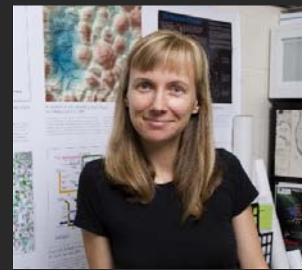
Computational Scientometrics That Informs Science Policy

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

International Council for Scientific and Technical Information Conference
Ottawa, Canada.
June 10, 2009



"Features that distinguish science from pseudoscience
are repeatability, economy, menuration, heuristics, and consilience."
E. O. Wilson in Consilience: The Unity of Knowledge (1998)

Introduction to Computational Scientometrics

General Scientometrics Workflow

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

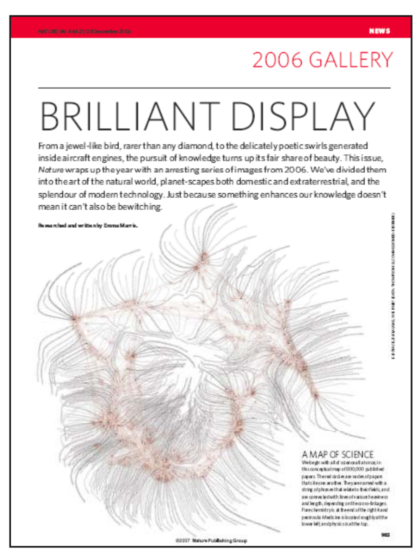
Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003) Visualizing Knowledge Domains. In Blaise Cronin (Ed.), [Annual Review of Information Science & Technology, Volume 37](), Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 5, pp. 179-255.

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



Scientometrics Opportunities

Advantages for Science Policy Makers/Funders

- Supports monitoring of (long-term) money flow and research developments, evaluation of funding strategies for different programs, decisions on project durations, funding patterns.
- Staff resources can be used for scientific program development, to identify areas for future development, and the stimulation of new research areas.

Advantages for Researchers

- Easy access to research results, relevant funding programs and their success rates, potential collaborators, competitors, related projects/publications **(research push).**
- More time for research and teaching.

Advantages for Industry

- Fast and easy access to major results, experts, etc.
- Can influence the direction of research by entering information on needed technologies **(industry-pull)**.

Advantages for Publishers

- Unique interface to their data.
- Publicly funded development of databases and their interlinkage.

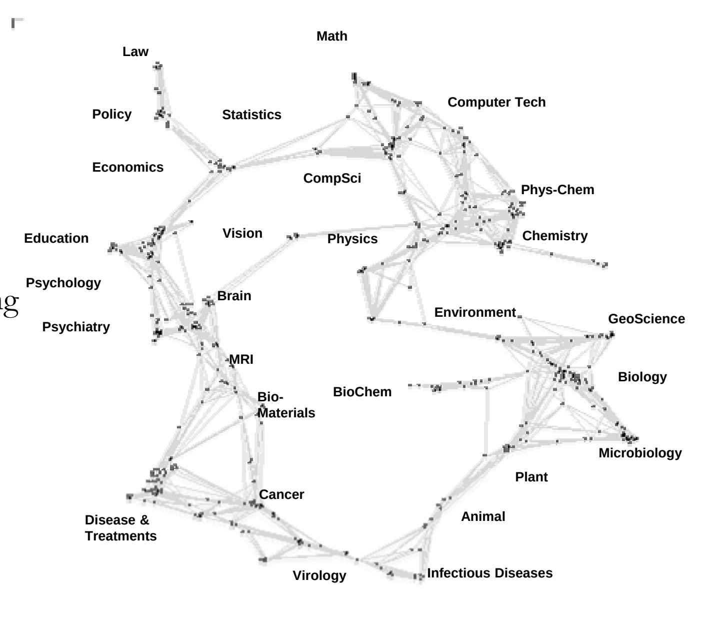
For Society

- Dramatically improved access to scientific knowledge and expertise.

2002 '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.

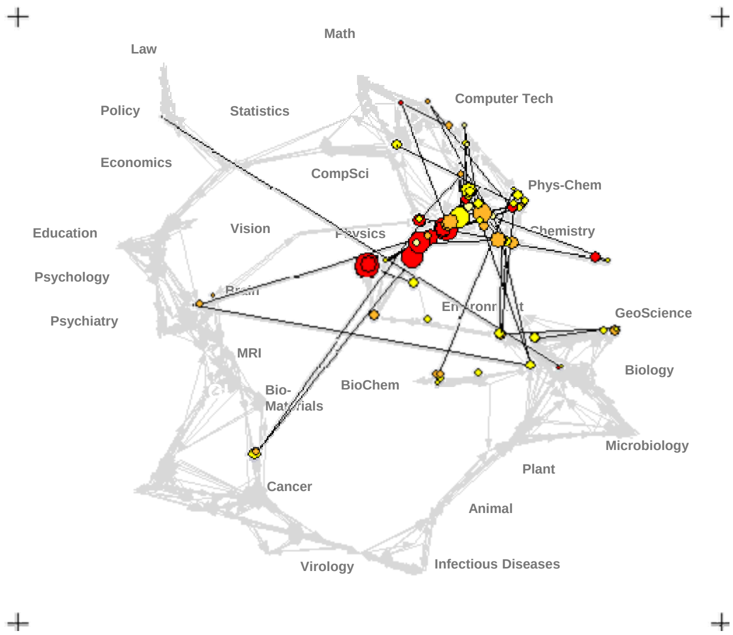
- Uses combined SCI/SSCI from 2002
  - 1.07M papers, 24.5M references, 7,300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts were reaggregated at the journal cluster level to calculate the
  - (x,y)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

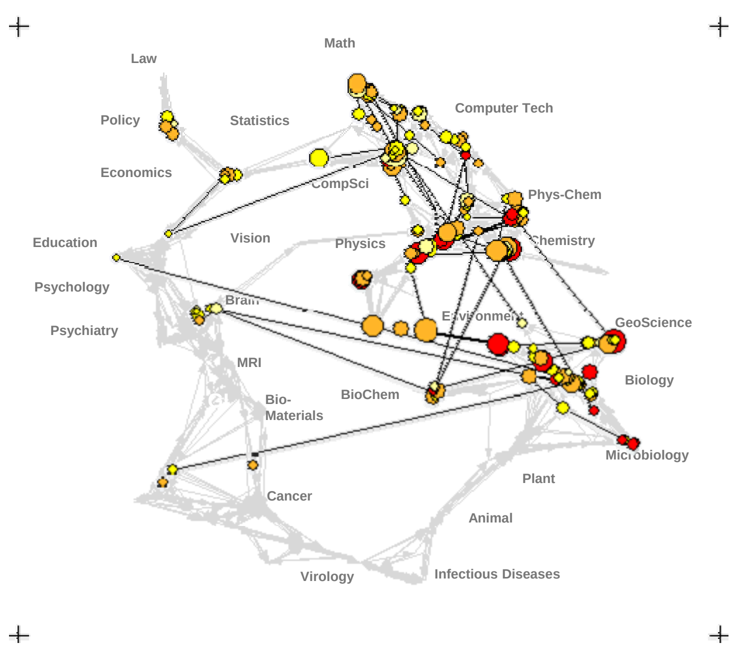
Funding patterns of the US Department of Energy (DOE)



Science map applications: Identifying core competency

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

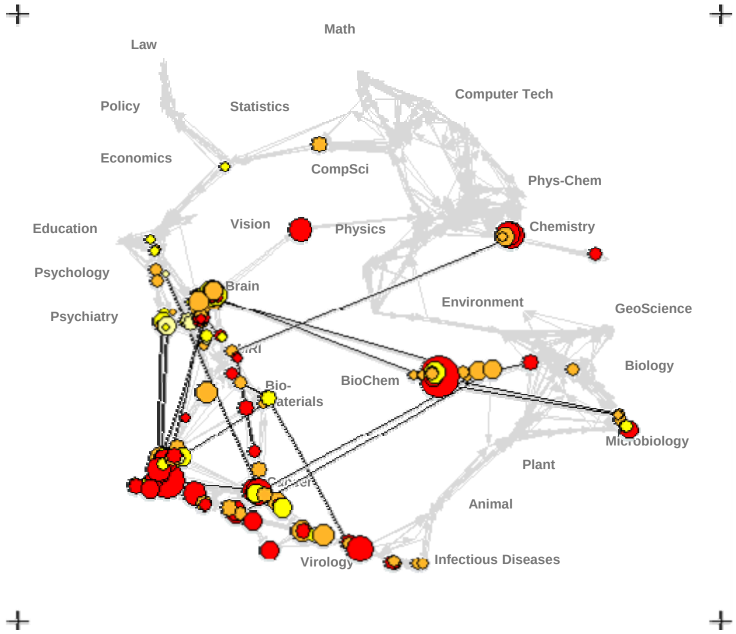
Funding Patterns of the National Science Foundation (NSF)



Science map applications: Identifying core competency

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

Funding Patterns of the National Institutes of Health (NIH)



What do Science Policy Makers want?

Needs Analysis

Reported are initial results of 34 interviews with science policy makers and researchers at

- Division director level at national, state, and private foundations (10),
- Program officer level (12),
- University campus level (8), and
- Science policy makers from Europe and Asia (4).

conducted between Feb. 8th, 2008 and Oct. 2nd, 2008.

Each interview comprised a 40 min, audio-taped, informal discussion on specific information needs, datasets and tools currently used, and on what a 'dream tool' might look and feel like. A pre-interview questionnaire was used to acquire demographics and a post-interview questionnaire recorded input on priorities.

Currently Used Datasets, Tools, and Hardware

In the pre-interview questionnaire subjects were asked “What databases do you use?”

- ***People databases*** such as agency internal PI & reviewer databases, human resources databases
- ***Publication databases*** such as WoS, Scopus; Dialogue (SCI, SSCI, Philosopher's Jadex), PUBmed/Pubmed Central, SciCit, IND, JStor, PsychInfo, Google scholar, agency/university library journal holdings (online), ISI/OIG databases, RePEc
- ***Patent databases*** such as PATSTAT, EPO, WPTO, and aggregators such as PatentLens, PatSTAT
- ***Intellectual property*** Public Intellectual Property Resource by UC Davis, SparcIP
- ***Funding databases*** such as NIH IMPACT II, SPIRES, QVR-internal NIH; NSF’s EIS, Proposal and Awards "PARS" "Electronic Jacket, IES Awards Database, USAspending.gov, Research.gov
- ***Federal reports*** such as SRS S&E Indicators, OECD data and statistics, Federal Budget databases, National Academies reports, AAAS reports, National Research Council (NRC) reports
- ***Survey data*** Taulbee Survey of CS salaries, NSF Surveys, EuroStats
- ***Internal proprietary databases*** at NSF, NIH, DOE
- ***Science databases such as*** FAO, USDA, GeneBank, TAIR, NCBI Plant genome
- ***Web data*** typically accessed via Google search
- ***News***, e.g., about federal budget decisions, Science Alerts from Science Magazine, Factiva, Technology Review, Science, Nature
- ***Expertise*** via stakeholder opinions, expert panels
- ***Management, trends, insights*** – from scientific societies, American Evaluation Association

Insight Needs

The pre-interview questionnaire asked “What would you most like to understand about the structure/evolution of science and why?” Responses can be grouped by

Science Structure and Dynamics
- Growth of interdisciplinary areas around a scientific field. Global growth of a scientific field.
- The development of disciplines and specialties (subdisciplines).
- how science is structured -- performers, funding sources, (international) collaborations.
- Grant size vs. productivity

Impact
- Criteria for quality. Scientific and public health impacts.
- Conditions for excellent science, use of scientific cooperation.
- Return on investment / impact spread of research discovery / impact of scientists on others.
- Does funding centers create a higher yield of knowledge than individual grants?

Feedback Cycles
- Linkages between S&E funding, educational and discovery outcomes, invention and technology development, economical and social benefit, at least generally applicable predictable system.
- The way institutional structures (funding/evaluation/career systems/agenda setting) influence the dynamics of science.
- Understanding the innovation cycle. Looking at history and identifying key technologies, surveying best practices for use today. Answer the question--"How best to foster innovation"?

Insights From Verbal Interviews

Different policy makers have very different tasks/priorities

Division directors

Rely mostly on experts, quick data access

Provide input to talks/testimonies, regulatory/legislator proposal reviews, advice/data

Compare US to other countries, identify emerging areas, determine impact of a decision on US innovation capacity, national security, health and longevity

Program officers

Rely more on data

Report to foundation, state, US tax payers

Identify ‘targets of opportunity' global), fund/support wisely (local), show impact (local+global)

University officials

Rely more on (internal) data

Make internal seed funding decisions, pool resources for major grant applications, attract the best students, get private/state support, offer best research climate/education.

All see people and projects as major “unit of analysis”.

Insights From Verbal Interviews

Major Task Types:

Connect

IP to companies, proposals to reviewers, experts to workshops, students to programs, researchers to project teams, innovation seekers to solution providers.

Impact and ROI Analysis

Scientific and public (health) impacts.

Real Time Monitoring

Funding/results, trajectories of people, bursts.

Longitudinal Studies

Understand dynamics of and delays in science system.

Computational Scientometrics Cyberinfrastructure

Provided by the Cyberinfrastructure for Network Science Center at Indiana University.

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

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

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

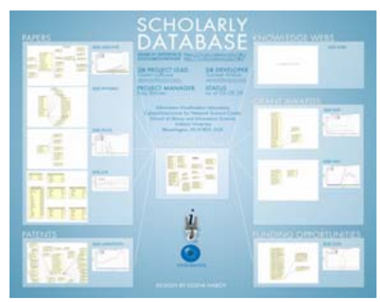
Scholarly Database
The Scholarly Database (SDB) at Indiana University aims to serve researchers and practitioners interested in the analysis, modeling, and visualization of large-scale scholarly datasets. The database currently provides access to over 20 million papers, patents and grants. Resulting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

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

http://sci.slis.indiana.edu

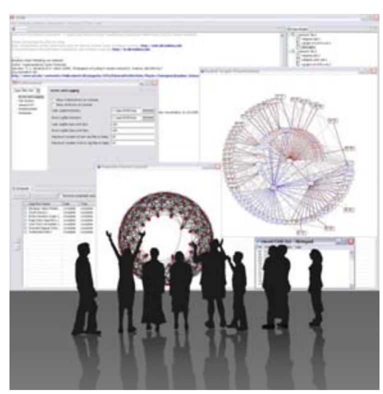
Scholarly Database of 23 million scholarly records
http://sdb.slis.indiana.edu

James S. McDonnell Foundation



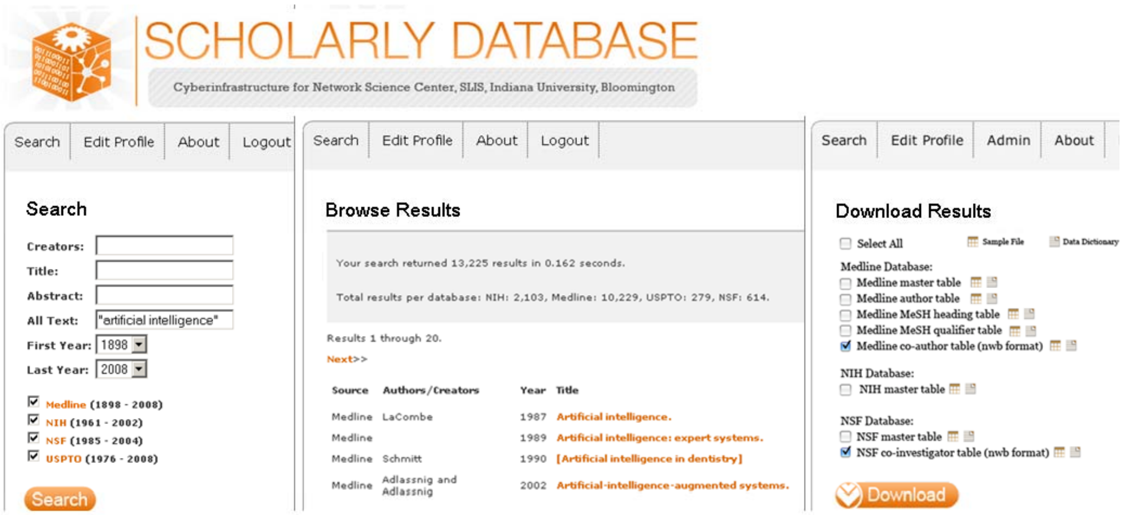
Information Visualization Cyberinfrastructure
http://iv.slis.indiana.edu

Network Workbench Tool and Community Wiki
NEW Scientometrics plugins
http://nwb.slis.indiana.edu



Epidemics Cyberinfrastructure
http://epic.slis.indiana.edu/

Scholarly Database: Web Interface

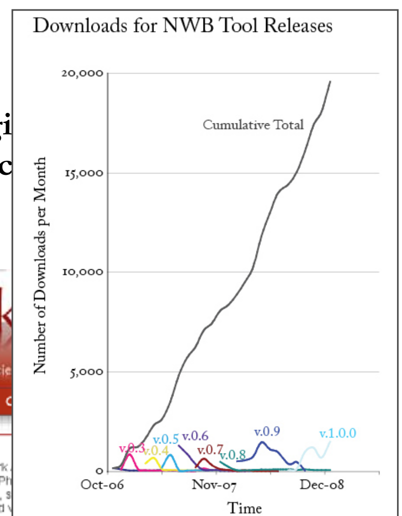


Anybody can register for free at https://sdb.slis.indiana.edu to search the about 23 million records and download results as data dumps.

In May 2009, SDB has over 170 registered users from academia, industry, and government from over 80 institutions and four continents.

Network Workbench Tool

An empty shell filled with algorithm plugi
will ultimately be 'packaged' as a SciPolic

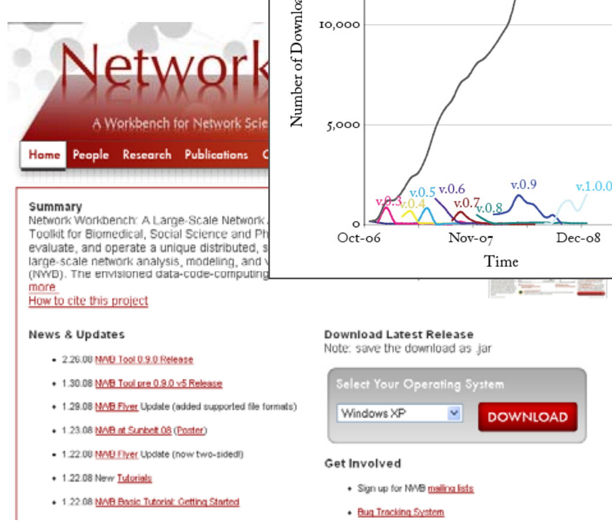


The Network Workbench (NWB) tool supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks.

In May 2009, the tool provides more 110 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

More than 40 of these plugins can be applied or were specifically designed for S&T studies.

It has been downloaded more than 18,000 times since Dec. 2006.



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

See https://nwb.slis.indiana.edu/community July 1st, 2008

Algorithms Currently Available

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?
- 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?

Unweighted & Undirected
- Based on degree/
  - Node Degree
  - Node Distribution
- Based on clustering
  - k-Nearest Neighbor
  - Watts Strogatz Clustering Coefficie
  - Watts Strogatz Clustering Coefficie
- 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?
  - Annotate K-Coreness?

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 Correla
- Unnamed Category?
  - Page Rank
- Based on local graph structure
  - Dyad Reciprocity?
  - Arc Reciprocity?
  - Adjacency Transitivity?
- Based on components
  - Weak Component Clustering
  - Extract Attractors?

Visualization Edit

Tools
- GUESS
- GnuPlot?

Predefined Positions Layout
- DrL (VxOrd)
- Pre-defined Positions (prefuse beta)?

Move
- Circular

Tree Layouts
- Radial Tree (prefuse alpha)
- Radial Tree with Annotations (prefuse beta)?
- Tree Map
- Tree View
- Balloon Graph (prefuse alpha)?

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)?
- LaNet (k-Core Decomposition)

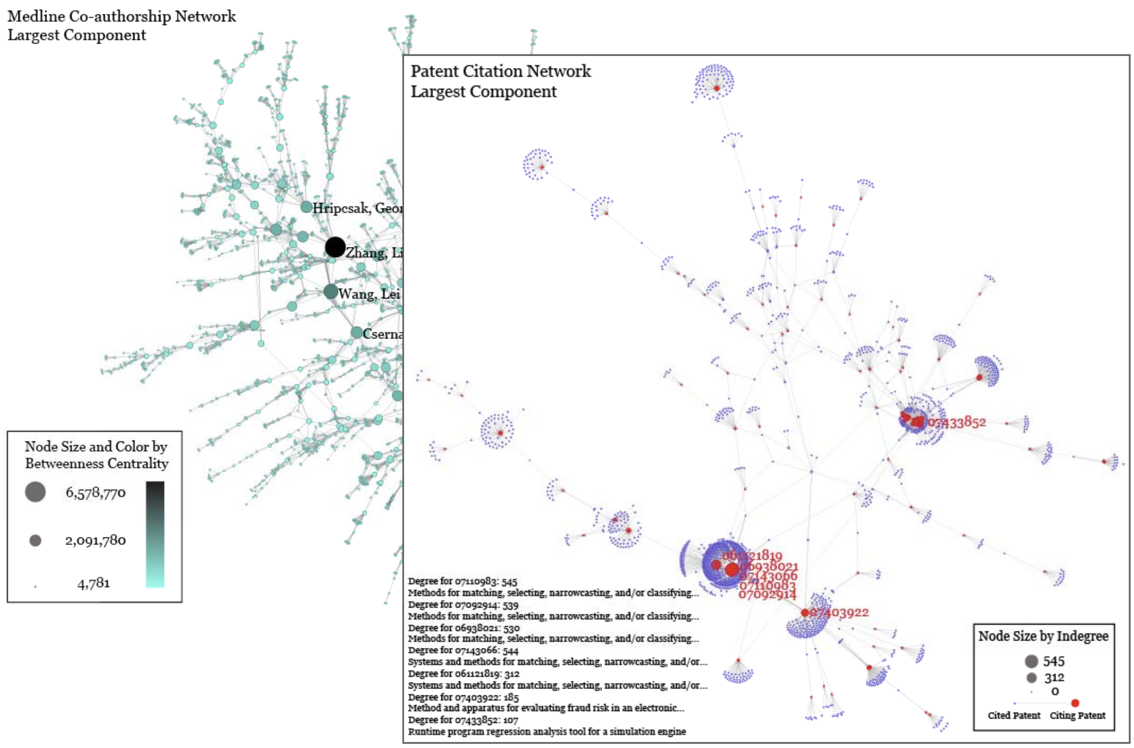
Scientometrics Edit

Extract Network From Table
- Extract Co-Authorship Network
- Extract Co-Occurrence Network From Table?
- Extract Directed Network From Table?

Extract Network From Another Network
- Extract Bibliographic Coupling Similarity Network
- Extract Co-Citation Similarity Network?

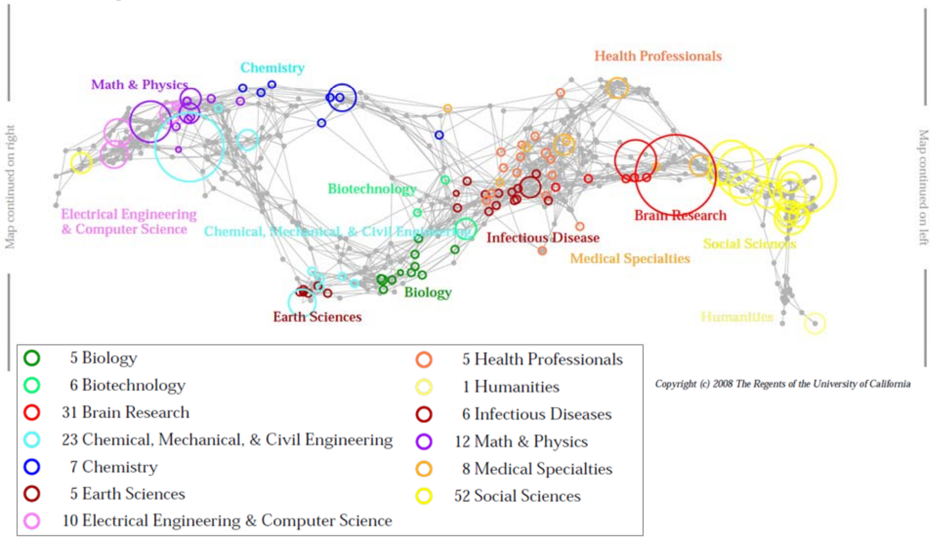
Cleaning
- Remove ISI Duplicate Records
- Detect Duplicate Nodes
- Remove Rows With Multitudinous Fields?

SciPolicy Studies - Using Open Data and Open Code

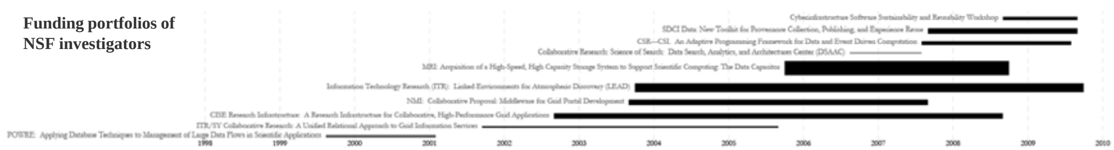


SciPolicy Studies - Using Open Data and Open Code

Science map overlays of base knowledge and/or core competencies



Funding portfolios of NSF investigators



Mapping Science Exhibit

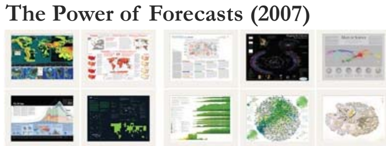
Mapping Science Exhibit – 10 Iterations in 10 years

http://scimaps.org/

The Power of Maps (2005)

The Power of Reference Systems (2006)

The Power of Forecasts (2007)



Science Maps for Economic Decision Makers (2008)

Science Maps for Science Policy Makers (2009)

Science Maps for Scholars (2010)

Science Maps as Visual Interfaces to Digital Libraries (2011)

Science Maps for Kids (2012)

Science Forecasts (2013)

How to Lie with Science Maps (2014)

Exhibit has been shown in 52 venues on four continents. Also at

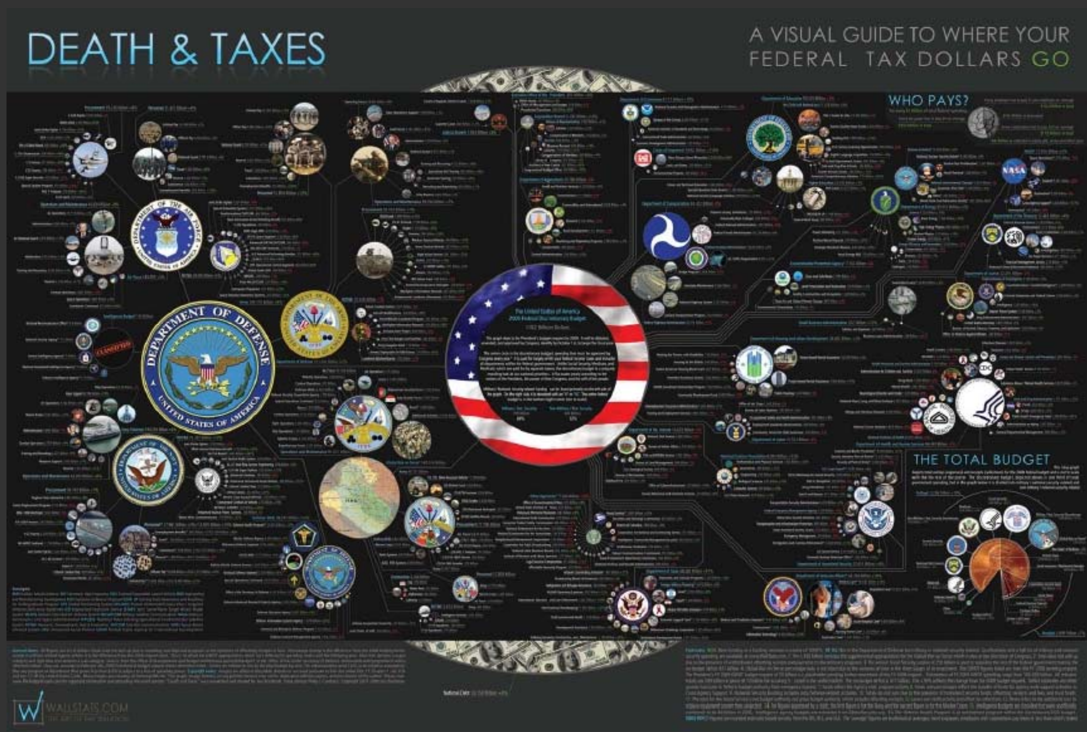
- NSF, 10th Floor, 4201 Wilson Boulevard, Arlington, VA.
- Chinese Academy of Sciences, China, May 17-Nov. 15, 2008.
- University of Alberta, Edmonton, Canada, Nov 10-Jan 31, 2009
- Center of Advanced European Studies and Research, Bonn, Germany, Dec. 11-19, 2008.

ORDER MAPS

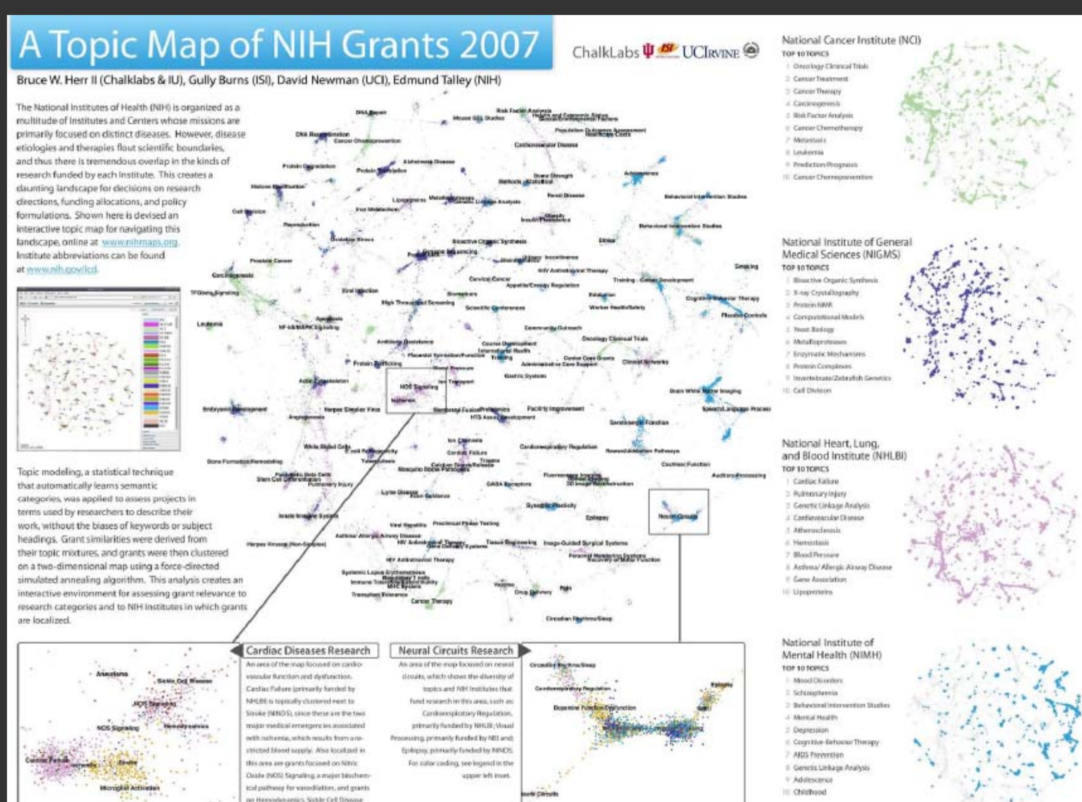
Debut of 5th Iteration of Mapping Science Exhibit at MEDIA X was on May 18, 2009 at Wallenberg Hall, Stanford University

http://mediax.stanford.edu

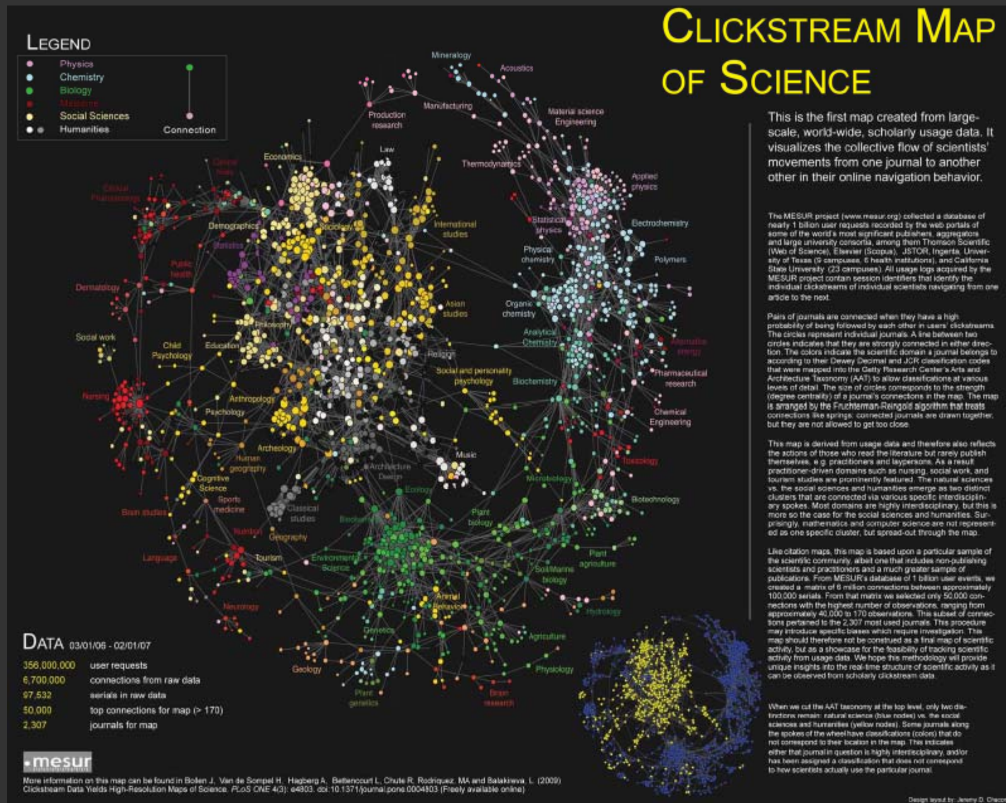
http://scaleindependentthought.typepad.com/photos/scimaps



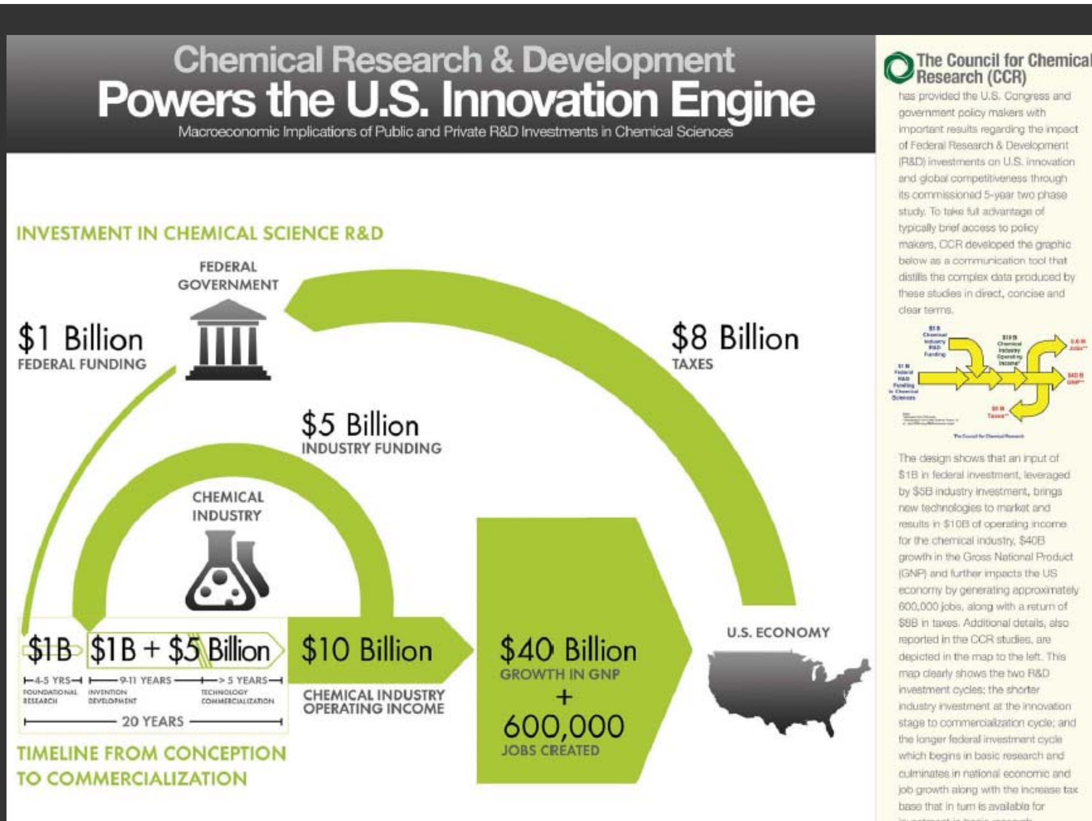
Death and Taxes 2009, by Jess Bachman
Courtesy of http://www.wallstats.com



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



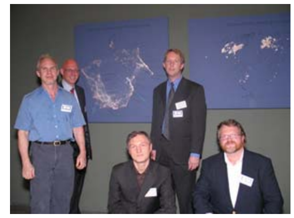
Bollen, Johan, Herbert Van de Sompel, Aric Hagberg, Luis M.A. Bettencourt, Ryan Chute, Marko A. Rodriquez, Lyudmila Balakireva. 2008. A Clickstream Map of Science.



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

Illuminated Diagram Display

W. Bradford Paley, Kevin W. Boyack, Richard Kalvans, and Katy Börner (2007) Mapping, Illuminating, and Interacting with Science. SIGGRAPH 2007.



Large-scale, high resolution prints illuminated via projector or screen.

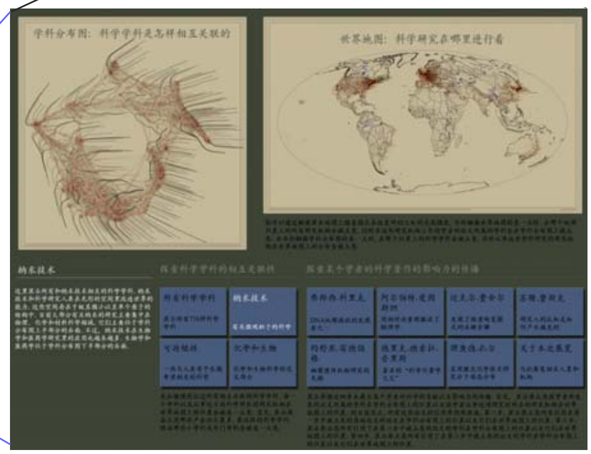
Questions:

- Who is doing research on what topic and where?
- What is the 'footprint' of interdisciplinary research fields?
- What impact have scientists?

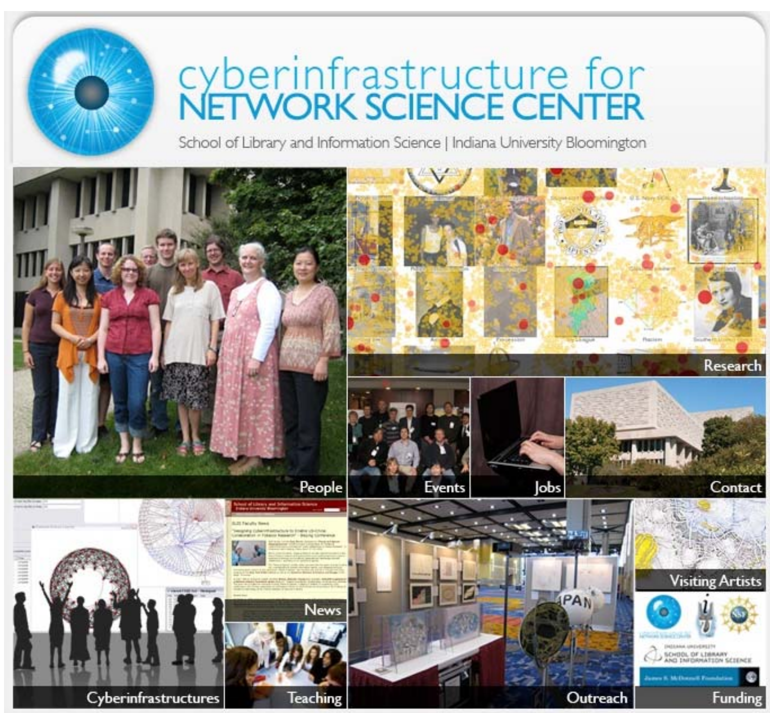
Interactive touch panel.

Contributions:

- Interactive, high resolution interface to access and make sense of data about scholarly activity.



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



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