Evolving and Emerging Populations and Topics Extracted from NSF Awards

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

With special thanks to Kevin W. Boyack, Micah Linnemeier, Russell J. Duhon, Patrick Phillips, Joseph Biberstine, Chintan Tank Nianli Ma, Hanning Guo, Mark A. Price, Scott Weingart

Virtual Presentation to NSF on June 7, 2010

Overview

Analyses using NSF award data from SharePoint

- Identify emerging areas in Career awards using burst analysis
- Evolving geospatial coverage of IGERT awards
- Topical/science coverage of MRI awards
- Co-Investigator network of all 51k NSF awards

Topic Analysis using data provided by David Newman

- Topics covered by NSF funding

Alternative Analyses that require additional data

- Flexible network extraction workflows
- Geospatial coding and visualization tools
- Temporal/textual analysis for examining topical trends
- RefMapper tool for analyzing the interdisciplinarity of grant proposals

This project uses the NSF SciSIP funded ***Science of Science (Sci²) Tool*** freely available as open source code and with tutorial at http://sci.slis.indiana.edu/sci2.

Evolving Networks

using NSF award data from SharePoint

Data Provided

All NSF awards that have been active at any time between Oct 2005 and Jan 2010 were provided by Paul Markovitz. The query was not limited by scientific term, program or program officer.

The data was retrieved from the Research Spending and Results (RS&R) service on Research.gov: http://www.research.gov/rgov/anonymous.portal?_nfpb=true&_pageLabel=page_research_funding_search&_nfls=false around Feb 3, 2010 (give or take a day).

The challenge with using RS&R for this purpose is that it is a search-based service, while you want all the records. The Excel spreadsheets were created by executing one query to retrieve all the information (contained in the spreadsheet columns) from the RS&S database(s) then exporting the results to Excel. The size of the resulting Excel spreadsheet was too large to upload to the SharePoint site (I think there is a 50M restriction on file uploads) so I divided the spreadsheet into 3 spreadsheets, named

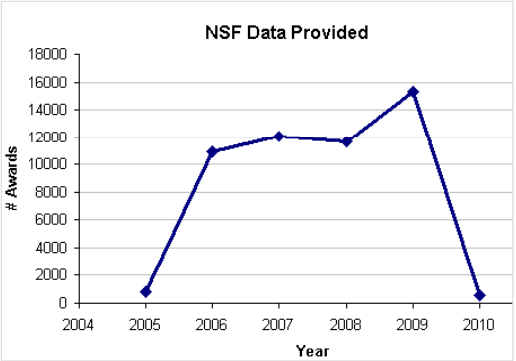
NSF awards Oct 2005_to_June 2007 as of Feb03.xlsx
NSF awards July 2007_to_Sept 2008 as of Feb03.xlsx
NSF awards Oct 2008_to_Jan 2010 as of Feb03.xlsx

then uploaded them.

Data Counts and Subsets

- NSF awards starting Oct 2005_to_June 2007 as of Feb03.xlsx (16,762 records)
- NSF awards starting July 2007_to_Sept 2008 as of Feb03.xlsx (18,000)
- NSF awards starting Oct 2008_to_Jan 2010 as of Feb03.xlsx (16,561 records)

Year	Count	First Awarded Date	Last Awarded Date
2005	831	10/5/2005	12/30/2005
2006	10942	1/3/2006	12/29/2006
2007	12031	1/3/2007	12/31/2007
2008	11622	1/2/2008	12/31/2008
2009	15312	1/2/2009	12/31/2009
2010	584	1/4/2010	2/2/2010
Total Count	**51322**		



4 complete years only.

From the 51,322 awards, we deleted "test proposals" resulting in 51,217 records.

We extracted

- IGERT awards (87 records)
- Career awards (2409 records)
- MRI awards (118 records)

Data Comparison with NSF Awards Search

Comparing dataset with queries run on NSF's Award search (http://www.nsf.gov/awardsearch) on 2010.05.13-14:

- IGERT (**87** records) NSFawardsearch retrieves **235** for "IGERT" in title with unchecked Historical Awards, Active Awards Only, Expired Awards Only. Excluded one research award on the impact of IGERTS. **114** awards start before Oct 05, **121** after.
- Career (**2409** records) NSFawardsearch retrieves **more than 3000** hits for active awards that have "career:" in the title. In *Feb03.xlsx there are **30** awards that started before Oct 2005. In the NSFawardsearch result there are **512** record that start in 2008 while the *Feb03.xlsx files show 530 records.
- MRI (**118** records) NSFawardsearch retrieves **1746** for "MRI" in title with unchecked Historical Awards, Active Awards Only, Expired Awards Only.

This data and the subsequent analyses should not be used for decision making before the accuracy of the data is confirmed.
SharePoint data is used subsequently.

Data Fields Comparison

** Dollar amount

statistical, temporal, geospatial, topical, and network analyses.

NSF's Awards Search

Award Number
Title
NSF Organization
Program(s)
Start Date
Last Amendment Date
Principal Investigator
State
Organization
Award Instrument
Program Manager
Expiration Date
Awarded Amount to Date**
Co-PI Name(s)
PI Email Address
Organization Street Address
Organization City
Organization State
Organization Zip
Organization Phone
NSF Directorate
Program Element Code(s)
Program Reference Code(s)
Field Of Application(s)
Award Number
Abstract

*Feb03.xlsx provided via Sharepoint

AWARDEE
DOING_BUSINESS_AS_NAME
PI_NAME
COPI
PI_PHONE, PI_EMAIL
AWARD_DATE
ESTIMATED_TOTAL_AWARD_AMOUNT
FUNDS_OBLIGATED_TO_DATE**
AWARD_START_DATE
AWARD_EXPIRATION_DATE
TRANSACTION_TYPE
AGENCY
CFDA_NUMBER
PRIMARY_PROGRAM_SOURCE
AWARD_TITLE_OR_DESCRIPTION
FEDERAL_AWARD_ID_NUMBER
DUNS_ID, PARENT_DUNS_ID
PROGRAM_NAME
PROGRAM_OFFICER_NAME, PROGRAM_OFFICER_PHONE, PROGRAM_OFFICER_EMAIL
AWARDEE_STREET_1, AWARDEE_STREET_2, AWARDEE_CITY, AWARDEE_STATE, AWARDEE_ZIP, AWARDEE_COUNTY, AWARDEE_COUNTRY, AWARDEE_CONG_DISTRICT
PERFORMING_ORG_NAME
PERFORMING_STREET_1, PERFORMING_STREET_2, PERFORMING_CITY, PERFORMING_STATE, PERFORMING_ZIP, PERFORMING_COUNTY, PERFORMING_COUNTRY, PERFORMING_CONG_DISTRICT
ABSTRACT_AT_TIME_OF_AWARD

Type of Analysis vs. Scale of Level of Analysis

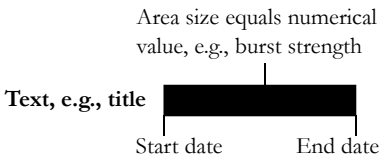
	Micro/Individual (1-100 records)	***Meso/Local (101–10,000 records)***	***Macro/Global (10,000 < records)***
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of USA, all of science.
Temporal Analysis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

Temporal Analysis (When)

Bursty terms in titles of Career awards

Burst Analysis using Kleinberg's algorithm.

Results are visualized using Horizontal Bar Graph



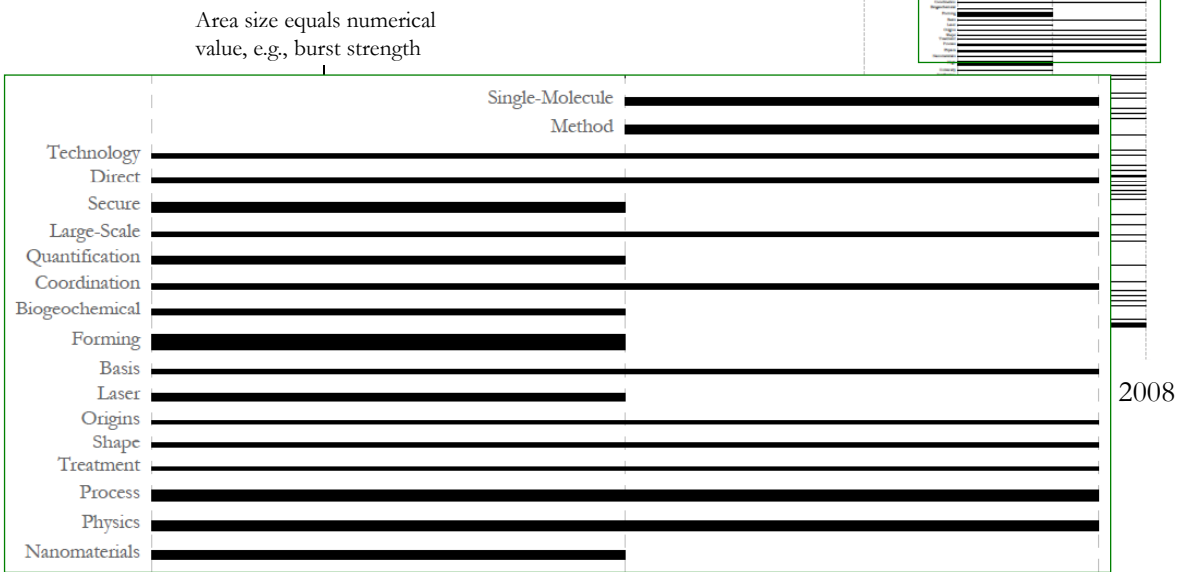
Terms that are still bursty are not shown.

Temporal Analysis (When)

Bursty terms in titles of Career awards

Burst Analysis using Kleinberg's algorithm.

Results are visualized using Horizontal Bar Graph

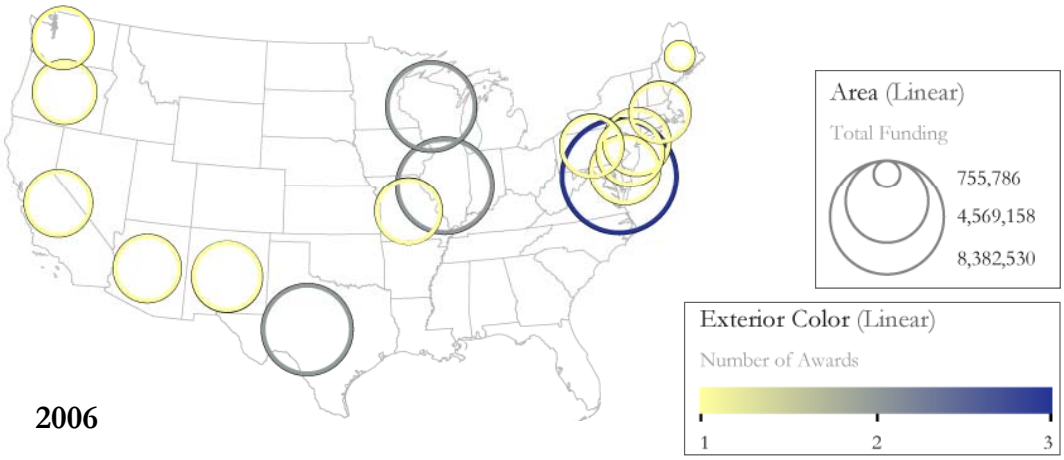


Geospatial Analysis (Where)

Evolving Geospatial Coverage of IGERT Awards

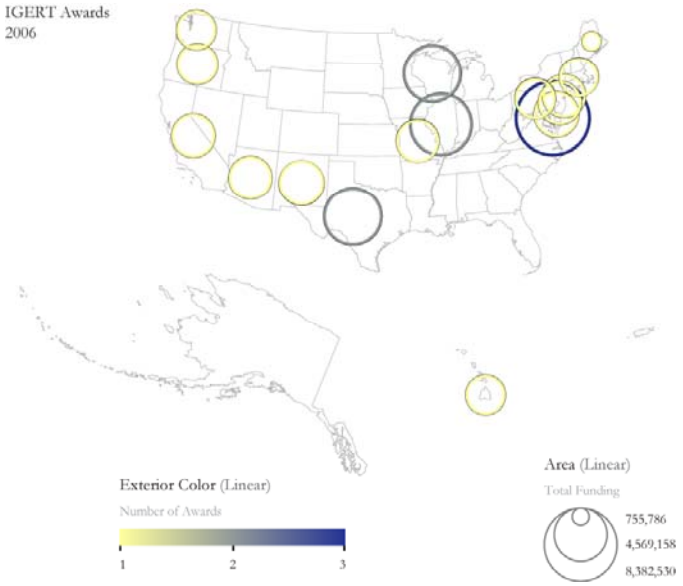
IGERT award data was aggregated by state and overlaid on geomap.

Circles are size coded by total dollar amount and colored by # awards made.

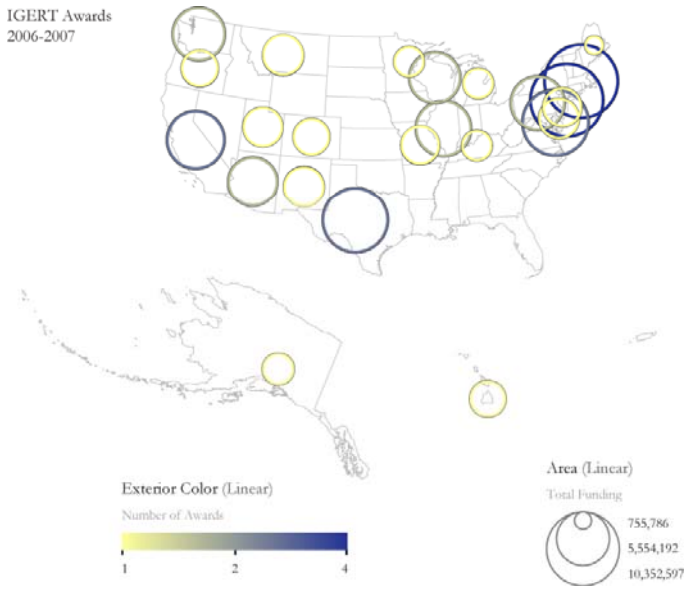


2006

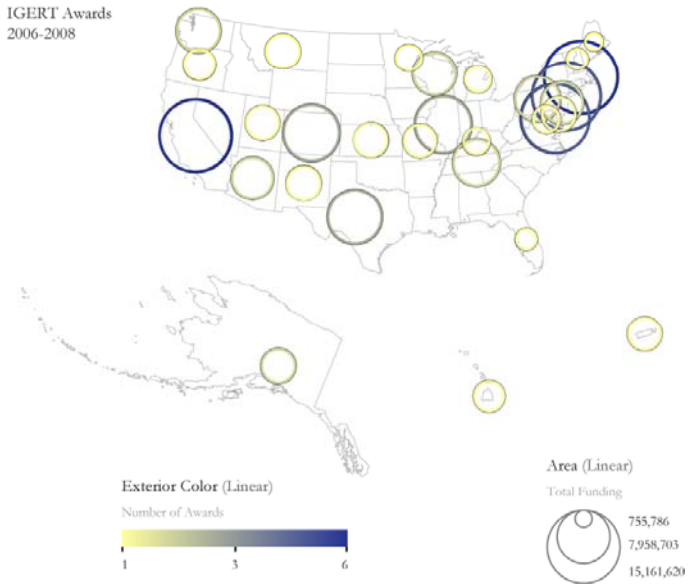
Geospatial Visualizations (IGERT 2006)



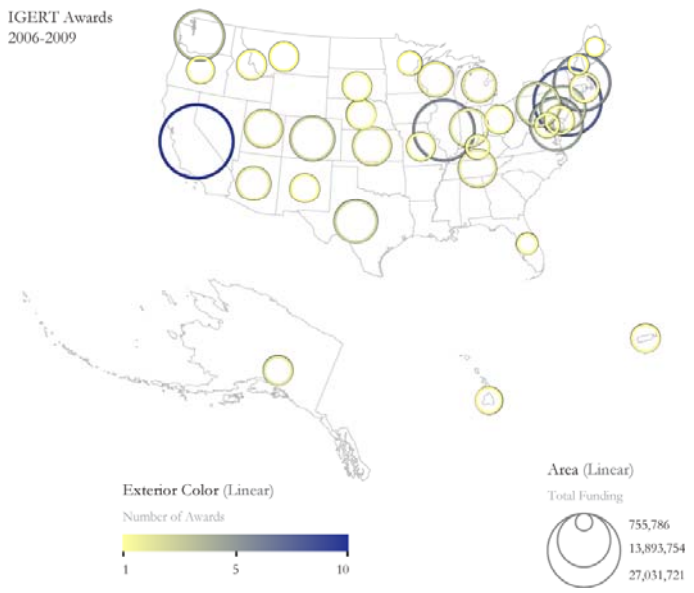
Geospatial Visualizations (IGERT 2006-2007)



Geospatial Visualizations (IGERT 2006-2008)



Geospatial Visualizations (IGERT 2006-2009)



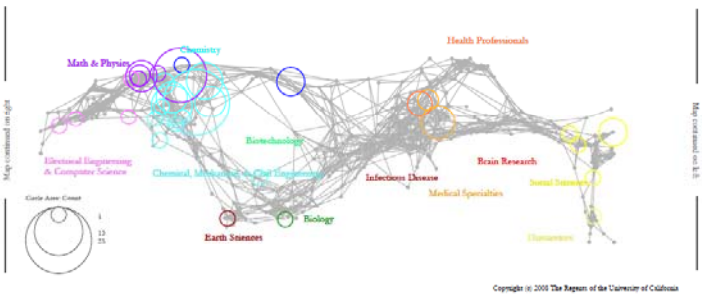
Note that circle sizes differ across time slices.

Topical Analysis (What)

UCSD Map of Science overlay of 118 MRI awards

The UCSD Map of Science was generated based on 7.2 million papers published in over 16,000 separate journals, proceedings, and series from Thomson Scientific and Scopus over the five year period from 2001 to 2005. Papers and journals were grouped into **554 clusters** of highly related journals. The **links** between the clusters show that some clusters are related to other clusters but are not as tightly connected as the journals that make up each cluster.

Each cluster is labeled both by the content area shared by the journals in the cluster and by the colored overarching scientific domain for that cluster. Data is science located by matching journal names or keywords associated with each of the 554 clusters.



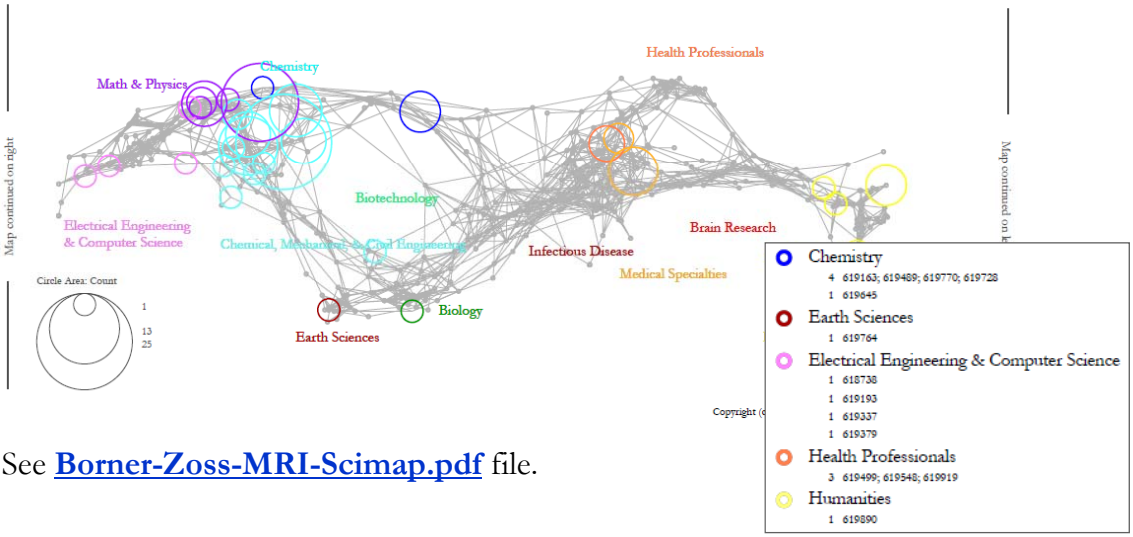
Topical Analysis (What)

UCSD Map of Science overlay of 118 MRI awards

Newman's terms were matched to keywords associated with the 554 clusters.

116 out of 118 records located.

These 116 records are associated with 10 of 13 disciplines of science and 34 of 554 research specialties in the UCSD Map of Science.



See **Borner-Zoss-MRI-Scimap.pdf** file.

Network Analysis (With Whom?)

Co-PI Network of all 51,217 NSF award records

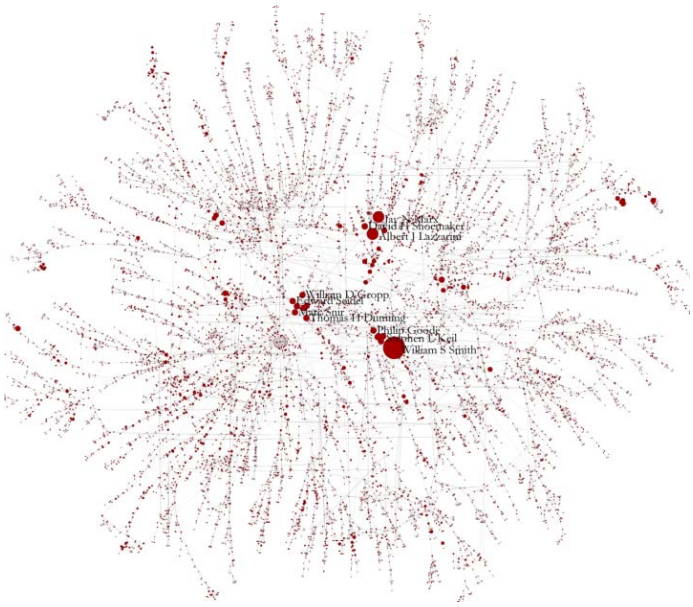
The co-PI network for all 51,217 NSF award records was extracted.

There are about 50,000 unique investigator names grouped in over 20,000 components (unconnected networks).

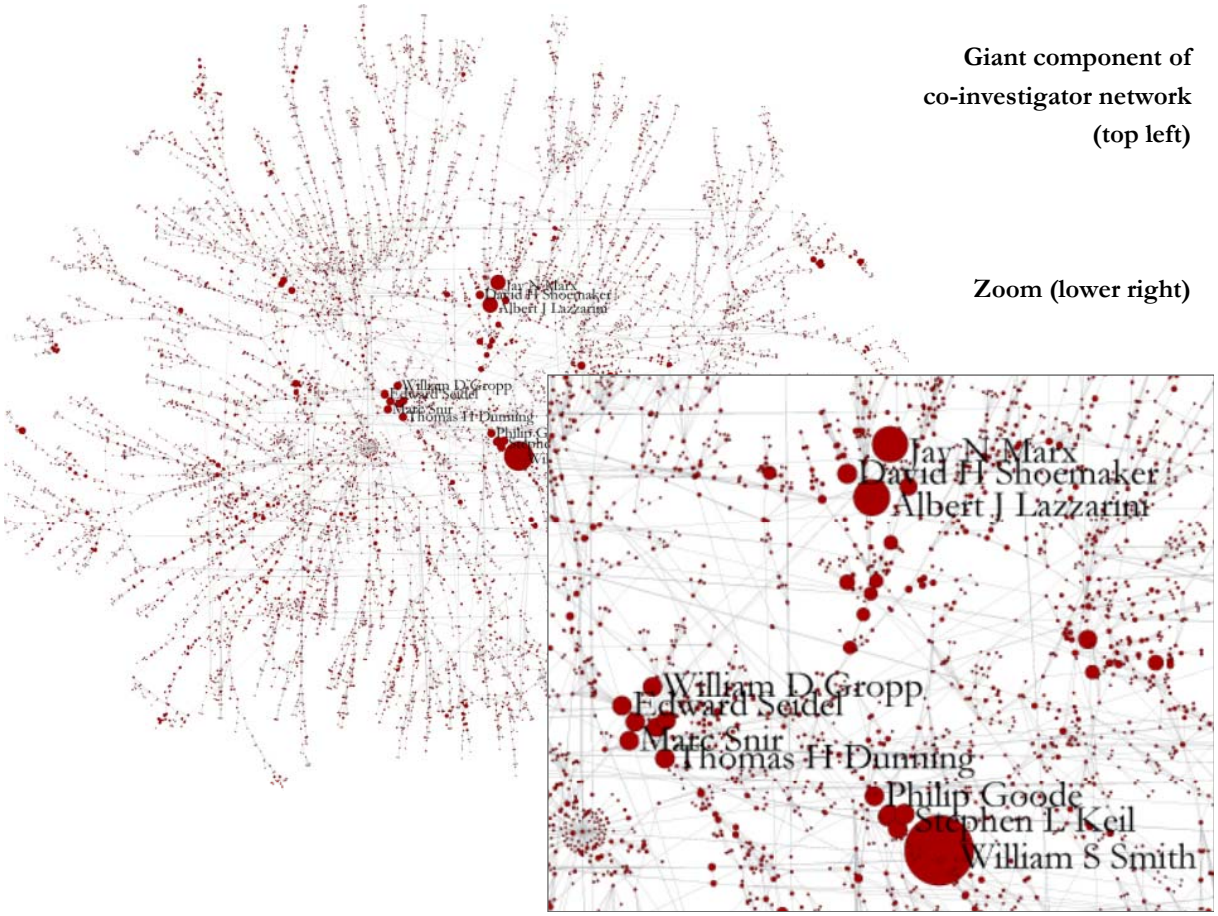
The largest (giant) component has more than 10,000 investigators and is shown here.

Nodes were size coded by funding amount—equally split among the investigators. That is, if there were 5 PIs on an award, each received 1/5 of the amount of the total award amount.

Investigator nodes are connected if they co-occurred on one award.



Disclaimer: Four years are very little time to grow networks.



Giant component of co-investigator network (top left)

Zoom (lower right)

Topic Analysis

using data provided by David Newman

Topic Data provided by David Newman, UCI on May 15, 2010

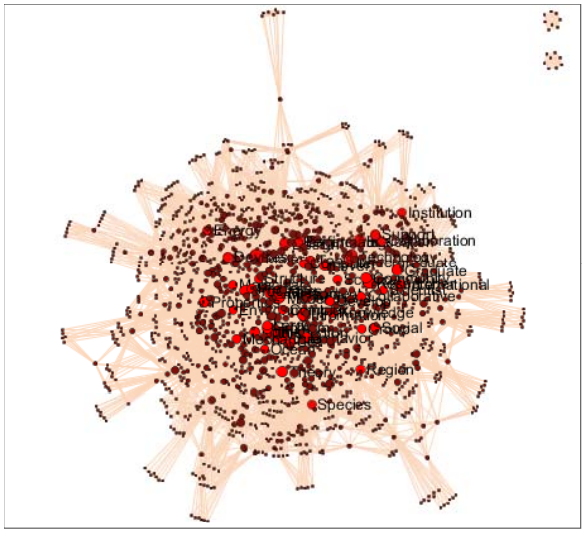
http://www.ics.uci.edu/~newman/nsf/may15/

Topic model (also known as Latent Dirichlet Allocation) is an unsupervised statistical algorithm for modeling document collections.

Procedure:

1. A bag-of-words representation was created from the 51k awards from the three xls files on the Sharepoint site by taking all the text in the title and abstract. Simple tokenization was performed, and stop words and infrequent terms were deleted.
2. A topic model was learned using T=400 topics (Gibbs sampler run for 800 iterations using const symmetric Dirichlet priors of beta=0.01 and alpha=0.05*N/(D*T))
3. Two results files were produced:

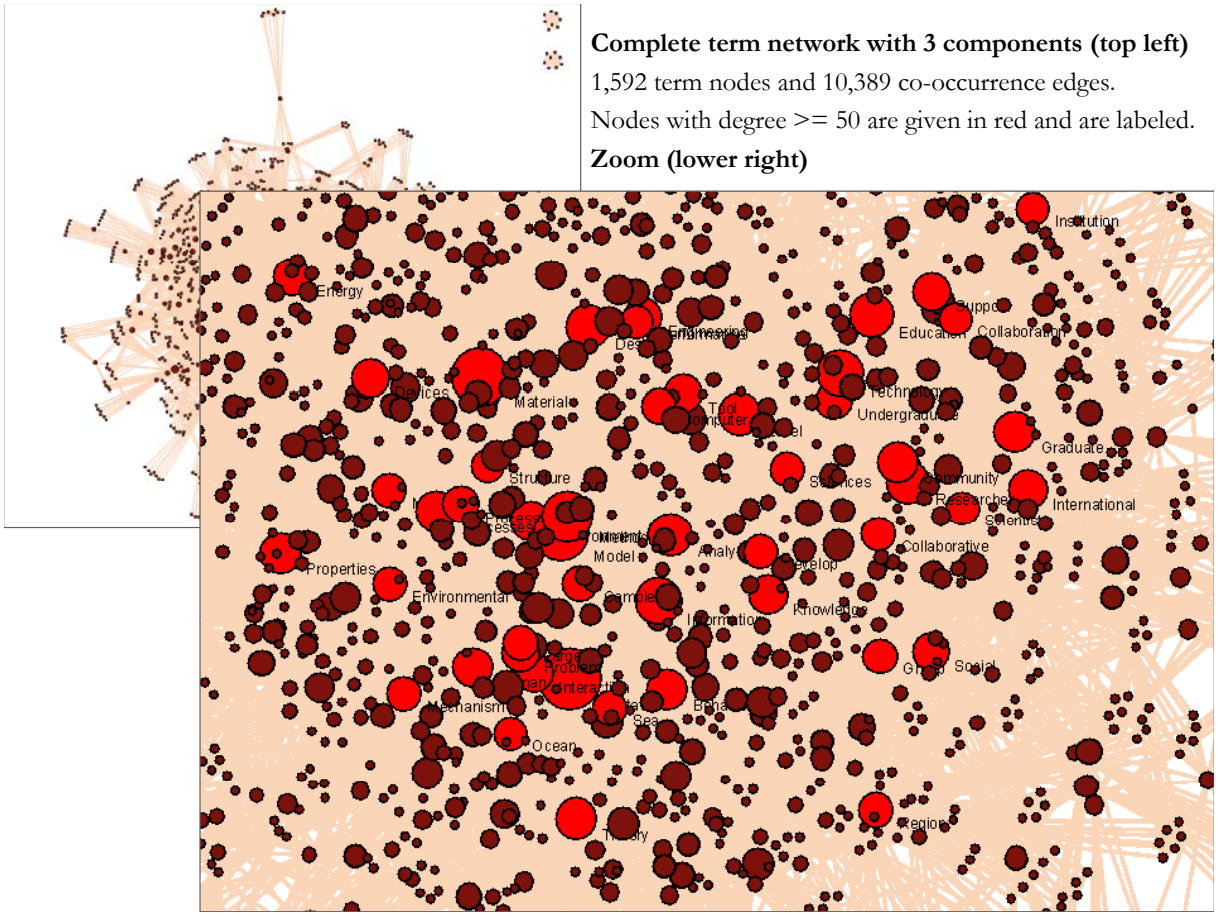
- prelim.topics.newman.txt (400 lines) contains a list of top-8 words in each of 400 topics learned by topic model.
- prelim.nsfid.top.topics.newman.txt comprises a list of up to top-four topics in each award abstract. Topics accounting for < 10% of document were suppressed. There are only 50,608 lines as some awards had had no clear topic tags.



Complete term network with 3 components (top left)

1,592 term nodes and 10,389 co-occurrence edges.

Nodes with degree >= 50 are given in red and are labeled.



Complete term network with 3 components (top left)

1,592 term nodes and 10,389 co-occurrence edges.

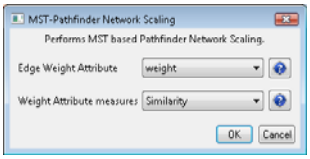
Nodes with degree >= 50 are given in red and are labeled.

Zoom (lower right)

Reducing edges via

MST-Pathfinder Network Scaling

Input Parameters:



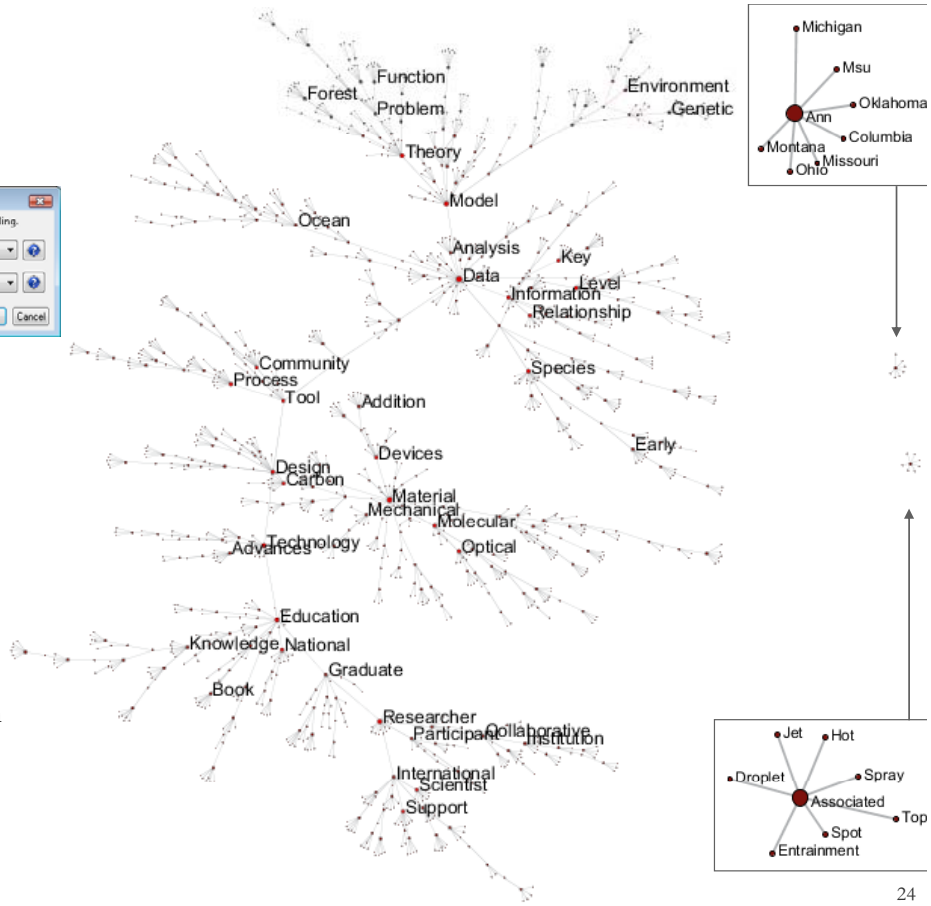
Result:

1,592 nodes and 8,560 edges.

Three components are shown in middle.

Nodes with degree > 7 are red and labeled.

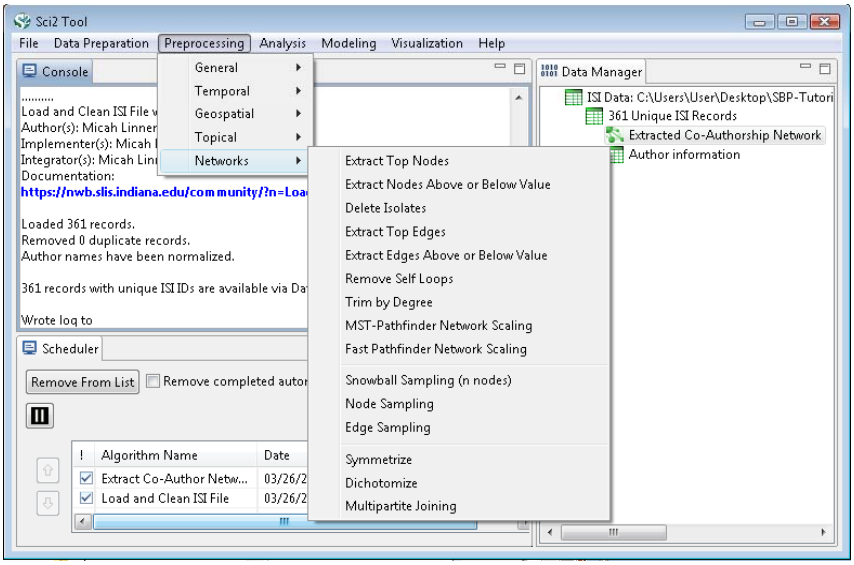
Zoom into small components are shown to the right.



Alternative Analyses

that require additional data

Sci² Tool for Science of Science Research and Practice



Acknowledgments

This work is supported in part by the Cyberinfrastructure for Network Science center and the School of Library and Information Science at Indiana University, the National Science Foundation under Grant No. SBE-0738111 and IIS-0513650, and the James S. McDonnell Foundation.

INDIANA UNIVERSITY SCHOOL OF LIBRARY AND INFORMATION SCIENCE

James S. McDonnell Foundation

Sci² Tool: Algorithms

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

Preprocessing
Extract Top N% Records
Extract Top N Records
Normalize Text
Slice Table by Line

Extract Top Nodes
Extract Nodes Above or Below Value
Delete Isolates

Extract top Edges
Extract Edges Above or Below Value
Remove Self Loops
Trim by Degree
MST-Pathfinder Network Scaling
Fast Pathfinder Network Scaling

Snowball Sampling (in nodes)
Node Sampling
Edge Sampling

Symmetrize
Dichotomize
Multipartite Joining

Geocoder

Extract ZIP Code

Modeling
Random Graph
Watts-Strogatz
Small World
Barabási-Albert Scale-Free
TARL

Analysis
Network Analysis Toolkit (NAT)
Unweighted & Undirected
- Node Degree
- Degree Distribution
- ---
- K-Nearest Neighbor (Java)
- Watts-Strogatz Clustering Coefficient
- Watts Strogatz Clustering Coefficient over K
- ---
- Diameter
- Average Shortest Path
- Shortest Path Distribution
- Node Betweenness Centrality
- ---
- Weak Component Clustering
- Global Connected Components
- ---
- Extract K-Core
- Annotate K-Coreness
- ---
- HITS

Weighted & Undirected
- Clustering Coefficient
- Nearest Neighbor Degree
- Strength vs Degree
- Degree & Strength
- Average Weight vs End-point Degree
- Strength Distribution
- Weight Distribution
- Randomize Weights
- ---
- Blondel Community Detection
- ---
- HITS

Unweighted & Directed
- Node Indegree
- Node Outdegree
- Indegree Distribution
- Outdegree Distribution
- ---
- K-Nearest Neighbor
- Single Node in-Out Degree Correlations
- ---
- Dyad Reciprocity
- Arc Reciprocity
- Adjacency Transitivity
- ---
- Weak Component Clustering
- Strong Component Clustering
- ---

Sci² Tool: Algorithms cont.

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

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

Weighted & Directed
- HITS
- Weighted PageRank

Textual
Burst Detection

Visualization
GnuPlot
GUESS
Image Viewer

Radial Tree/Graph (prefuse alpha)
Radial Tree/Graph with Annotation (prefuse beta)
Tree View (prefuse beta)
Tree Map (prefuse beta)
Force Directed with Annotation (prefuse beta)
Fruchterman-Reingold with Annotation (prefuse beta)

DrL (VxOrd)
Specified (prefuse beta)

Horizontal Line Graph
Circular Hierarchy
Geo Map (Circle Annotation Style)
Geo Map (Colored-Region Annotation Style)
***Science Map (Circle Annotation)**

Scientometrics
Remove ISI Duplicate Records
Remove Rows with Multitudinous Fields
Detect Duplicate Nodes
Update Network by Merging Nodes

Extract Directed Network
Extract Paper Citation Network
Extract Author Paper Network

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

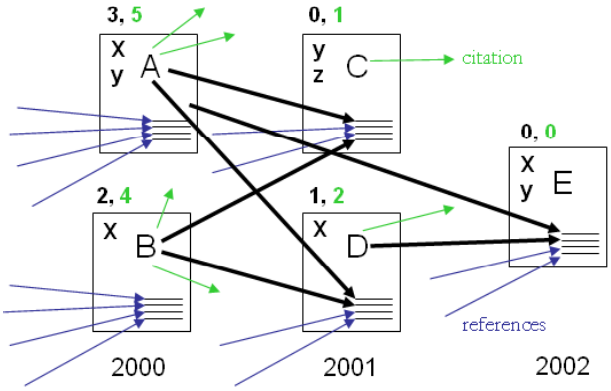
* Requires permission from UCSD
All four+ save into Postscript files.

General Network extraction

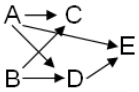
Network Extraction

Sample paper network (left) and four different network types derived from it (right). From ISI files, about 30 different networks can be extracted.

Papers A-E written by authors x, y, z over 3 years.
Each paper happens to have 4 references.



Paper-Paper Citation Network
Papers are connected via direct citation links. Arrows represent information flow from older papers to younger papers.



Author-Author (Co-Author) Network
x and y co-author papers A and E together
y and z co-author papers A and E

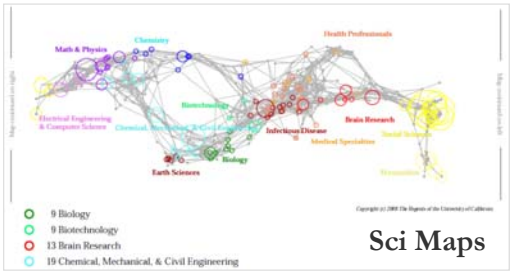
Document Co-Citation (DCA) Network
A and B are co-cited by C and D
A and D are co-cited by E

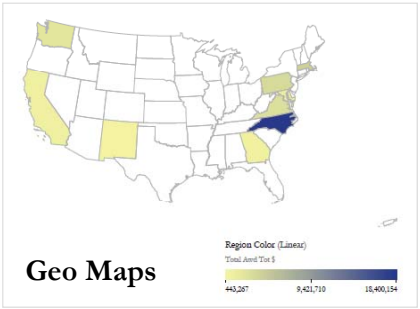
Reference Co-Occurrence (Bibliographic Coupling) Network
C and D are bibliographically coupled as they both cite/reference A and B.

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

Sci² Tool

Plugins that render into Postscript files:





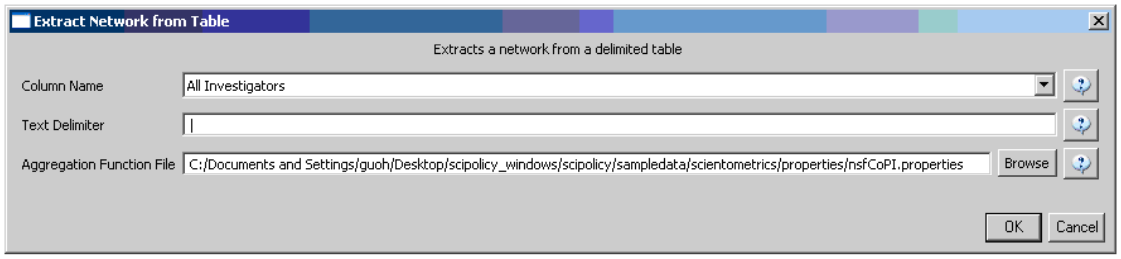
Börner, Katy, Huang, Weixia (Bonnie), Linnemeier, Micah, Duhon, Russell Jackson, Phillips, Patrick, Ma, Nianli, Zoss, Angela, Guo, Hanning & Price, Mark. (2009). Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Scholarly Database and the Network Workbench Tool. Proceedings of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

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

Funding Data Analysis

Free online services such as NSF's Award Search (See Section 4.2.2.1 NSF Award Search) support the retrieval of ego-centric funding profiles. Here, a search was exemplarily conducted for "Katy Borner" in the "Principal Investigator" field while keeping the "Include CO-PI" box checked.

The resulting data is available at '**yoursci2directory*/sampledata/scientometrics/nsf/KatyBorner.nsf*.' Load the data using '*File > Load*', select the loaded dataset in the Data Manager window, and run *'Data Preparation > Text Files > Extract Co-Occurrence Network'* using these parameters:



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

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

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

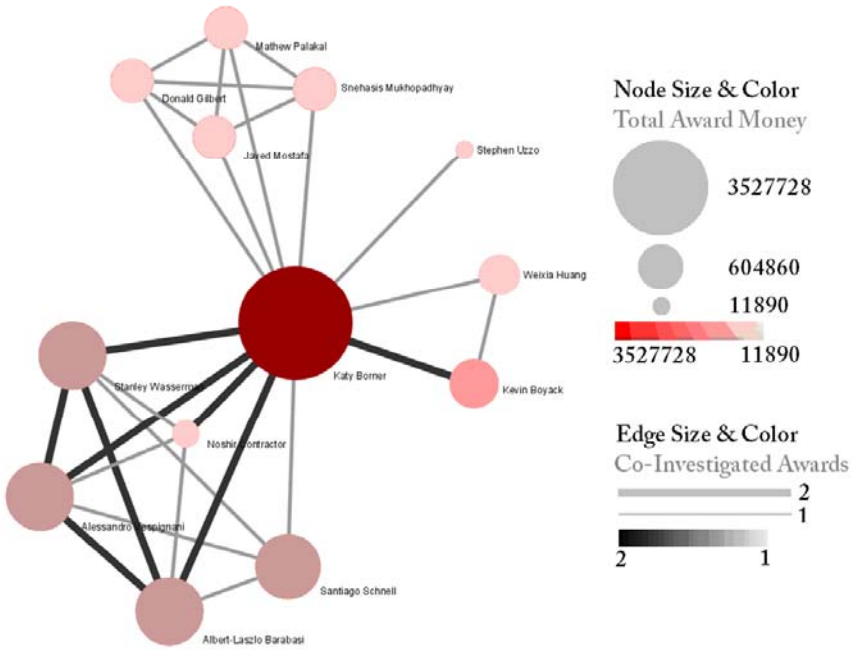
1. Resize Linear > Nodes > totalawardmoney > From: 5 To: 35 > Do Resize Linear
2. Resize Linear > Edges > coinvestigatedawards From: 1 To: 2 > Do Resize Linear
3. Colorize > Nodes > totalawardmoney From : To: > Do Colorize
4. Colorize > Edges > coinvestigatedawards From: To: > Do Colorize
5. Object: all nodes > Show Label
6. Type in Interpreter:

```
>for n in g.nodes:
…         n.strokecolor = n.color
```

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

Co-PI Network

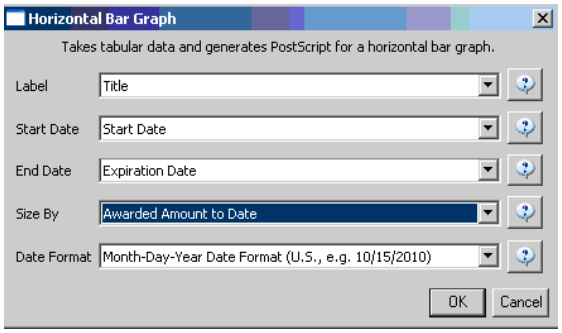
This is a so called ego-centric network, i.e., almost complete data is available and shown for exactly one ego. The funding records for all other people in the network are most likely incomplete.



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

Award Durations and Totals

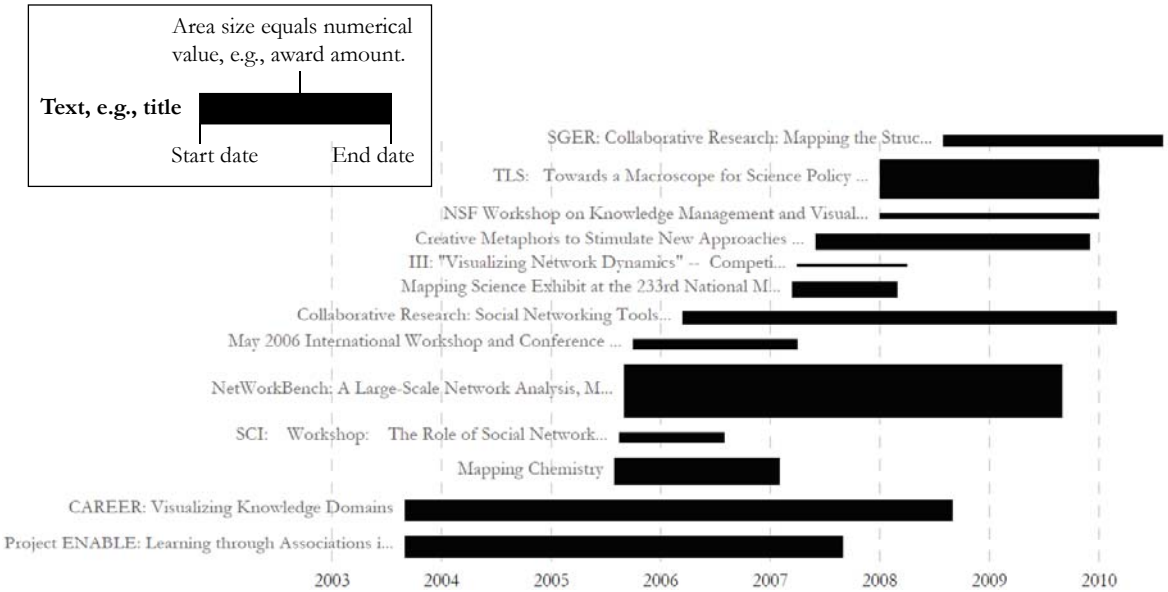
For a summary of the grants themselves, with a visual representation of their award amount, select the NSF csv file in the Data Manager and run '*Visualization > Temporal > Horizontal Bar Graph*', entering the following parameters:



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

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

Award Durations and Totals

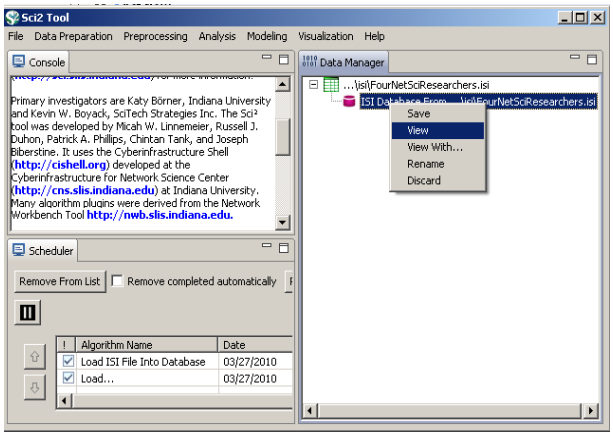


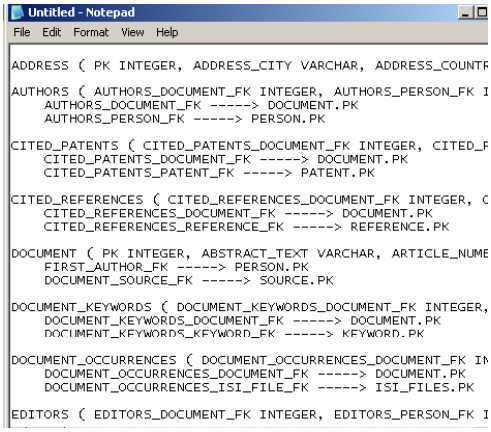
Studying Four Major NetSci Researchers (ISI Data) using Database *(section 5.1.5)*

Replicate Studies Using Database Support

Load *'*yoursci2directory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi'*, using '*File > Load*' instead of '*File > Load and Clean ISI File*'.

Run *'File > Load Into Database > Load ISI File Into Database'*. View the database schema by right-clicking on the loaded database in the Data Manager and clicking "View".

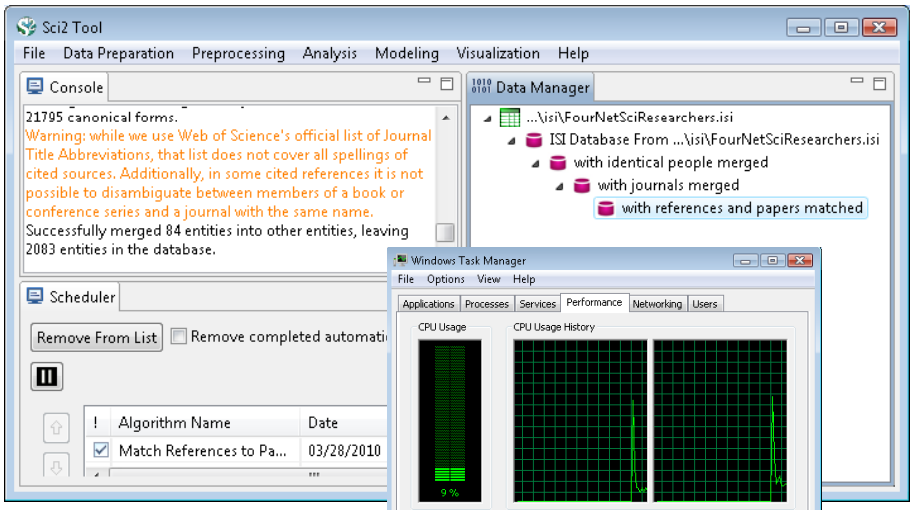




Studying Four Major NetSci Researchers (ISI Data) using Database *(section 5.1.5)*

Replicate Studies Using Database Support – Unification

Run '*Data Preparation > Database > ISI > Merge Identical ISI People*', followed by '*Data Preparation > Database > ISI > Merge Journals'* and '*Data Preparation > Database > ISI > Match References to Papers*'. Make sure to wait until each cleaning step is complete before beginning the next one. Read red warnings.



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

Using Database Support – Extract Basic Properties

Run '*Data Preparation > Database > ISI > Extract Authors*' and right-click on the resulting table to view all the authors from FourNetSciResearchers.isi. The table also has columns with information on how many papers each person in the dataset authored, their Global Citation Count (how many times they have been cited according to ISI), and their Local Citation Count (how many times they were cited in the current dataset).

	A	B	C	D	E	F	G	H	I	J	K
1	UNSPLIT_NAME	PAPERS_	GLOBAL_	LOCAL_CI	ADDITION	FAMILY_NAME	FIRST_INI	FULL_NAM	MIDDLE_I	PERSONAL_NAME	
2	Barthelemy, M	9	454	12		Barthelemy	M				
3	Barrat, A	13	480	14		Barrat	A				
4	Pastor-satorras, R	24	1769	48		Pastor-satorras	R				
5	Vespignani, A	101	3811	213		Vespignani	A				
6	Wasserman, S	32	675	109		Wasserman	S				
7	Daruka, I	7	392	11		Daruka	I				
8	Makeev, MA	8	198	19		Makeev	M		A		
9	Sidoretti, S	1	1	1		Sidoretti	S				
10	Iacobucci, D	6	115	33		Iacobucci	D				
11	Vazquez, A	10	620	5		Vazquez	A				
12	Oliveira, JG	2	20	0		Oliveira	J		G		
13	Farkas, I	3	47	1		Farkas	I				
14	Jeong, H	17	4160	143		Jeong	H				
15	Oltvai, ZN	17	2961	59		Oltvai	Z		N		
16	Cuerno, R	2	267	11		Cuerno	R				
17	Dobrin, R	2	85	2		Dobrin	R				
18	Beg, QK	1	41	0		Beg	Q		K		
19	Pudovkin, AI	5	32	6		Pudovkin	A		I		

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

Using Database Support – Records over time

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

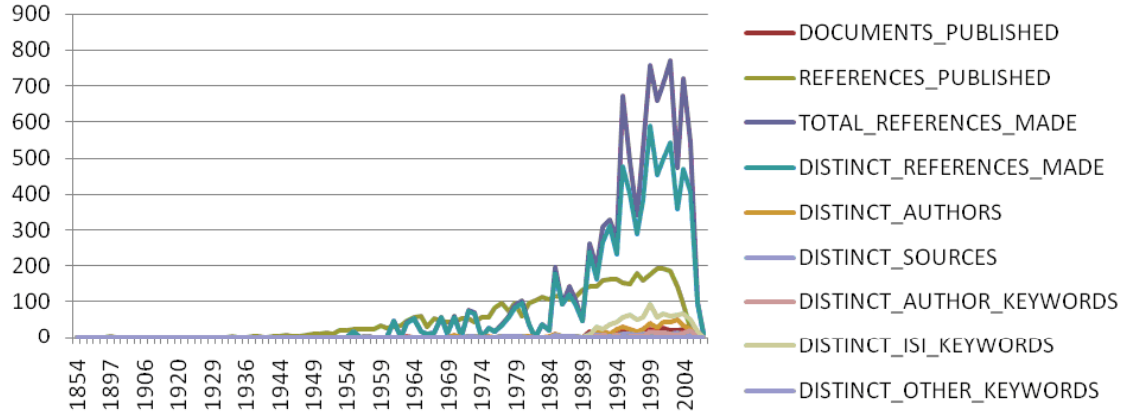
F1 DISTINCT_AUTHORS

	A	B	C	D	E	F	G	H	I	J	K	L
1	YR	DOCUMENTS_	REFERENCES_	TOTAL_REFERI	DISTINCT_REF	DISTINCT_AU	DISTINCT_	DISTINCT_	DISTINCT_	DISTINCT_	OTHER_KEYWOR	
83	1995	19	153	672	477	32	9	0	57	0		
84	1996	14	148	490	401	23	9	3	62	0		
85	1997	13	179	343	289	16	6	4	49	0		
86	1998	19	159	527	383	23	9	4	57	0		
87	1999	24	176	757	590	39	11	18	94	0		
88	2000	19	191	660	455	28	9	13	57	0		
89	2001	28	192	706	497	44	13	13	68	0		
90	2002	21	186	770	542	44	11	12	61	0		
91	2003	21	144	474	358	51	15	8	62	0		
92	2004	23	94	723	471	34	12	14	68	0		
93	2005	20	24	542	406	25	13	20	49	0		
94	2006	3	1	100	94	9	3	3	17	0		
95	2007	1	0	12	12	1	1	1	2	0		

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

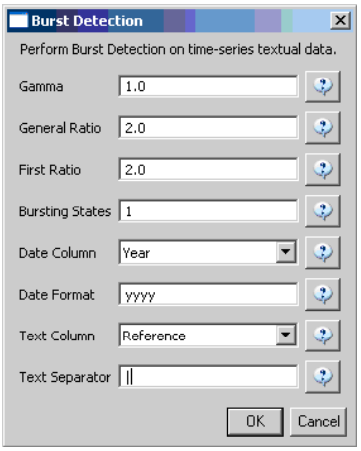
Using Database Support – Records over time

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

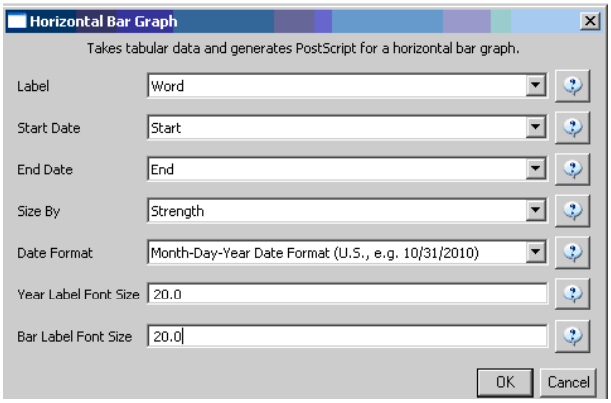


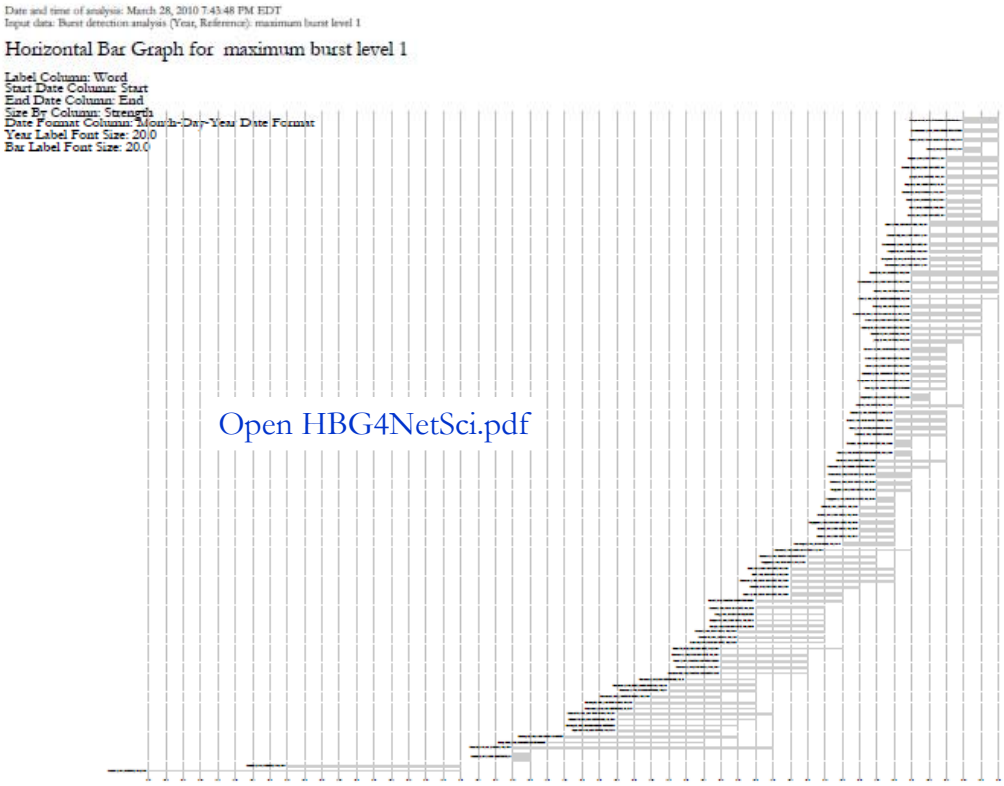
Using Database Support – Burst Analysis for References

The queries can also output data specifically tailored for the burst detection algorithm (see Section 4.6.1 Burst Detection). Run '*Data Preparation* > *Database* > *ISI* > *Extract Authors* > *Extract References by Year for Burst Detection*' on the cleaned database followed by '*Analysis* > *Topical* > *Burst Detection*' with parameters on left and then run '*Visualize* > *Temporal* > *Horizontal Bar Graph*' with parameters on right.



Watch those red warnings!



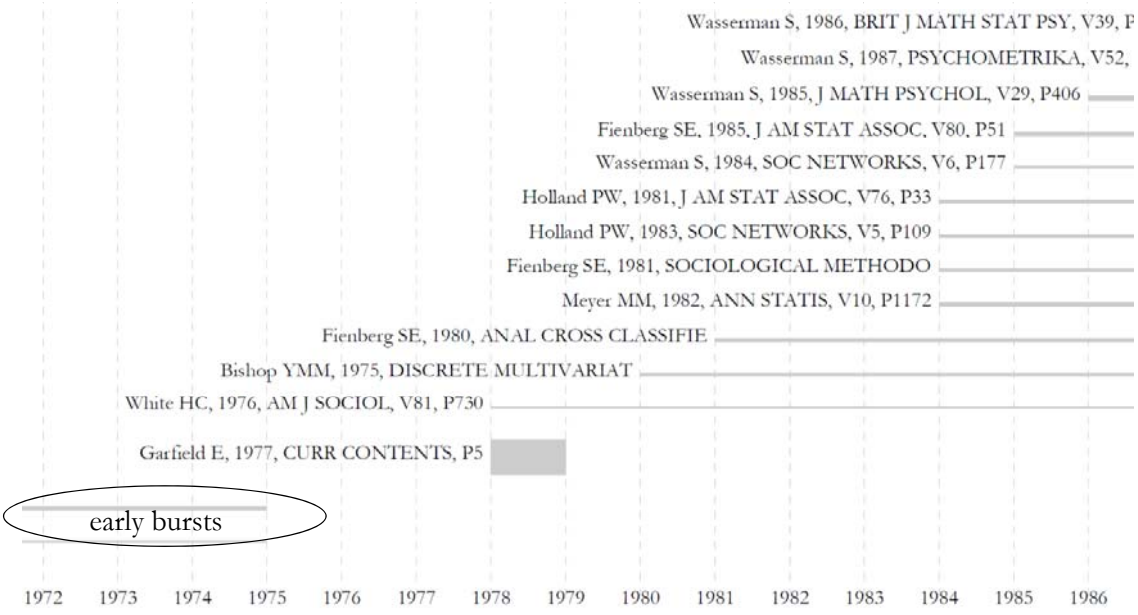


Open HBG4NetSci.pdf

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

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

Using Database Support – Burst Analysis Result

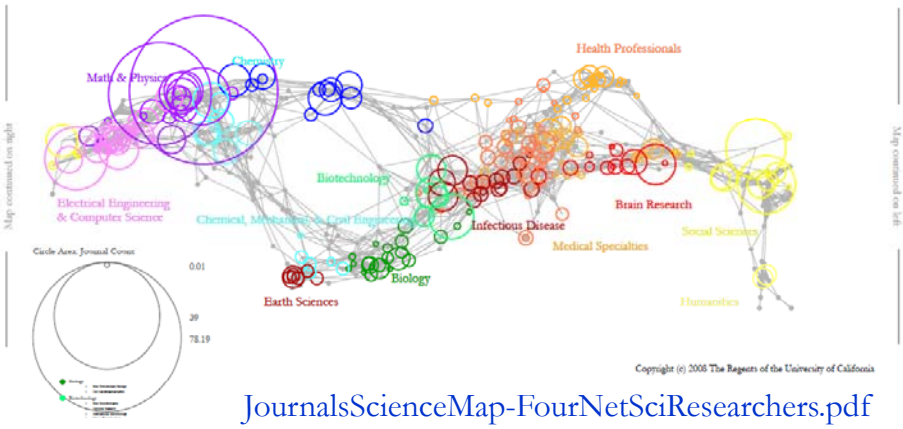


Topic Mapping: UCSD Science Map

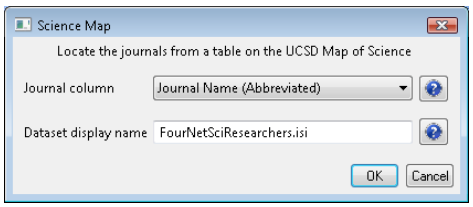
Science Map via Journals for FourNetSciResearchers.isi

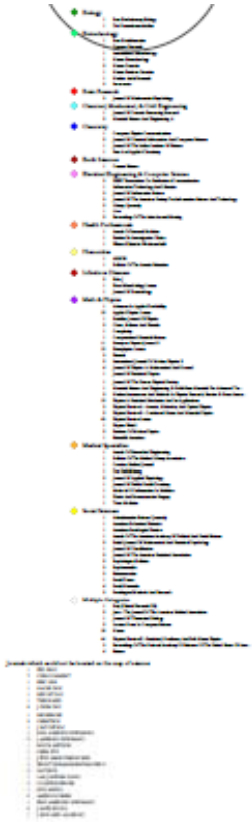
314 journal references matched out of 361 found.

These 314 references are associated with 13 of 13 disciplines of science and 255 of 554 research specialties in the UCSD Map of Science.



JournalsScienceMap-FourNetSciResearchers.pdf





Biomedical Funding Profile of NSF (NSF Data)

(section 5.2.4)

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

What organizations and programs at the National Science Foundation support projects that deal with medical and health related topics? Data was downloaded from the NSF Awards Search SIRE (http://www.nsf.gov/awardsearch) on Nov 23rd, 2009, using the query "medical AND health" in the title, abstract, and awards field, with "Active awards only" checked (see section 4.2.2.1 NSF Award Search for data retrieval details).

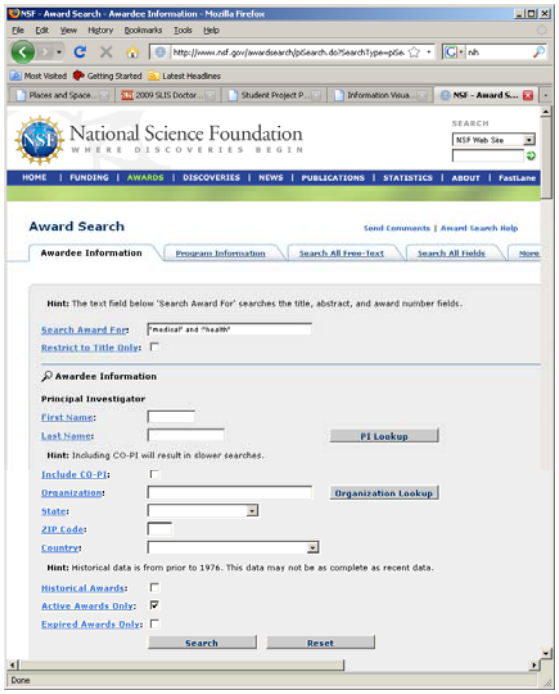
Biomedical Funding Profile of NSF (NSF Data)

(section 5.2.4)

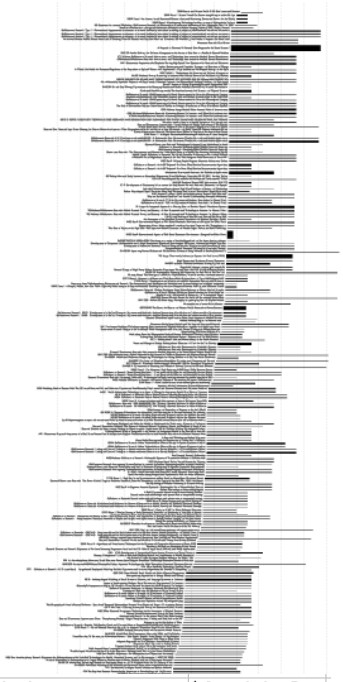
Using NSF Awards Search:
http://www.nsf.gov/awardsearch
download relevant NSF awards that have "medical" AND "health" in title, abstract, and awards. Active awards only.

Number of awards: 283 awards
Total awarded amount to date: $152,015,288

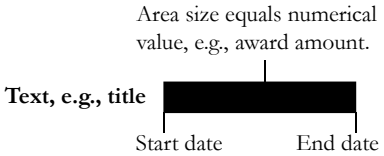
Retrieved on Oct 18, 2009



Biomedical Funding Profile
(section 5.2.4)



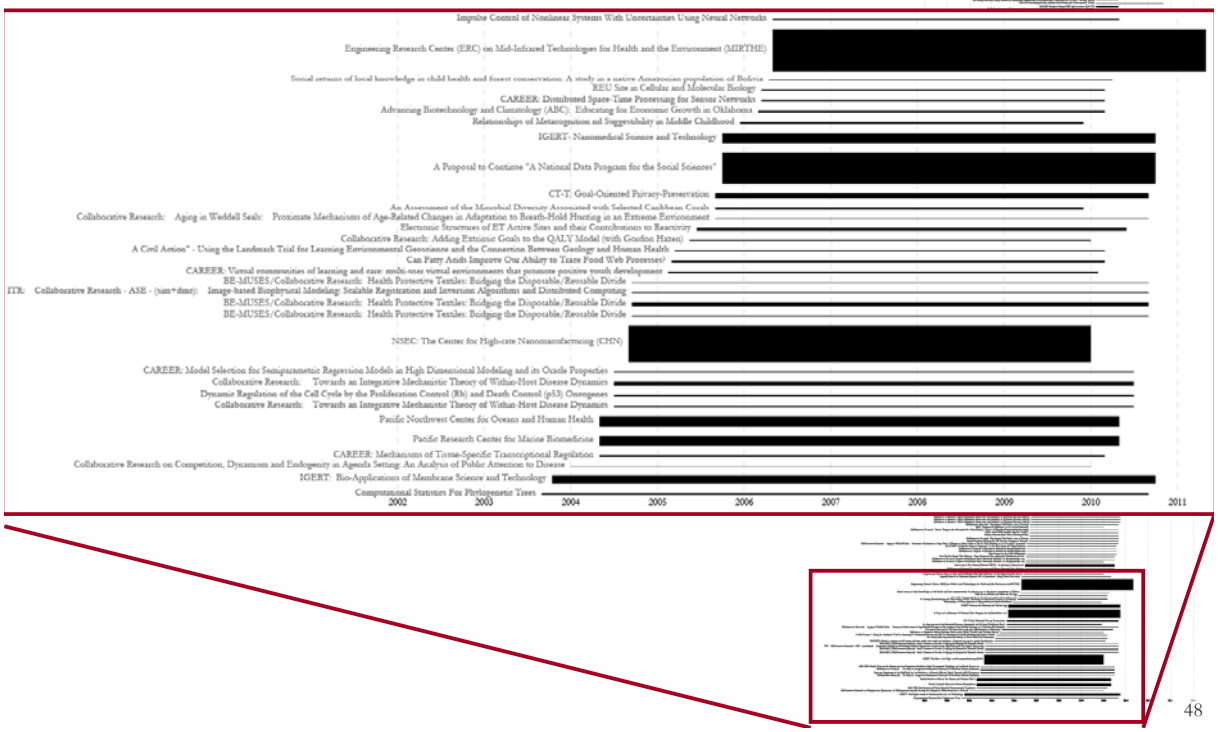
Horizontal Bargraph



Top-10 grants with highest $Awarded to Date:

Title	NSF Org.	Program(s)	PI	State	Organization	$ Awarded to Date
University of NewMexico/Harvard PREM: Leadership in Bi	DMR	PREM|MATERIAL	Lopez, Gabriel	NM	University of New Mexico	2,037,500
TC: Large: Trustworthy Information Systems for Healthcare	CNS	TRUSTWORTHY C	Kotz, David	NH	Dartmouth College	2,999,999
IGERT- Nanomedical Science and Technology	DGE	||||||IGERT FULL PI	Sridhar, Srinivas	MA	Northeastern University	3,323,891
IGERT: Bio-Applications of Membrane Science and Techr	DGE	HUMAN RESOURI	Fried, Joel	OH	University of Cincinnati Main Campu	3,644,410
Pacific Research Center for Marine Biomedicine	OCE	CHEMICAL OCEA	Laws, Edward	HI	University of Hawaii	3,816,943
Pacific Northwest Center for Oceans and Human Health	OCE	CHEMICAL OCEA	Faustman, Elaine	WA	University of Washington	4,026,968
A Proposal to Continue 'A National Data Program for the S	SES	||SCIENCE & ENG	Smith, Tom	IL	National Opinion Research Center	5,835,140
A Proposal to Continue "A National Data Program for the S	SES	||||SCIENCE & EN(	Davis, James	IL	National Opinion Research Center	10,053,668
NSEC: The Center for High-rate Nanomanufacturing (CHN)	EEC	Studies of Policy S	Busnaina, Ahmed	MA	Northeastern University	13,047,758
Engineering Research Center (ERC) on Mid-Infrared Techn	EEC	COLLABORATIVE	Gmachl, Claire	NJ	Princeton University	13,681,994

Biomedical Funding Profile
(section 5.2.4)



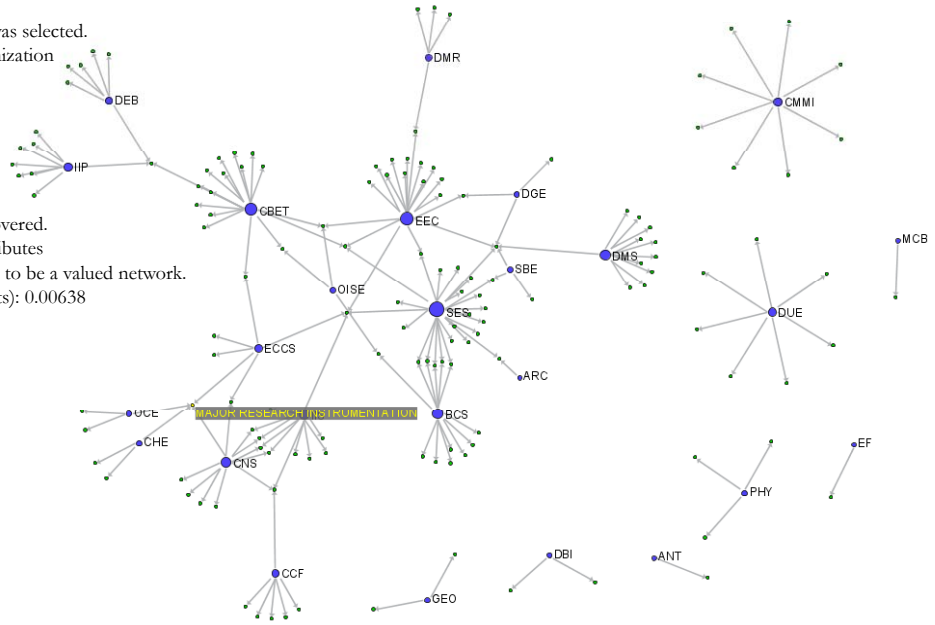
Biomedical Funding Profile of NSF (NSF Data)

(section 5.2.4)

Bimodal Network of NSF Organization to Program(s)

Extract Directed Network was selected.
Source Column: NSF Organization
Text Delimiter: |
Target Column: Program(s)

Nodes: 167
Isolated nodes: 0
Edges: 177
No parallel edges were discovered.
Did not detect any edge attributes
This network does not seem to be a valued network.
Density (disregarding weights): 0.00638



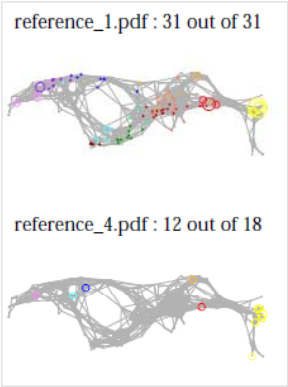
Measuring (Input/Output) Interdisciplinarity

Can be <u>measured</u> based on

- Title, abstract or full proposal text using **simple text analysis or linguistic techniques.** Difficult across disciplinary boundaries as writing styles and word usage are very different.
- Proposal references (cited base knowledge) using **RefMapper.**
- Keywords provided by investigators, e.g., for CDI proposals.
- Background/departments/publications /prior funding of PI/Co-PIs—requires unique people IDs and resume like information. VIVO might help here (http://vivoweb.org).
- Publications and other results reported in NSF progress reports.

Can be <u>visualized</u> as

- Tables with cluster assignments.
- TopicMaps—visual groupings of awards that are similar.
- Science Map overlays, see below and next slide.



Reference Mapper

Duhon & Börner, forthcoming.

(a) Overview

Date and input directory

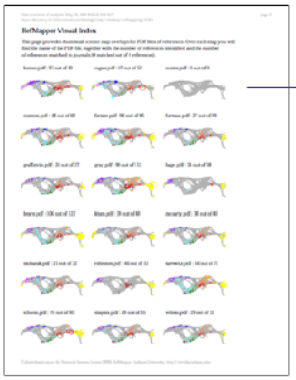
Basic counts

Overlay of all matched journal references from all PDF files on 554 scientific disciplines (nodes) in UCSD Map of Science

Circle size denotes # references

Listing of all references grouped by 13 science areas

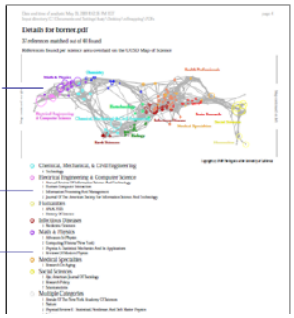
(b) Visual Index



For each PDF file:
Basic counts and thumbnail science map

Max 18 per page

(c) Details



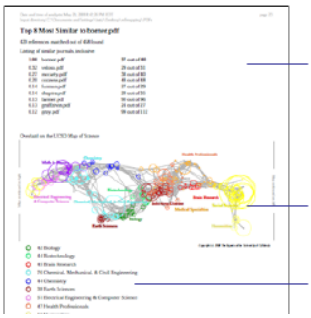
For each PDF file:
Overlay of all matched journal references on 554 scientific fields (nodes) in UCSD Map of Science

Circle size denotes # references

Colors and names of science areas that are cited

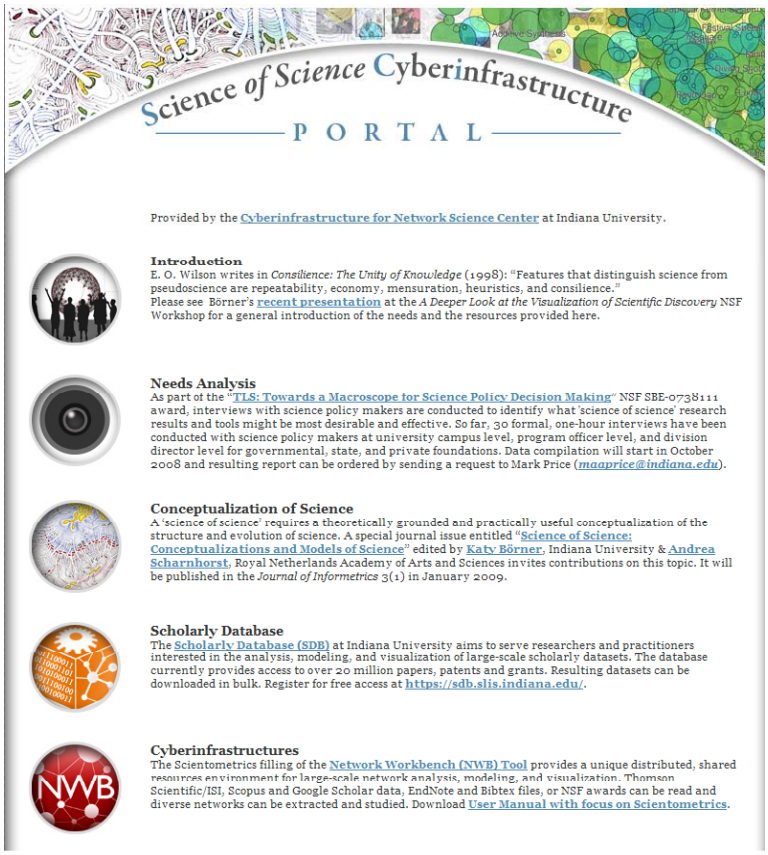
Alphabetic listing of cited journals and # of times cited

(d) Top-10 Most Similar



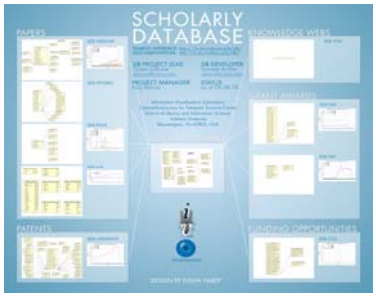
Top-n most similar PDF files identified based on journal name co-occurrences
The similarity of each PDF file to itself is 1

Overlay of matched journal references from all above listed PDF files on UCSD Map of Science and grouping by 13 science areas



http://sci.slis.indiana.edu

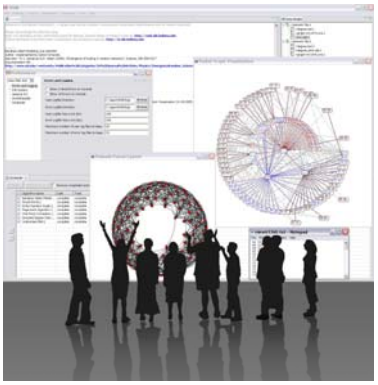
CIs Developed and Served by CNS



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

James S. McDonnell Foundation

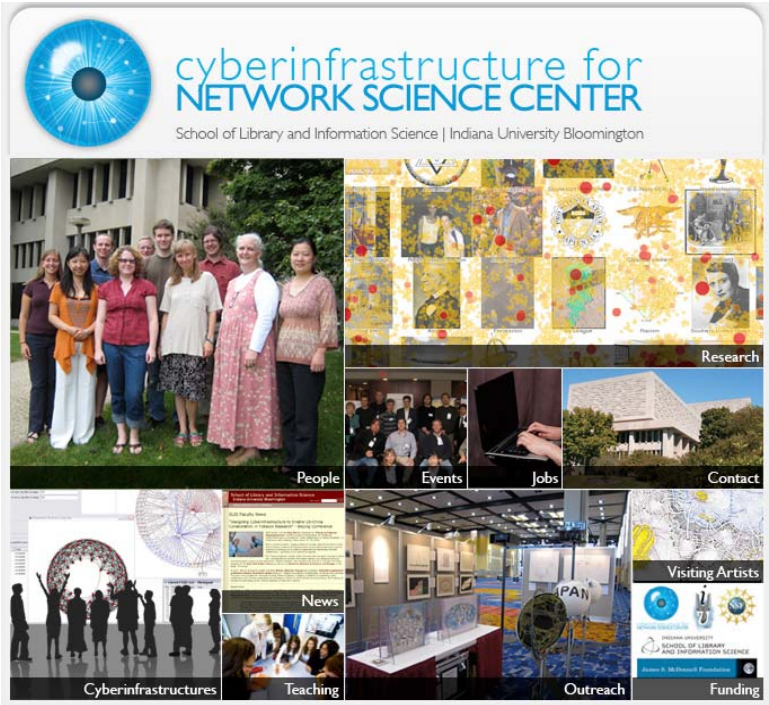
Information Visualization Cyberinfrastructure
http://iv.slis.indiana.edu



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

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

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



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