Serving the Needs of Science Policy Decision Makers

(What Science Policy Makers Really Want and Resulting Research Challenges)

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

Thursday, May 14, 2009, 11:00 AM-12:30 PM EST

National Institutes of Health, Building 1, Room 151, Bethesda, MD 20892

Overview

1. Needs Analysis
Interview Results

2. Demonstrations
Scholarly Database (SDB) (http://sdb.slis.indiana.edu)
Science Policy plug-ins in Network Workbench Tool (http://nwb.slis.indiana.edu)

3. Discussion and Outlook
Shopping Catalog
Science of Science Cyberinfrastructure (http://sci.slis.indiana.edu)
Science Exhibit (http://scimaps.org)

1. Needs Analysis

Reported here 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.

Data compilation is in progress, should be completed in July 2009, and will be submitted as a journal paper.

1.1 Demographics

➤ Nine of the subjects were woman all others men.
➤ Most (31) checked English as their native language, the other four listed French, German, Dutch, and Japanese.
➤ Subjects' ages ranged from 31-40 (4), 41-50 (7), 51-60 (15), 60 (6), other subjects did not reveal age.

1.2 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

1.3 Currently Used Datasets, Tools, and Hardware

Asked to identify what tools they use in their daily work, subjects responded:
- MS Office   16
- MS Excel   11
- MS Word   7
- MS Powerpoint   5
- MS Access   4
- Internet (browser)   4
- SPSS   4
- Google   3
- SQL   3
- UCINET   3
- Adobe Acrobat   2
- Image editing software such as Photoshop 2
- Pajek   2

Only tools mentioned at least two times are listed here.

1.4 Currently Used Datasets, Tools, and Hardware

Asked to identify what hardware they use in their daily work, subjects responded:

- Windows PC   20
- Laptop   11
- Blackberry   6
- Mac   5
- PDS   2
- Cell phone   1
- IPod   1
- Printer   1

Five subjects reported that they use PC and Laptop and a Blackberry.

1.5 Desired Datasets and Tools

Major responses (* denotes existing datasets/tools)

Datasets

- Soc Sci Citation index, Scientific Citation Index, Impact Factors*
- DB of all faculty and industrial experts in a scientific field
- DB of academic careers, memberships in academic communities, reviews/refereeing histories
- DB that links government funding, patent, and IP databases
- DB that links publications and citations to funding awards
- "DB that collates from all dbs I currently access"

Tools

- Webcrawler, etc.
- Bio/timeslines of academic careers, outputs, impacts, career trajectories
- Virtual analytic software that is user friendly
- Visualization software / advanced graphics
- Videoconferencing capability*

1.6a 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"?

1.6b Insight Needs

The post-interview questionnaire asked What are your initial thoughts regarding the utility of science of science studies for improving decision making? How would access to datasets and tool speed up and increase the quality of your work?"

Excerpts of answers:
- Two areas have great potential: Understanding S&T as a dynamic system, means to display, visualize and manipulate large interrelated amounts of data in maps that allow better intuitive understanding.
- Look for new areas of research to encourage growth/broader impacts of research--how to assess/ transformative science--what scientific results transformed the field or created a new field/ finding panelists/reviews/ how much to invested until a plateau in knowledge generation is reached/how to define programs in the division.
- Scientometrics as cartography of the evolution of scientific practice that no single actor (even Nobel Laureates) can have. Databases provide a macro-view of the whole of scientific field and its structure. This is needed to make rational decision at the level of countries/states/provinces/regions.
- Understanding where funded scientists are positioned in the global map of science.
- Self-knowledge about effects of funding/ self-knowledge about how to improve funding schemes.
- Ability to see connections between people and ideas, integrate research findings, metadata, clustering career measurement, workforce models, impact (economic/social) on society-interactions between levels of science; lab, institution, agency, Fed Budget, public interests.
- It would be valuable to have tools that would allow one automatically to generate co-citation, co-authorship maps…I am particularly interested in network dynamics.

- It would enable more quantitative decision making in place of an "impression-based" system, and provide a way to track trends, which is not done now.
- When NSF started SciSIP, I was skeptical, but I am more disposed to the idea behind it now although I still don't have a clear idea what scientific metrics will be…..how they will apply across disciplines and whether it's really possible to predict with any accuracy the consequences of any particular decision of a grant award.
- SoS potentially useful to policymakers by providing qualitative and quantitative data on the impacts of science toward government policy goals…ideally these studies would enable policy makers to make better decisions for linking science to progress toward policy goals.
- Tracking faculty's work over time to determine what factors get in the way of productivity and which enhance, e.g. course-releases to allow more time--does this really work or do people who want to achieve do so in spite of barriers.
- I'm not sure that this has relevance to my decision-making. There is a huge need for more reliable data about my organization and similar ones, but that seems distinct from data and tools to study science.
- It would assist me enormously.
- Help to give precedents that would rationalize decisions--help to assess research outside one's major area. Ways of assessing innovation, ways of assessing interactions (among researchers, across areas, outside academia).
- It would allow me to answer questions from members of congress provide visual presentations of data for them.
- Very positive step--could fill important need in understanding innovation systems and organizations.

1.7 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".

All seem to need better data and tool access.

1.7 Insights From Verbal Interviews

Types of Tasks

Connect

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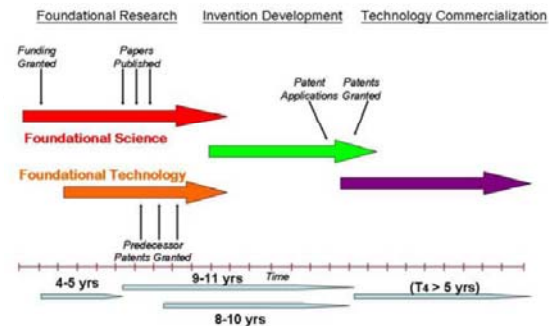providers

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.

http://www.ccrhq.org/publications_docs/CCRPhaseIIStudyReport.pdf

1.8 Conclusions

Science policy makers have very concrete needs yet little time/expertise to identify
the best datasets/tools.

There are several re-occurring themes such as the need for

- Scientific theories on the structure, dynamics, or cycles in science.

  (But see Science of Science & Innovation Policy listserv scisip@lists.nsf.gov,
  and Special Issue of *Journal of Informetrics,* 3(3), 2009 on "Science of Science:
  Conceptualizations and Models of Science". Editorial is available at
  http://ivl.slis.indiana.edu/km/pub/2009-borner-scharnhorst-joi-sos-intro.pdf)

- Higher data resolution, quality, coverage, and interlinkage.
- Easy way to try out/compare algorithms/tools.

Overview

1. Needs Analysis
Interview Results

2. Demonstrations
Scholarly Database (SDB) (http://sdb.slis.indiana.edu)
Science Policy plug-ins in Network Workbench Tool (http://nwb.slis.indiana.edu)

3. Discussion and Outlook
Shopping Catalog
Science of Science Cyberinfrastructure (http://sci.slis.indiana.edu)
Science Exhibit (http://scimaps.org)

2.1 Scholarly Database
http://sdb.slis.indiana.edu

Nianli Ma

"From Data Silos to Wind Chimes"
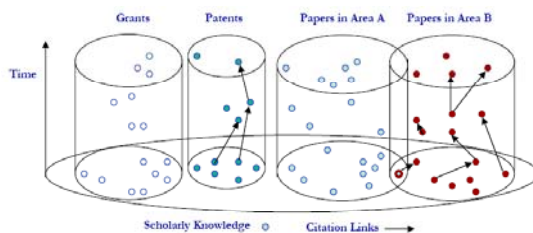
Figure 1: The interoperability and cross linkage problem. Many but not all of today's scholarly datasets, e.g., papers, patents, grants, are stored and made available so that 'vertical' citation linkages can be traversed. There are very few instances in which datasets of different origin and/or type are 'horizontally' interlinked.
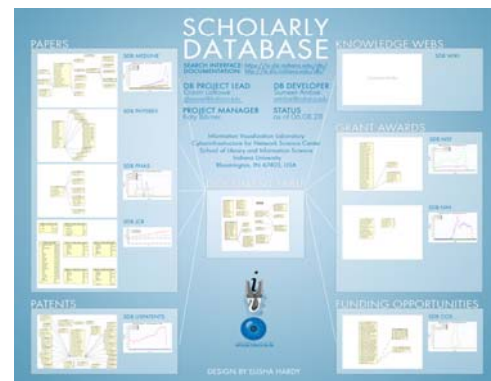
➢ Create public databases that any scholar can use. Share the burden of data cleaning and federation.
➢ Interlink creators, data, software/tools, publications, patents, funding, etc.

La Rowe, Gavin, Ambre, Sumeet, Burgoon, John, Ke, Weimao and Börner, Katy. (2007) The Scholarly Database and Its Utility for Scientometrics Research. In Proceedings of the 11th International Conference on Scientometrics and Informetrics, Madrid, Spain, June 25-27, 2007, pp. 457-462. http://ella.slis.indiana.edu/~katy/paper/07-issi-sdb.pdf

Scholarly Database: # Records & Years Covered

Datasets available via the Scholarly Database (* internally)

Dataset	# Records	Years Covered	Updated	Restricted Access
Medline	17,764,826	1898-2008	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, 710,952	1976-2008	Yes*	
NSF	174,835	1985-2002	Yes*	
NIH	1,043,804	1961-2002	Yes*	
Total	**23,167,642**	**1893-2006**	**4**	**3**

Aim for comprehensive time, geospatial, and topic coverage.

Scholarly Database: Web Interface

Anybody can register for free to search the about 23 million records and download results as data dumps.
Currently the system has over 120 registered users from academia, industry, and government from over 60 institutions and four continents.

Since March 2009:
Users can download networks:
 Co-author
 Co-investigator
 Co-inventor
  Patent citation
and tables for
burst analysis in NWB.
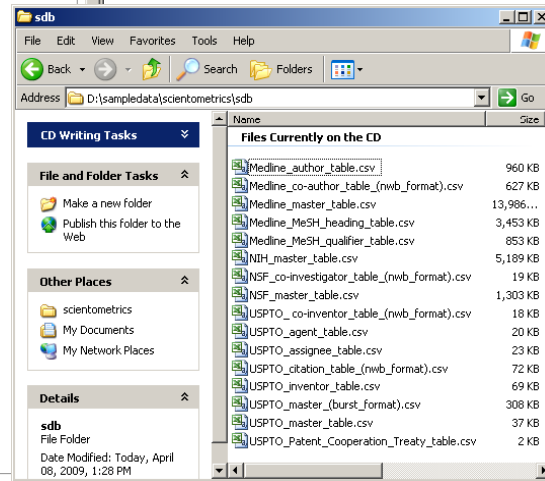
2.2 Scientometrics Filling of Network Workbench Tool

will ultimately be 'packaged' as a SciPolicy' 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 Feb. 2009, the tool provides more 100 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 19,000 times since Dec. 2006.

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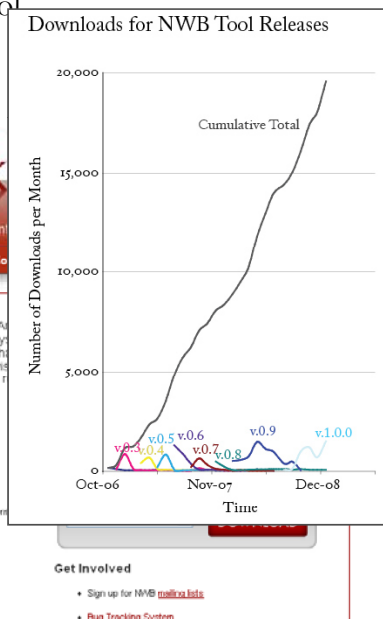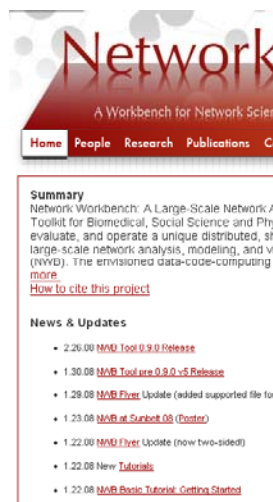
Herr II, Bruce W., Huang, Weixia (Bonnie), Penumarthy, Shashikant & Börner, Katy. (2007). Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. In Bainbridge, William S. & Roco, Mihail C. (Eds.), Progress in Convergence - Technologies for Human Wellbeing (Vol. 1093, pp. 161-179),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

NetworkWorkbench
A Workbench for Network Scientists

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, Mark Price, 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

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

21

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[?]
- 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

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[?]

NWB Tool: Output Formats

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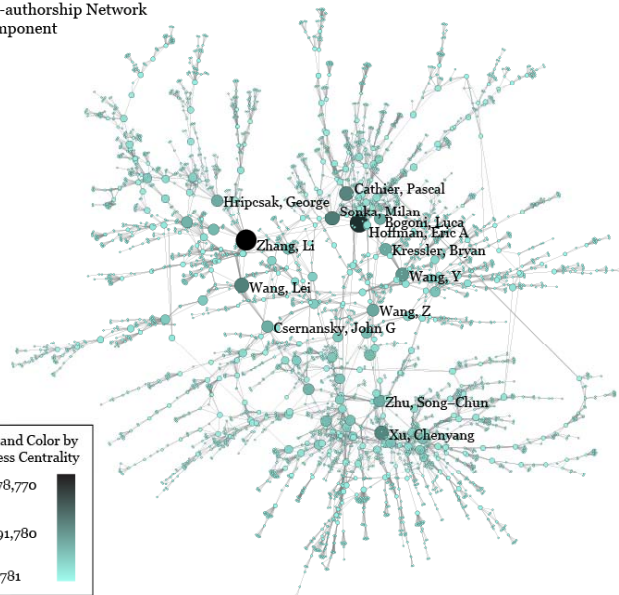- Pajek (.mat)
- GraphML (.xml or .graphml)
- XGMML (.xml)

GUESS
- Supports export of images into common image file formats.

Horizontal Bar Graphs
- saves out raster and ps files.

Medline Co-authorship Network Largest Component

Node Size and Color by Betweenness Centrality
- 6,578,770
- 2,091,780
- 4,781

Exemplary Analyses and Visualizations

Individual Level

A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.

B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

C. Indiana U, Cornell U, and Michigan U, extracting, and comparing Co-PI networks.

Scientific Field Level

D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.

Exemplary Analyses and Visualizations

Individual Level

A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.

B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

C. Indiana U, Cornell U, and Michigan U, extracting, and comparing Co-PI networks.

Scientific Field Level

D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.

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
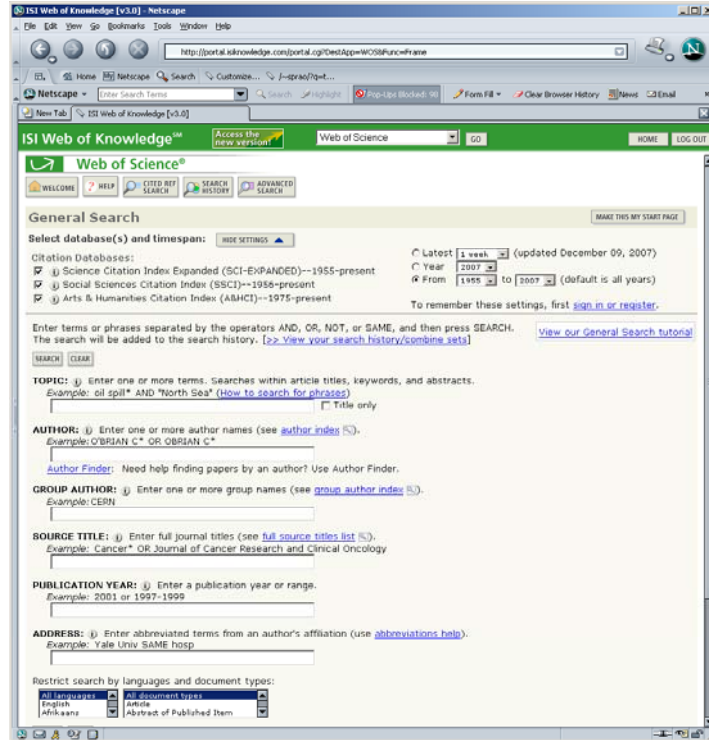
Comparison of Counts
No books and other non-WoS publications are covered.

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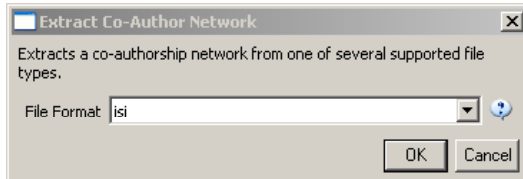
The result is an undirected network of co-authors in the Data Manager. It has 247 nodes and 891 edges.

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.*

Comparison of Co-Author Networks
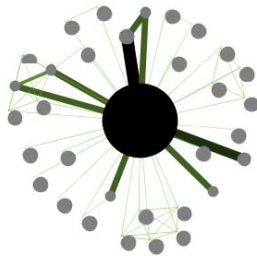
Eugene Garfield
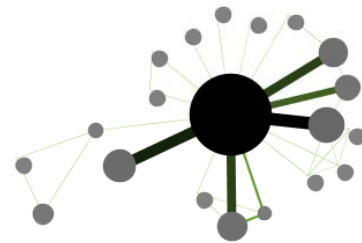
Stanley Wasserman
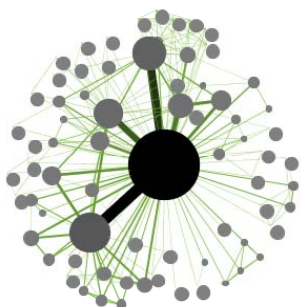
Alessandro Vespignani
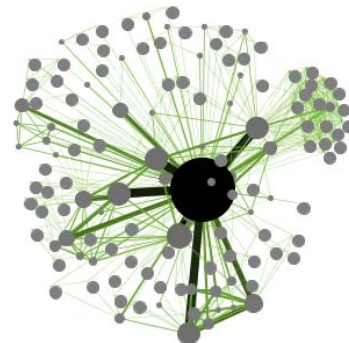
Albert-László Barabási

Joint Co-Author Network of all Four NetsSci Researchers

Node Size & Color
Number of Papers

Edge Size & Color
Number of Times
Co-Authored
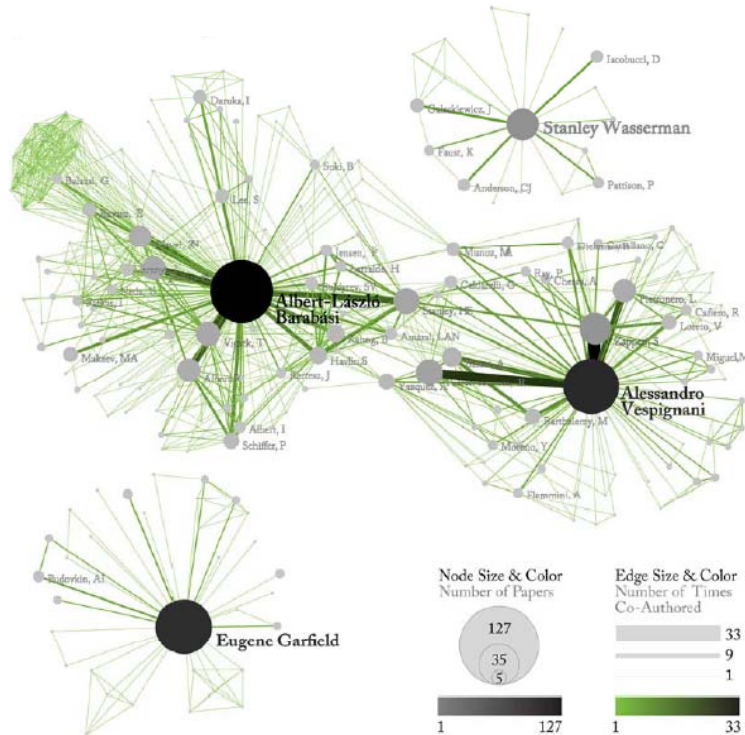
Paper-Citation Network Layout

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'*.
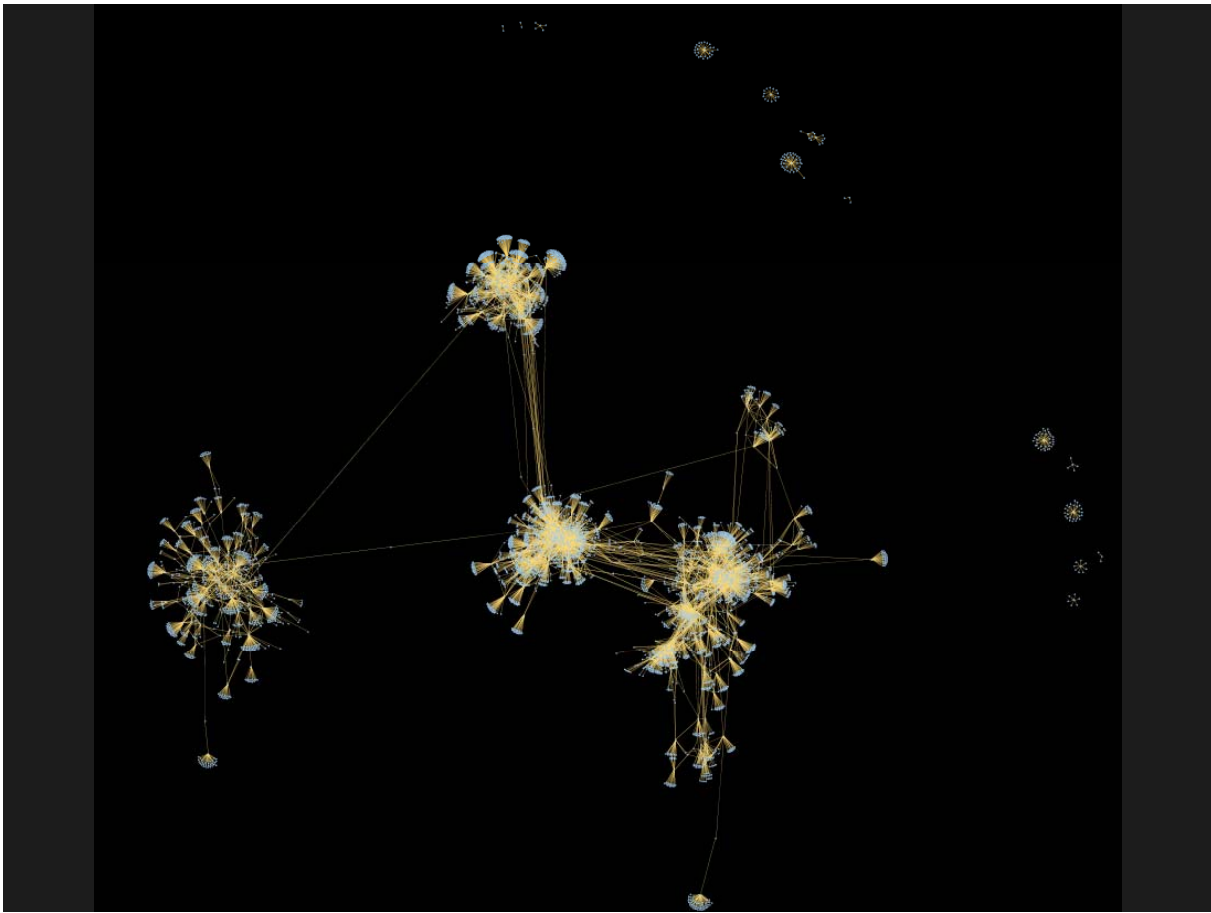
Exemplary Analyses and Visualizations

Individual Level

A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.

B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

C. Indiana U, Cornell U, and Michigan U, extracting, and comparing Co-PI networks.

Scientific Field Level

D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.

NSF Awards Search via http://www.nsf.gov/awardsearch

NSF Awards Search Results

Name	# Awards	First A. Starts	Total Amount to Date
Geoffrey Fox	27	Aug 1978	12,196,260
Michael McRobbie	8	July 1997	19,611,178
Beth Plale	10	Aug 2005	7,224,522

Disclaimer:

Only NSF funding, no funding in which they were senior personnel, only as good as NSF's internal record keeping and unique person ID. If there are 'collaborative' awards then only their portion of the project (award) will be included.

➢ Load into NWB, open file to count records, compute total award amount.

➢ Run '*Scientometrics > Extract Co-Occurrence Network*' using parameters:

➢ Select *"Extracted Network .."* and run *'Analysis > Network Analysis Toolkit (NAT)'*

➢ Remove unconnected nodes via *'Preprocessing > Delete Isolates'*.

➢ *'Visualization > GUESS'*, layout with GEM
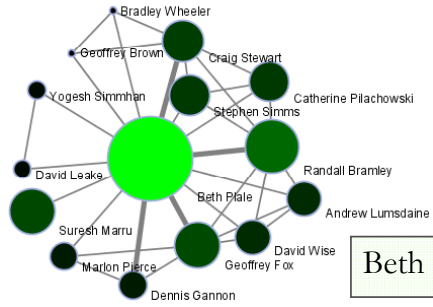
➢ Run 'co-PI-nw.py' GUESS script to color/size code.
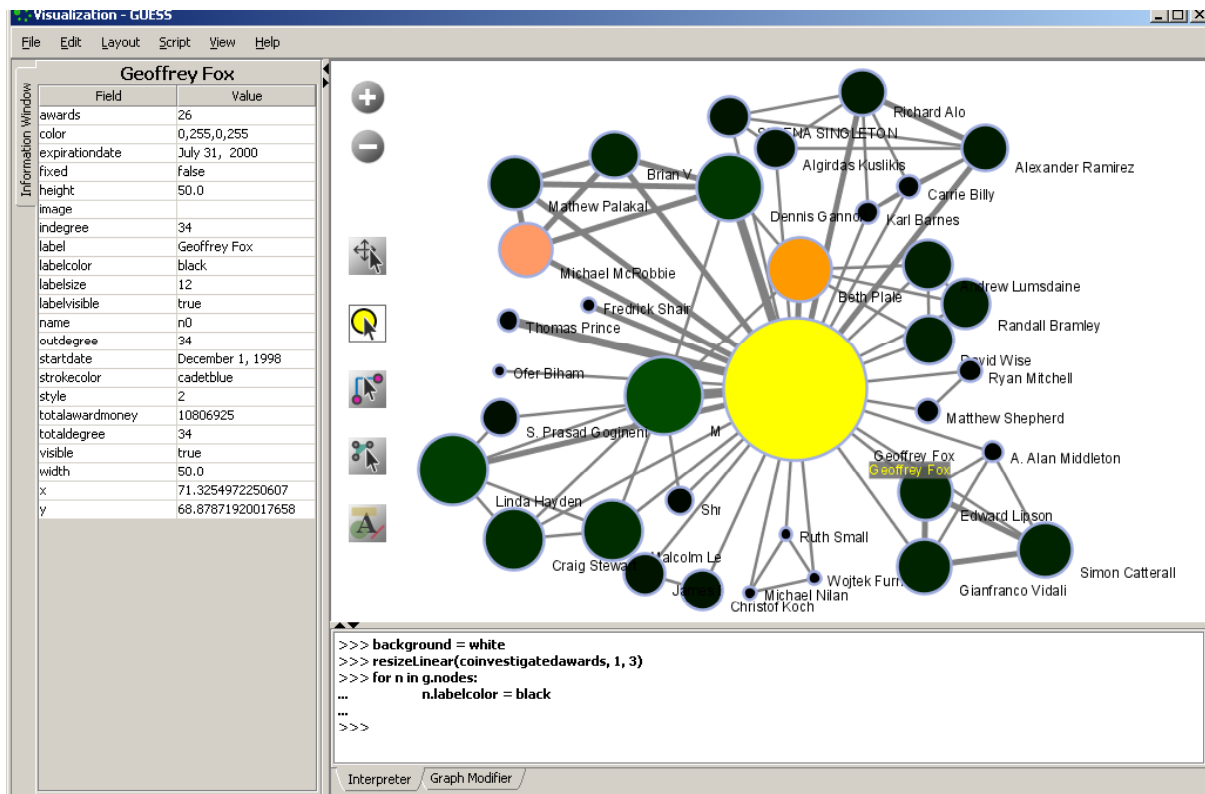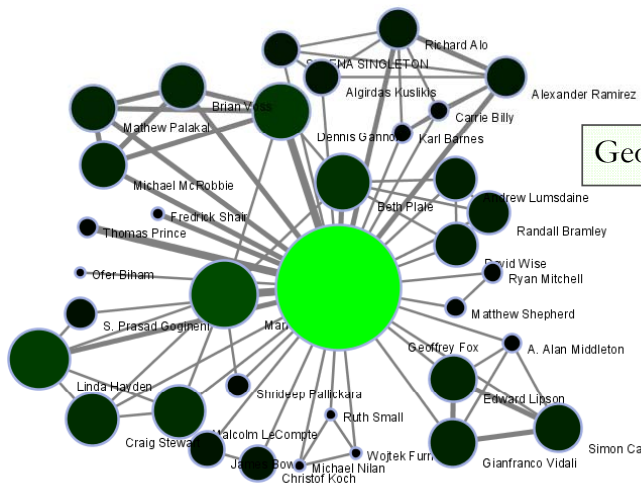
Geoffrey Fox
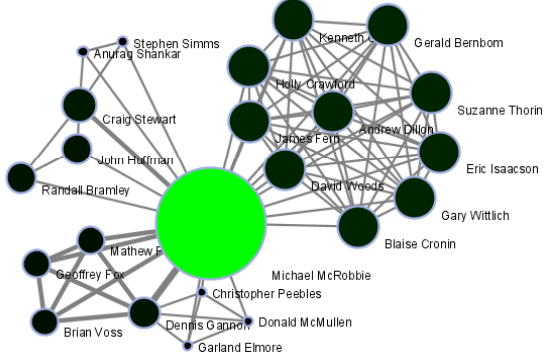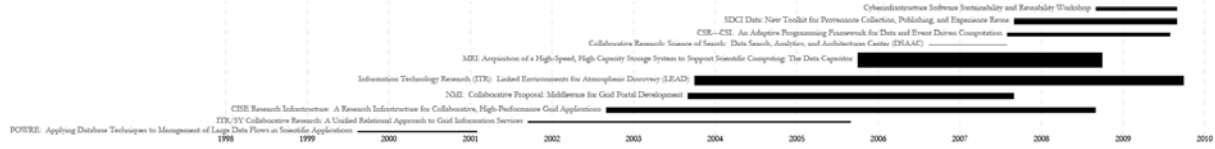
Michael McRobbie

Beth Plale

Geoffrey Fox

Last Expiration date

July 10

Michael McRobbie

Feb 10

Beth Plale

Sept 09

Exemplary Analyses and Visualizations

Individual Level

A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.

B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

C. Indiana U, Cornell U, and Michigan U, extracting, and comparing Co-PI networks.

Scientific Field Level

D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.

NSF Awards Search via http://www.nsf.gov/awardsearch

Save in CSV format as *institution*.nsf

Active NSF Awards on 11/07/2008:

➢ Indiana University   257

(there is also Indiana University at South Bend Indiana University Foundation, Indiana University Northwest, Indiana University-Purdue University at Fort Wayne, Indiana University-Purdue University at Indianapolis, Indiana University-Purdue University School of Medicine)

➢ Cornell University   501

(there is also Cornell University – State, Joan and Sanford I. Weill Medical College of Cornell University)

➢ University of Michigan Ann Arbor   619

(there is also University of Michigan Central Office, University of Michigan Dearborn, University of Michigan Flint, University of Michigan Medical School)

Save files as csv but rename into .nsf.

Or simply use the files saved in *'*yournwbdirectory*/sampledata/scientometrics/nsf/'*.

Extracting Co-PI Networks

Load NSF data, selecting the loaded dataset in the Data Manager window, run *'Scientometrics > Extract Co-Occurrence Network'* using parameters:
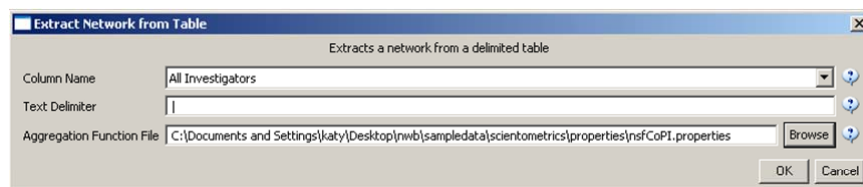
Two derived files will appear in the Data Manager window: the co-PI network and a merge table. In the network, nodes represent investigators and edges denote their co-PI relationships. The merge table can be used to further clean PI names.

Running the *'Analysis > Network Analysis Toolkit (NAT)'* reveals that the number of nodes and edges but also of isolate nodes that can be removed running *'Preprocessing > Delete Isolates'*.

Select *'Visualization > GUESS'* to visualize. Run 'co-PI-nw.py' script.

Indiana U: 223 nodes, 312 edges, 52 components

U of Michigan: 497 nodes, 672 edges, 117 c
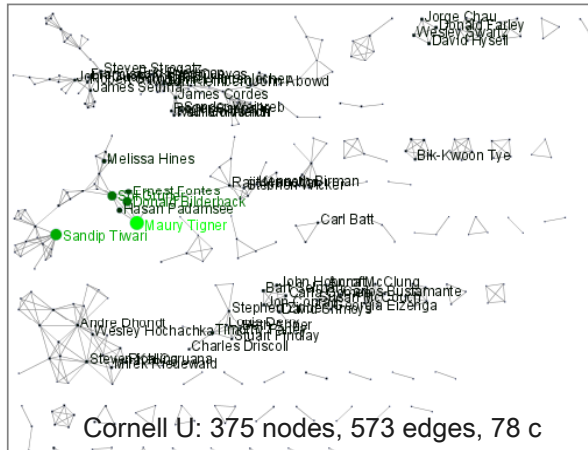
Cornell U: 375 nodes, 573 edges, 78 c

Extract Giant Component

Select network after removing isolates and run *'Analysis > Unweighted and Undirected > Weak Component Clustering'* with parameter
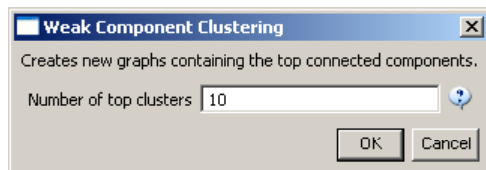
Indiana's largest component has 19 nodes, Cornell's has 67 nodes, Michigan's has 55 nodes.

Visualize Cornell network in GUESS using same .py script and save via *'File > Export Image'* as jpg.

Largest component of
Cornell U co-PI network

Node size/color ~ totalawardmoney
Top-50 totalawardmoney nodes are labeled.

Top-10 Investigators by Total Award Money

```
for i in range(0, 10):
    print str(nodesbytotalawardmoney[i].label) + ": " +
    str(nodesbytotalawardmoney[i].totalawardmoney)
```

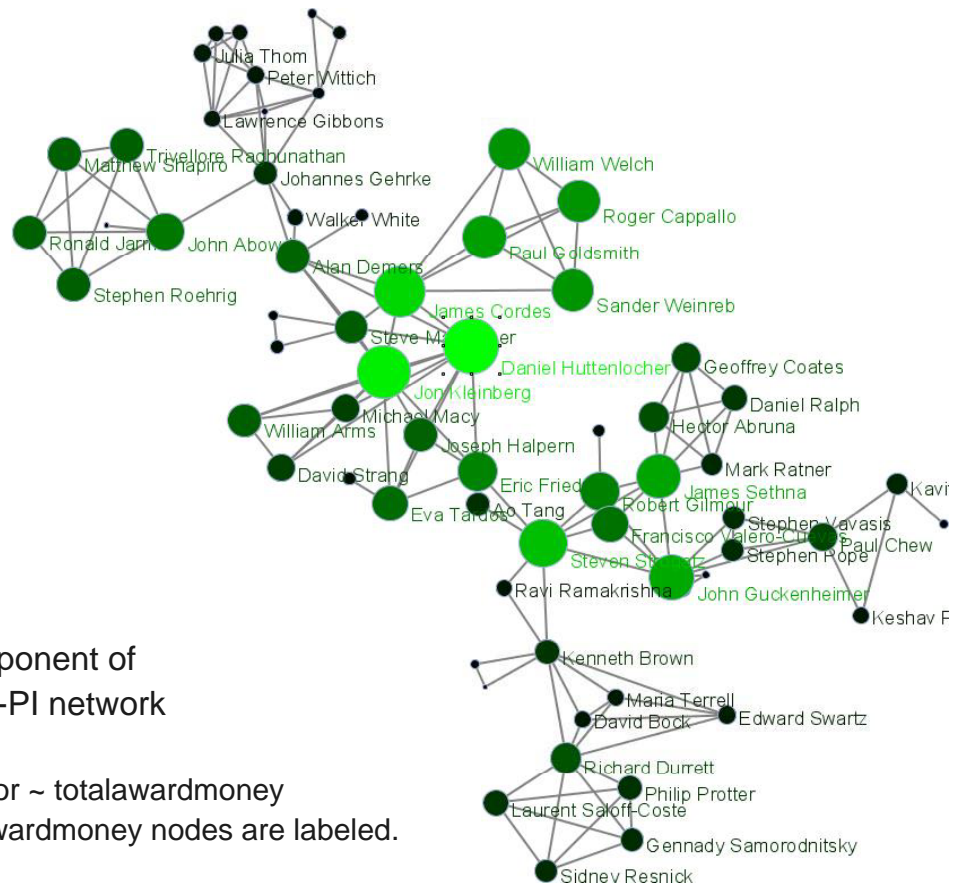| Indiana University		Cornell University		Michigan University	
Curtis Lively:	7,436,828	Maury Tigner:	107,216,976	Khalil Najafi:	32,541,158
Frank Lester:	6,402,330	Sandip Tiwari:	72,094,578	Kensall Wise:	32,164,404
Maynard Thompson:	6,402,330	Sol Gruner:	48,469,991	Jacquelynne Eccles:	25,890,711
Michael Lynch:	6,361,796	Donald Bilderback:	47,360,053	Georg Raithel:	23,832,421
Craig Stewart:	6,216,352	Ernest Fontes:	29,380,053	Roseanne Sension:	23,812,921
William Snow:	5,434,796	Hasan Padamsee:	18,292,000	Theodore Norris:	23,35,0921
Douglas V. Houweling:	5,068,122	Melissa Hines:	13,099,545	Paul Berman:	23,350,921
James Williams:	5,068,122	Daniel Huttenlocher:	7,614,326	Roberto Merlin:	23,350,921
Miriam Zolan:	5,000,627	Timothy Fahey:	7,223,112	Robert Schoeni:	21,991,140
Carla Caceres:	5,000,627	Jon Kleinberg:	7,165,507	Wei-Jun Jean Yeung:	21,991,140

Exemplary Analyses and Visualizations

Individual Level

A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.

B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

C. Indiana U, Cornell U, and Michigan U, extracting, and comparing Co-PI networks.

Scientific Field Level

D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.

Medline Co-authorship Network
Largest Component

Cathier, Pascal
Hripcsak, George
Sonka, Milan
Bogoni, Luca
Hoffman, Eric A
Zhang, Li
Kressler, Bryan
Wang, Y
Wang, Lei
Wang, Z
Csernansky, John G
Zhu, Song–Chun
Xu, Chenyang

Node Size and Color by
Betweenness Centrality

6,578,770

2,091,780

4,781

Patent Citation Network
Largest Component

07433852

061121819
06938021
07143066
07110983
07092914

07403922

Degree for 07110983: 545
Methods for matching, selecting, narrowcasting, and/or classifying...
Degree for 07092914: 539
Methods for matching, selecting, narrowcasting, and/or classifying...
Degree for 06938021: 530
Methods for matching, selecting, narrowcasting, and/or classifying...
Degree for 07143066: 544
Systems and methods for matching, selecting, narrowcasting, and/or...
Degree for 061121819: 312
Systems and methods for matching, selecting, narrowcasting, and/or...
Degree for 07403922: 185
Method and apparatus for evaluating fraud risk in an electronic...
Degree for 07433852: 107
Runtime program regression analysis tool for a simulation engine

Node Size by Indegree

545

312

0

Cited Patent   Citing Patent

Top-10 burst terms from abstracts of the AI search results.

Medline

Word	Length	Weight	Start	End
medical	17	299.7924	1983	1999
knowledge	5	293.9375	1991	1995
knowledge	6	215.2407	1997	2002
expert	13	171.0443	1985	1997
systems	15	170.3306	1985	1999
intelligence	21	123.9794	1981	2001
patient	21	123.9297	1982	2002
care	12	106.5522	1990	2001
registration	5	104.8139	2005	
knowledge-based	16	98.83778	1987	2002

NIH

Word	Length	Weight	Start	End
Phase	8	117.2205	1993	2000
commercial	9	87.57158	1995	
proposed	9	87.57158	1995	
mass	3	83.36952	1978	1980
protein	1	72.15788	1988	1988
networks	4	71.252	1993	1996
patterns	3	66.44826	1977	1979
being	8	66.29254	1971	1978
reasoning	2	65.68178	1984	1985
expert	4	60.49935	1987	1990

NSF

Word	Length	Weight	Start	End
their	6	47.05097	1999	
gray	2	28.19808	2000	2001
learning	2	27.40728	1997	1998
human	5	25.4525	2000	
control	2	24.07877	1992	1993
knowledge	1	21.48756	1998	1998
students	1	21.07674	1997	1997
problems	2	20.77133	1998	1999
more	2	19.96109	2000	2001
use	1	19.38503	2001	2001

USPTO

Word	Length	Weight	Start	End
human	3	19.03937321	2004	2006
video	3	15.32736425	1998	2000
disclosed	2	14.06694671	1999	2000
neural	3	13.30105906	2004	2006
"correct"	2	12.4336047	1999	2000
unit	2	12.35745838	2002	2003
material	1	12.08487035	2000	2000
feedback	1	12.07730195	2000	2000
rule	1	12.07730195	2000	2000
elevator	4	11.83351857	1991	1994

Overview

1. Needs Analysis
Interview Results

2. Demonstrations
Scholarly Database (SDB) (http://sdb.slis.indiana.edu)
Science Policy plug-ins in Network Workbench Tool (http://nwb.slis.indiana.edu)

3. Discussion and Outlook
Shopping Catalog
Science of Science Cyberinfrastructure (http://sci.slis.indiana.edu)
Science Exhibit (http://scimaps.org)

3.1 Shopping Catalog

A registry of existing datasets, tools, services, expertise and their
- Utility (insights provided, time savings based on scientific research/evaluations)
- Cost (dollars but also expertise/installation/learning time)
- How to learn more/order

Many datasets and tools are freely available. There will be (seasonal) special offers.

Catalog will be available in print (to peruse in plane) and online (to get download counts for ranking) but also comments, ratings.

Print version is funded by NSF's SciSIP program and should come out in Aug 2009. Feel free to sign up for it.

3.2 Science of Science Cyberinfrastructure

That builds on industry standards such as OSGi (NWB, soon also Cytoscape, MyExperiment), Joomla! (ZeroHUB).

Is staged: research -> development -> production code that comes with 24/7 support.

Addresses the needs of science policy makers and is easy to use.

Science of Science Cyberinfrastructure
PORTAL

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

3.3 Mapping Science Exhibit – 10 Iterations in 10 years
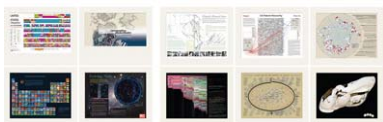http://scimaps.org/

The Power of Maps (2005)

Science Maps for Economic Decision Makers (2008)

The Power of Reference Systems (2006)

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)

The Power of Forecasts (2007)

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.

Debut of 5th Iteration of Mapping Science Exhibit at MEDIA X, Stanford University
May 18, 5-6:30pm Reception, Wallenberg Hall
http://mediax.stanford.edu
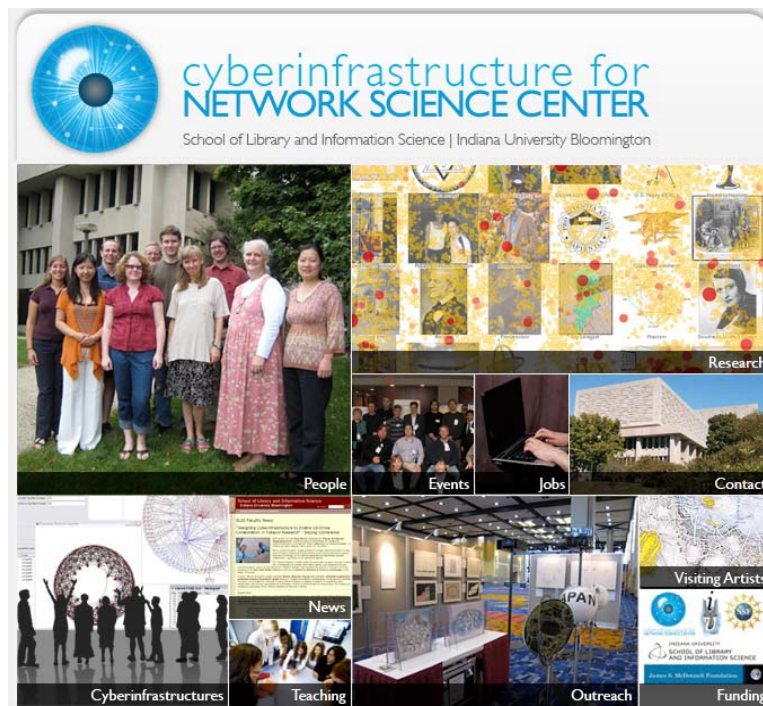http://scaleindependentthought.typepad.com/photos/scimaps

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

This is the only mockup in this slide show.

Everything else is available today.

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