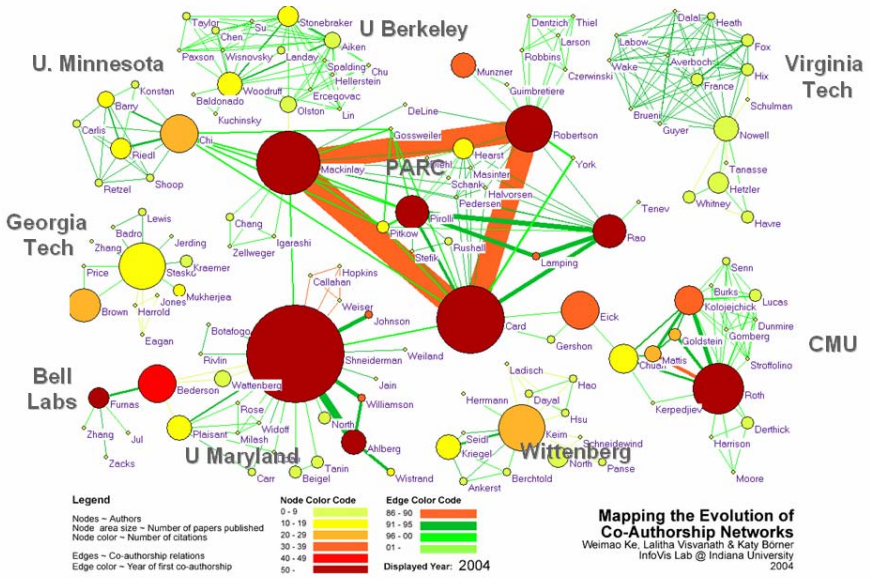
Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest.



62

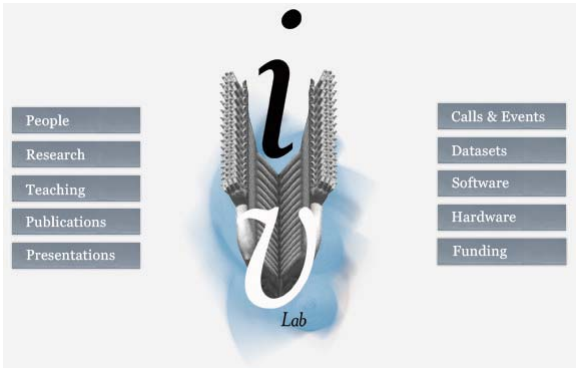
Mapping the Evolution of Co-Authorship Networks

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



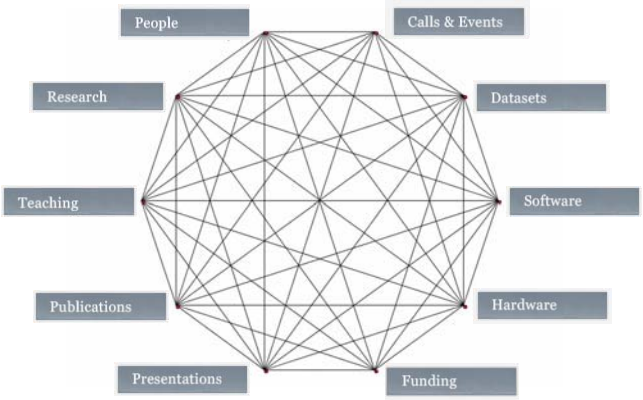
63

Lab/Center Management System

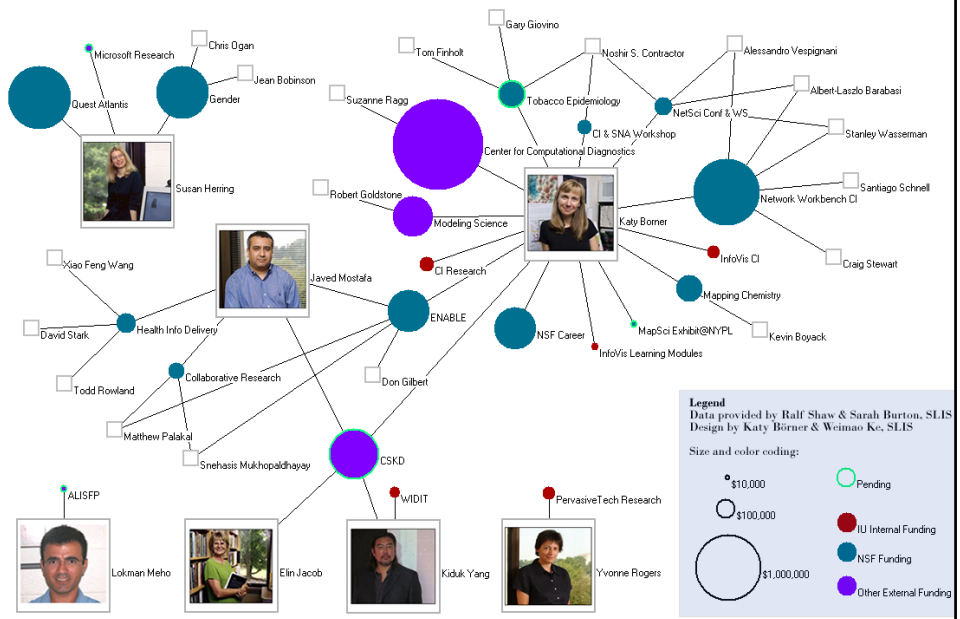


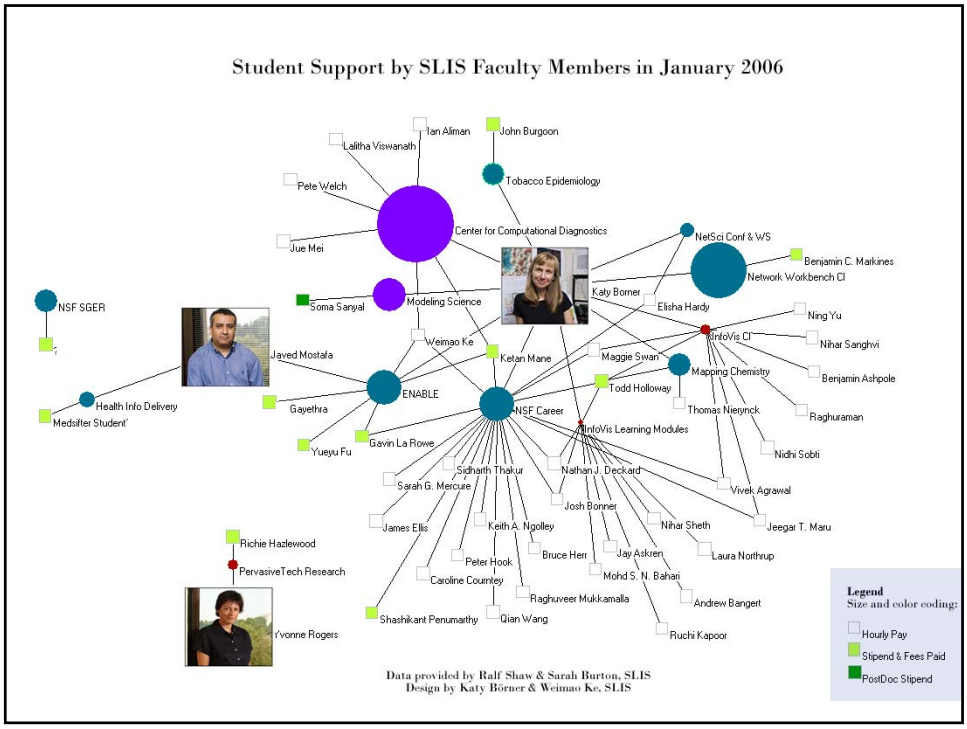
https://ivl.slis.indiana.edu

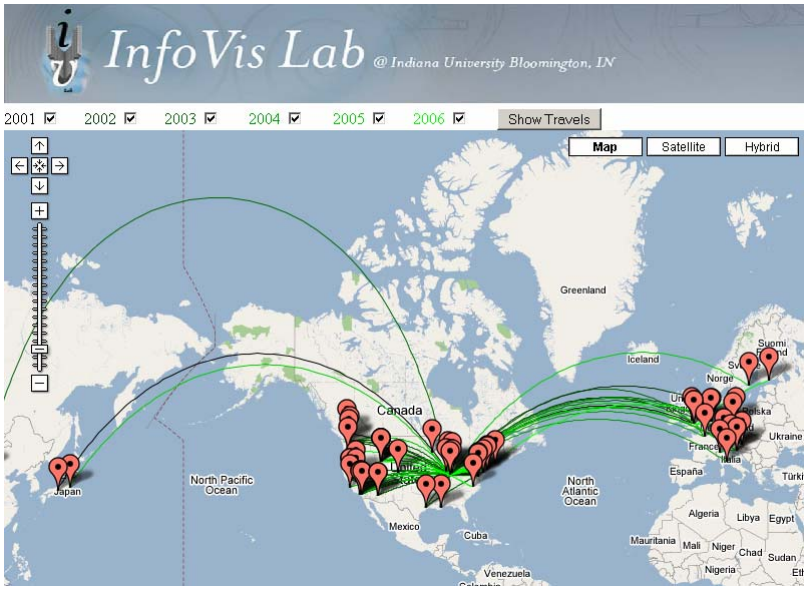
Data Entities and Interlinkages



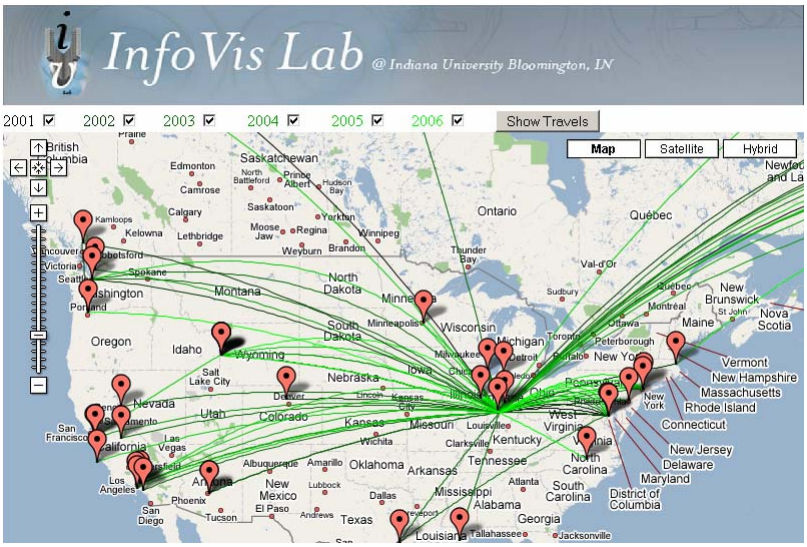
Awarded & Pending Funding of SLIS Faculty Members in January 2006



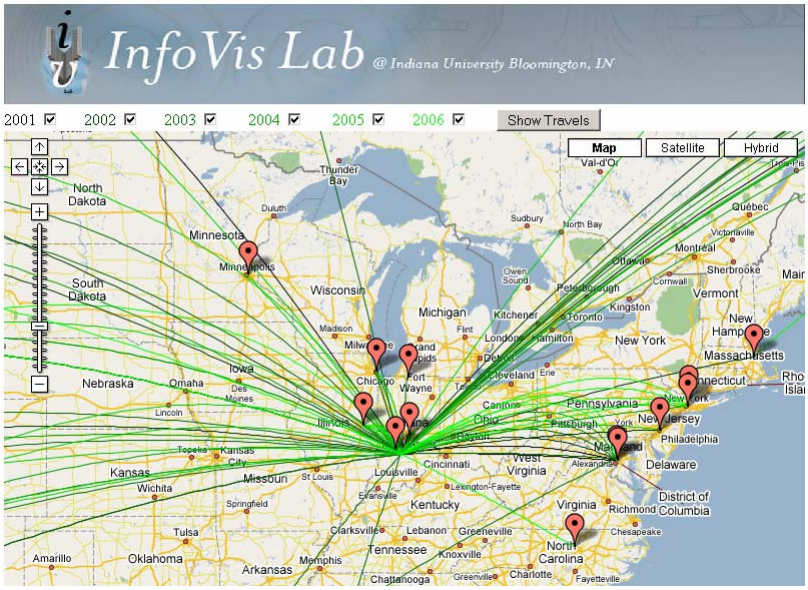




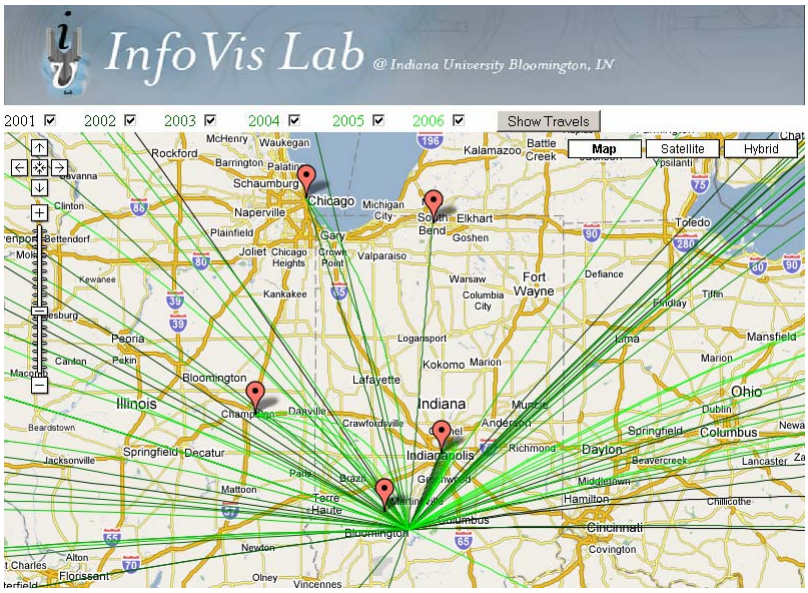
By Thomas Neirynck, 2007.



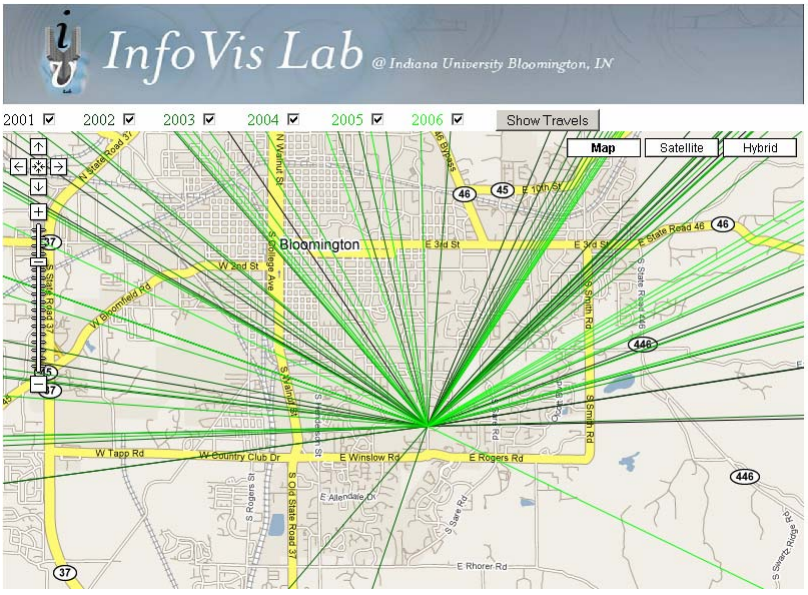
By Thomas Neirynck, 2007.



By Thomas Neirynck, 2007.

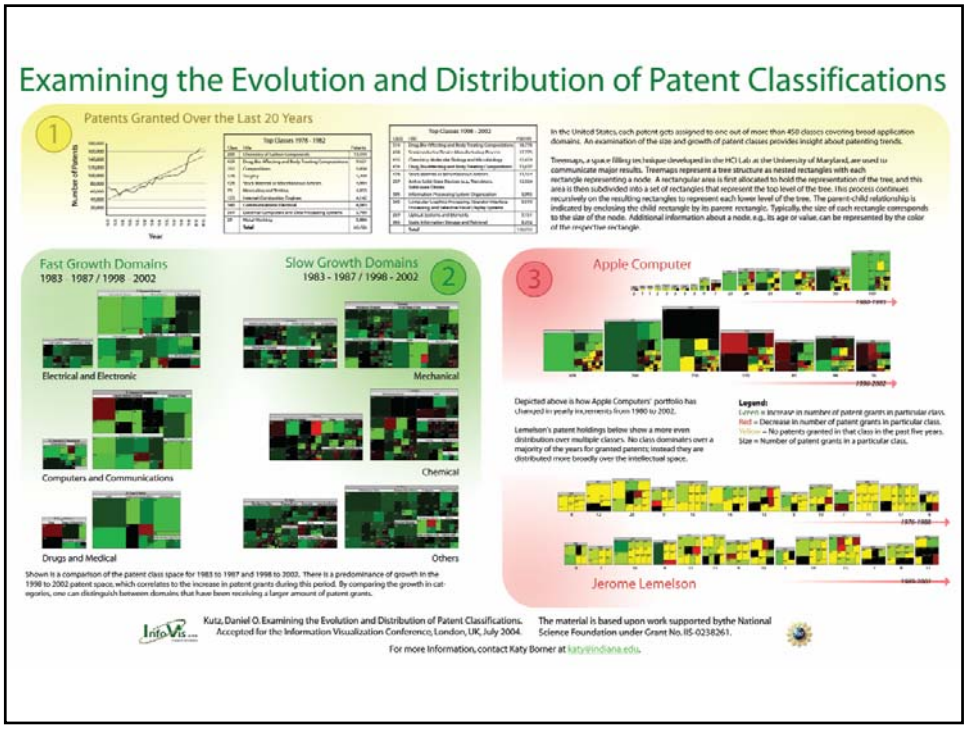


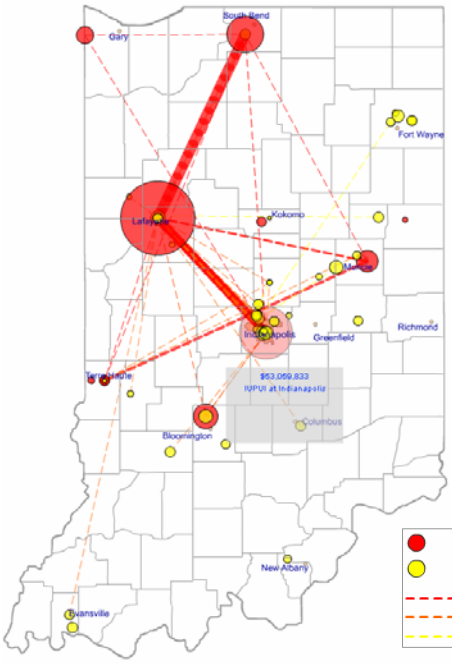
By Thomas Neirynck, 2007.



By Thomas Neirynck, 2007.

… and INDUSTRY too.





Mapping Indiana's Intellectual Space

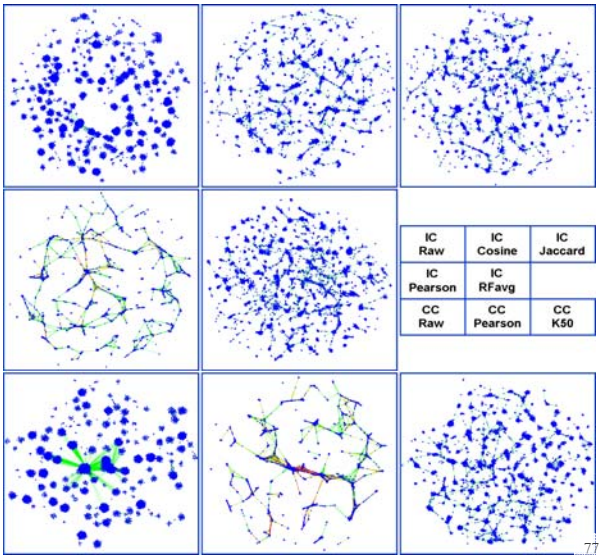
Identify

- Pockets of innovation
- Pathways from ideas to products
- Interplay of industry and academia

The End.

Comparison of Similarity Metrics

- ISI file year 2000, SCI and SSCI: 7,121 journals.
- Different similarity metrics
  - Inter-citation (raw counts, cosine, modified cosine, Jaccard, RF, Pearson)
  - Co-citation (raw counts, cosine, modified cosine, Pearson)
- Maps were compared based on
  - regional accuracy,
  - the scalability of the similarity algorithm, and
  - the readability of the layouts.

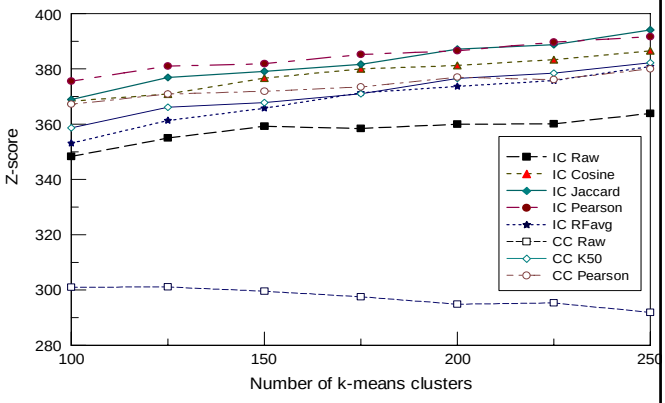


Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. Scientometrics. 64(3), 351-374.

77

Selecting the similarity measure with the best regional accuracy

- For each similarity measure, the VxOrd layout was subjected to k-means clustering using different numbers of clusters.
- Resulting cluster/category memberships were compared to actual category memberships using entropy/mutual information method by Gibbons & Roth, 2002.
- Increasing Z-score indicates increasing distance from a random solution.
- Most similarity measures are within several percent of each other.

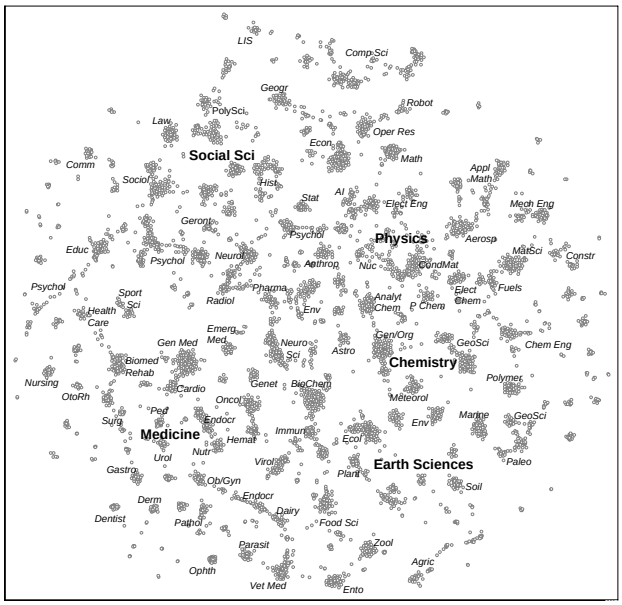


Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. Scientometrics. 64(3), 351-374.

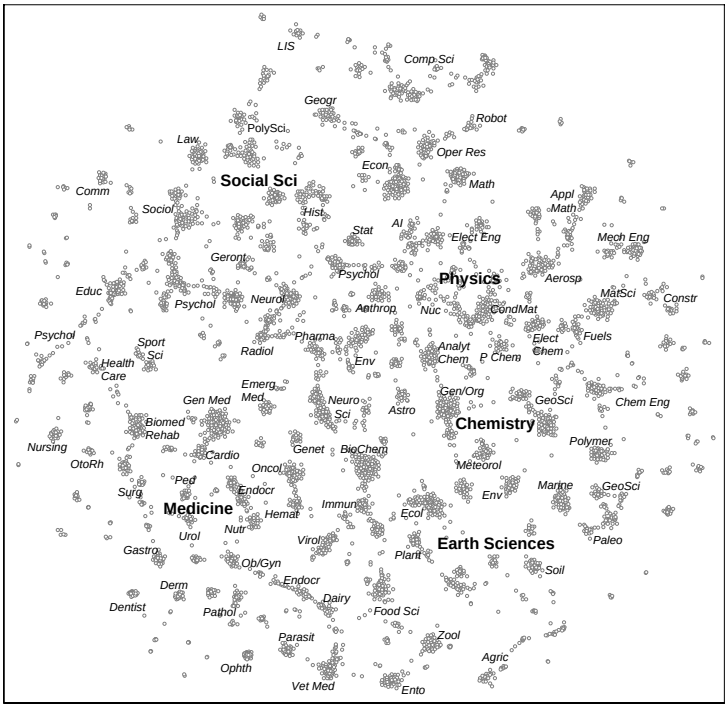
78

A 'Backbone' Map of Science & Social Science

- The map is comprised of 7,121 journals from year 2000.
- Each dot is one journal
- An *IC-Jaccard* similarity measure was used.
- Journals group by discipline.
- Groups are labeled by hand.
- Large font size labels identify major areas of science.
- Small labels denote the disciplinary topics of nearby large clusters of journals.



79

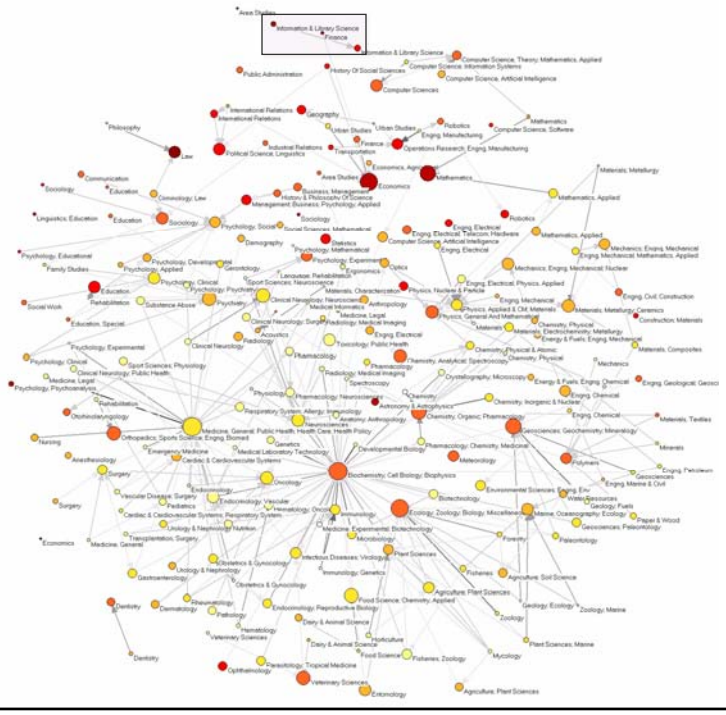


80

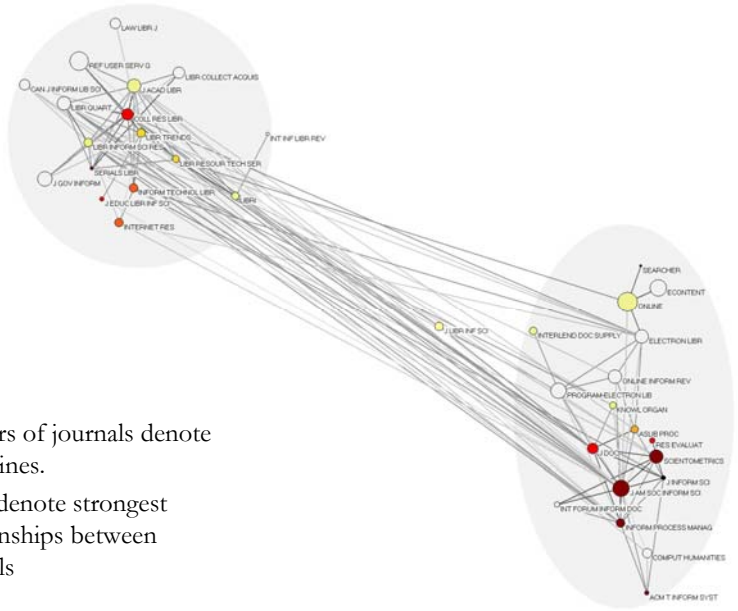
Structural Map: Studying Disciplinary Diffusion

- The 212 nodes represent clusters of journals for different disciplines.
- Nodes are labeled with their dominant ISI category name.
- Circle sizes (area) denote the number of journals in each cluster.
- Circle color depicts the independence of each cluster, with darker colors depicting greater independence.
- Lines denote strongest relationships between disciplines (citing cluster gives more than 7.5% of its total citations to the cited cluster).

81

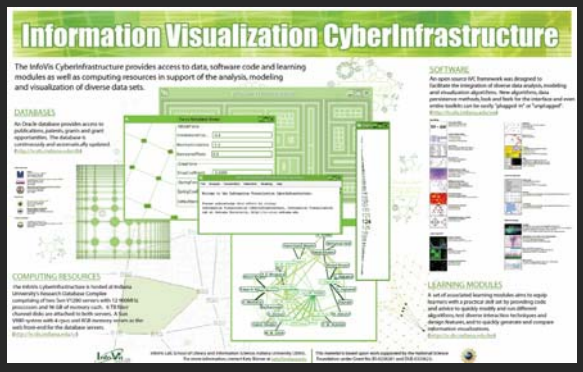


Zoom Into Structural Map



- Clusters of journals denote disciplines.
- Lines denote strongest relationships between journals

83



CAREER: Visualizing Knowledge Domains. NSF IIS-0238261 award (Katy Börner, $440,000) 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, Craig Stewart (Senior Personnel), $1,120,926) Sept. 05 - Aug. 08. https://nwb.slis.indiana.edu

84