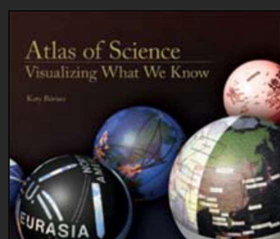
Mining, Mapping, and Accelerating Science and Technology

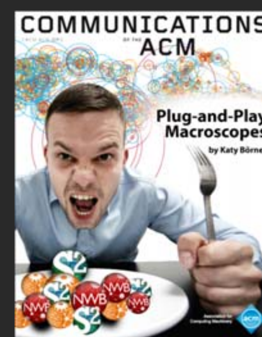
Katy Börner
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
katy@indiana.edu

With special thanks to the members at the Cyberinfrastructure for Network Science Center; the Sci2, NWB, and EpiC team; and the VIVO Collaboration

iSchool Faculty Lecture
University of Pittsburgh
Pittsburgh, PA

November 10, 2011



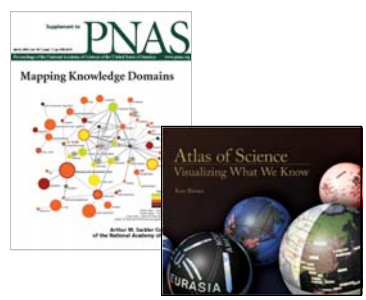


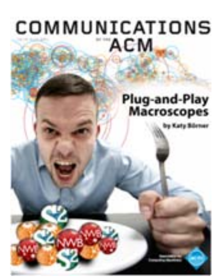
Overview

1. **Data mining and visualization research** that aims to increase our scientific understanding of the structure and dynamics of science and technology.

2. **Novel approaches and services** that improve information access, researcher networking, and research management.

3. **Data services and plug-and-play macroscope tools** that commoditize data mining and visualization.





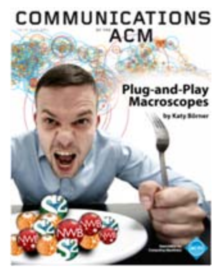
cyberinfrastructure for
NETWORK SCIENCE CENTER
School of Library and Information Science | Indiana University Bloomington

Overview

1. **Data mining and visualization research** that aims to increase our scientific understanding of the structure and dynamics of science and technology.

2. **Novel approaches and services** that improve information access, researcher networking, and research management.

3. **Data services and plug-and-play macroscope tools** that commoditize data mining and visualization.



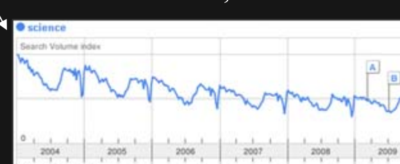
cyberinfrastructure for NETWORK SCIENCE CENTER
School of Library and Information Science | Indiana University Bloomington

Take terra bytes of data

Descriptive & Predictive Models

Find your way

Find collaborators, friends



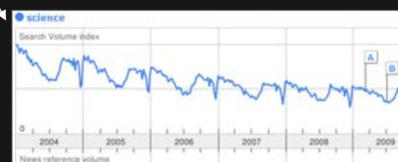
Identify trends

Take terra bytes of data

Plug-and-Play Macroscopes

Find your way

Find collaborators, friends



Identify trends

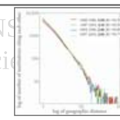
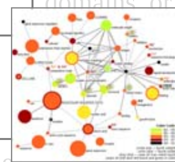
Type of Analysis vs. Level of Analysis

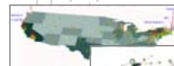
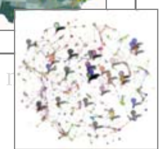
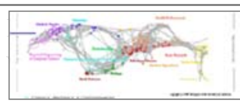
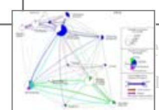
	Micro/Individual (1-100 records)	***Meso/Local (101–10,000 records)***	***Macro/Global (10,000 < records)***
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of USA, all of science.
Temporal Analysis (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	NIH's core competency

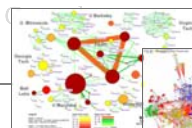
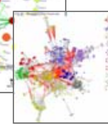
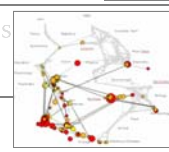
cyberinfrastructure for NETWORK SCIENCE CENTER
School of Library and Information Science | Indiana University Bloomington

Type of Analysis vs. Level of Analysis

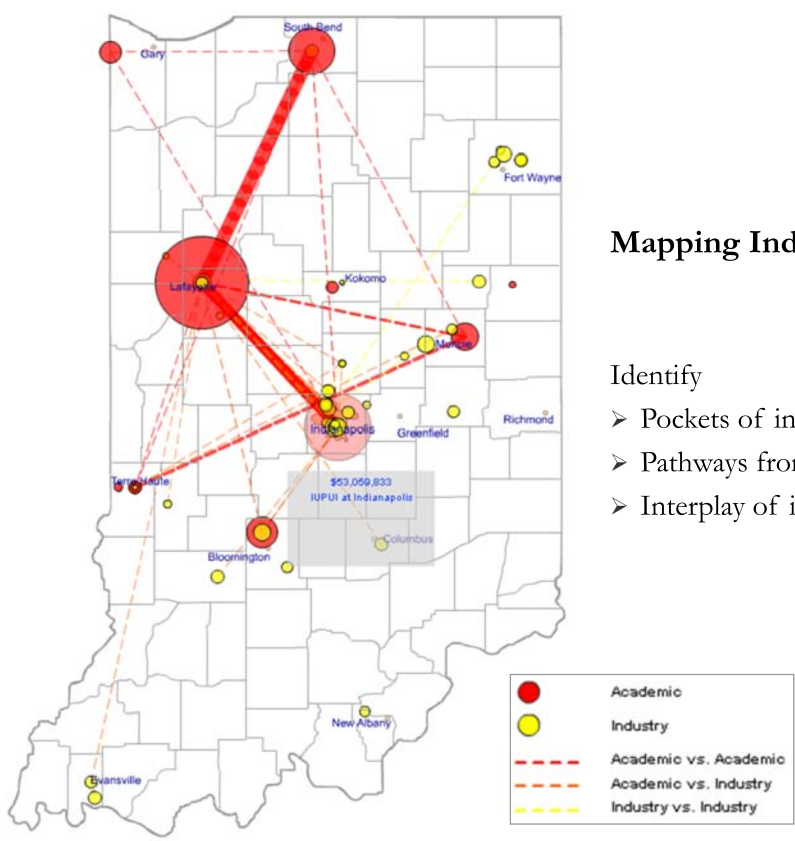
	Micro/Individual (1-100 records)	*Meso/Local (101–10,000 records)*	*Macro/Global (10,000 < records)*
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NS SA, all of sci
Temporal Analysis (When)	Funding portfolio of one individual	ic bursts f PNAS	113 Years of P Research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a st intellectual l	PNAS
Topical Analysis (What)	s.	research	VxOrd/Topic r NIH funding
Network Analysis (With Whom?)	NSF work of one	C k	NIH's cy

cyberinfrastructure for NETWORK SCIENCE CENTER
School of Library and Information Science | Indiana University Bloomington



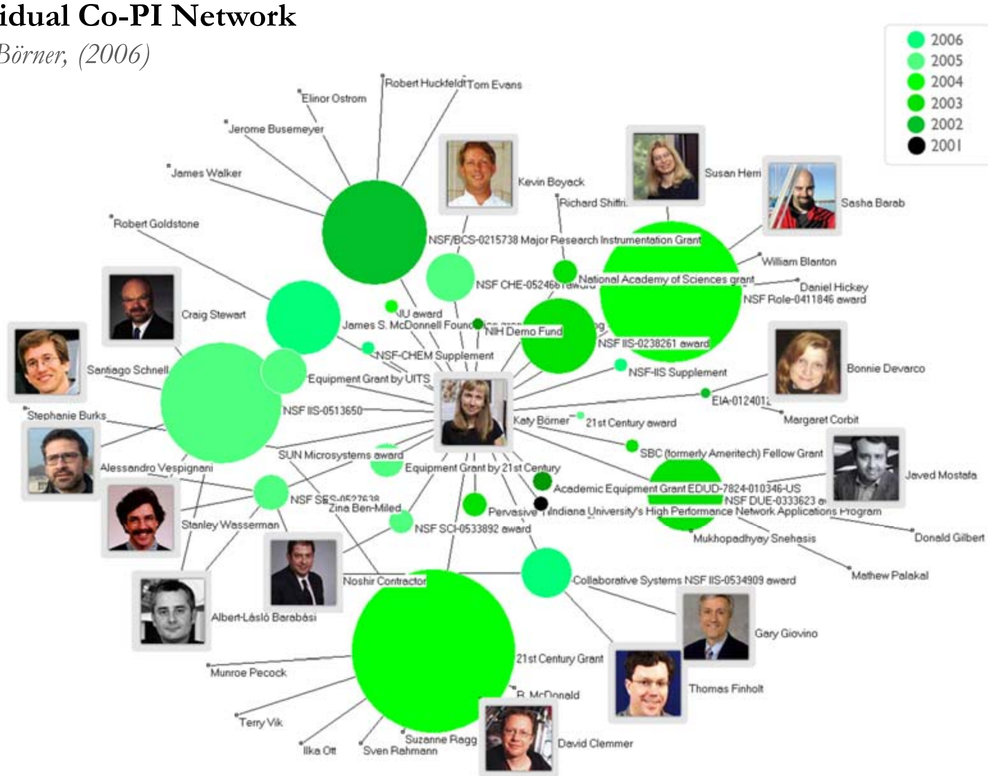
Mapping Indiana's Intellectual Space

Identify

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

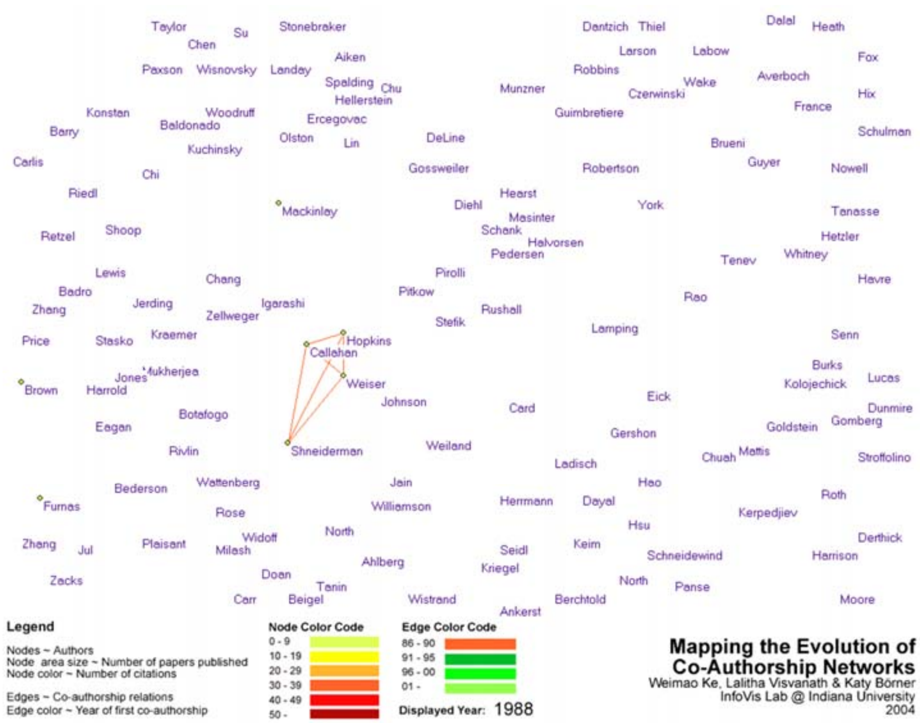
Individual Co-PI Network

Ke & Börner, (2006)



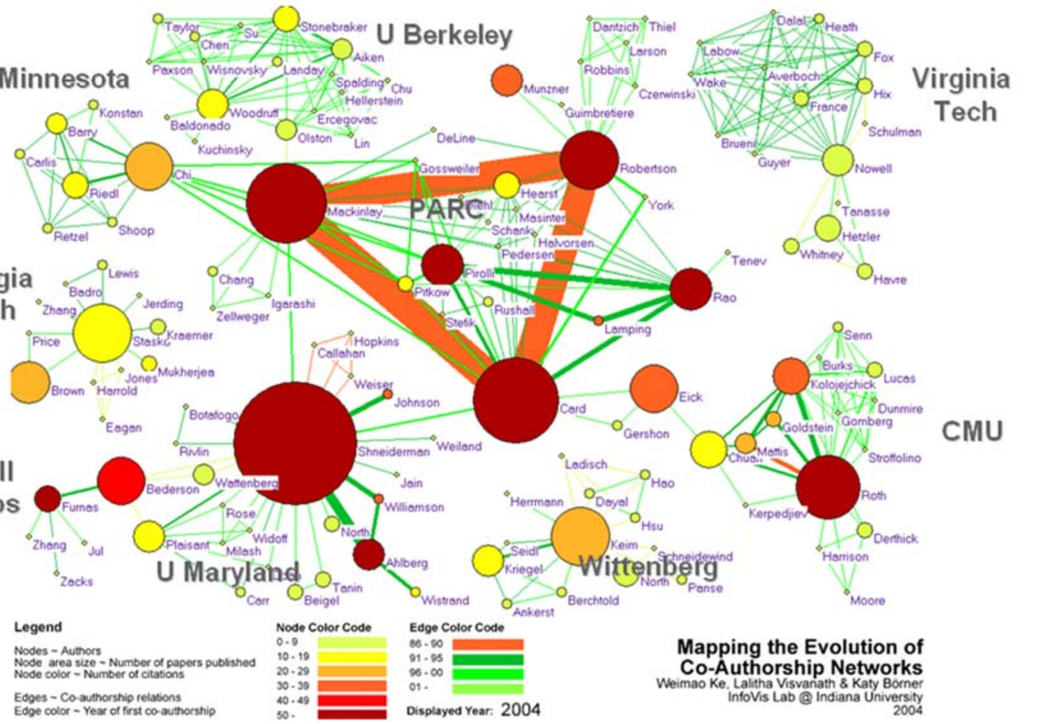
Mapping the Evolution of Co-Authorship Networks

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



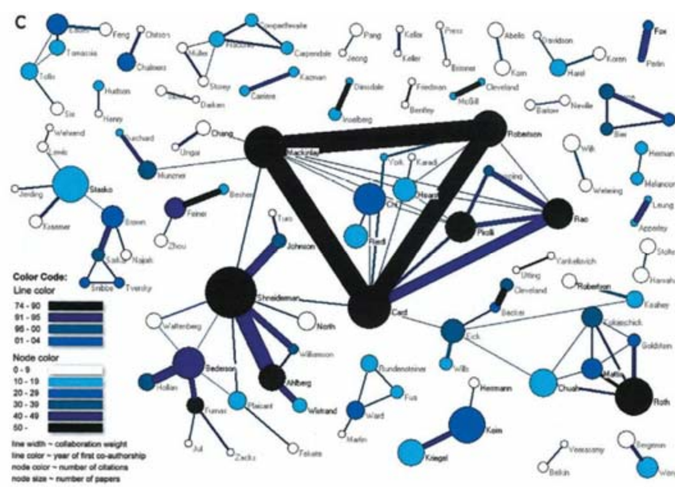
Mapping the Evolution of Co-Authorship Networks

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



Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams

Börner, Dall'Asta, Ke & Vespignani (2005)
Complexity, 10(4):58-67.



Research question:

- Is science driven by prolific single experts or by high-impact co-authorship teams?

Contributions:

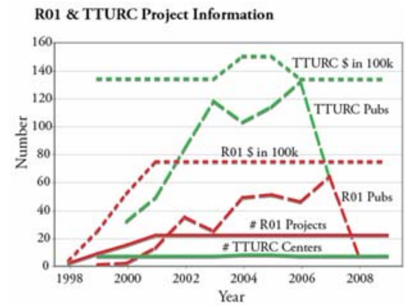
- New approach to allocate citational credit.
- Novel weighted graph representation.
- Visualization of the growth of weighted co-author network.
- Centrality measures to identify author impact.
- Global statistical analysis of paper production and citations in correlation with co-authorship team size over time.
- Local, author-centered entropy measure.

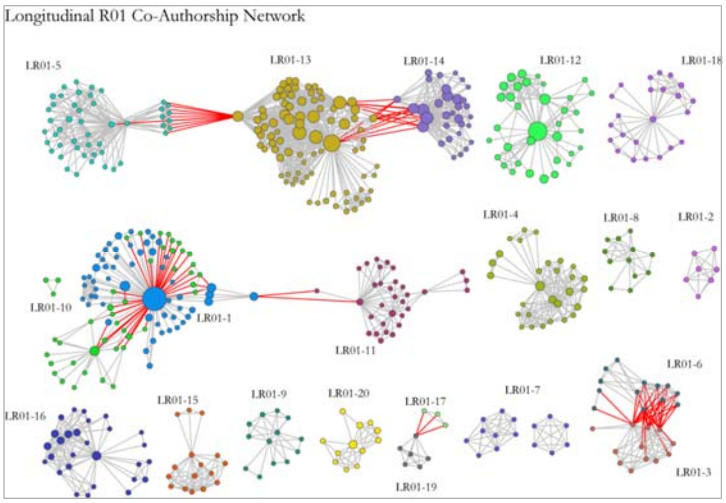
Mapping Transdisciplinary Tobacco Use Research Centers Publications

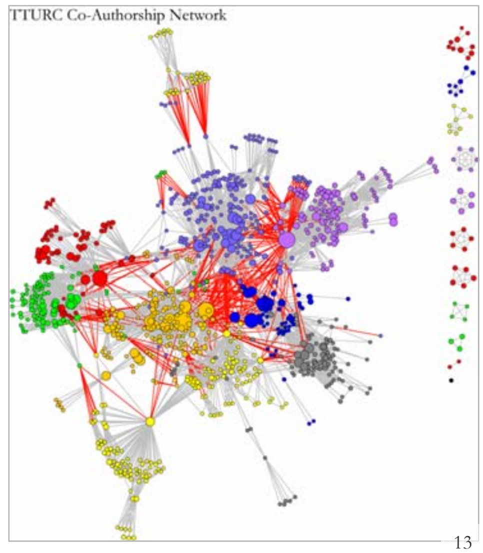
Compare R01 investigator based funding with TTURC Center awards in terms of number of publications and evolving co-author networks.

Zoss & Börner, forthcoming.

Supported by NIH/NCI Contract HHSN261200800812

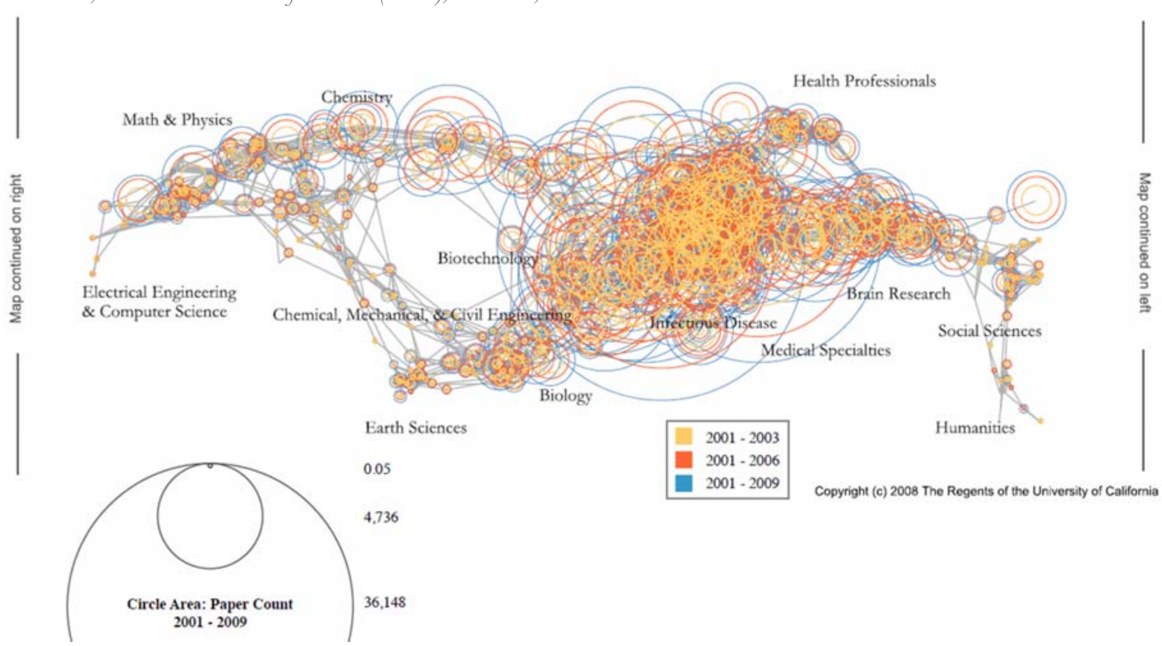






MEDLINE Publication Output by The National Institutes of Health (NIH) Using Nine Years of ExPORTER Data

Katy Börner, Nianli Ma, Joseph R. Biberstine, Cyberinfrastructure for Network Science Center, SLIS, Indiana University, Robin M. Wagner, Rediet Berhane, Hong Jiang, Susan E. Ivey, Katrina Pearson and Carl McCabe, Reporting Branch, Division of Information Services, Office of Research Information Systems, Office of Extramural Research, Office of the Director, National Institutes of Health (NIH), Bethesda, MD.

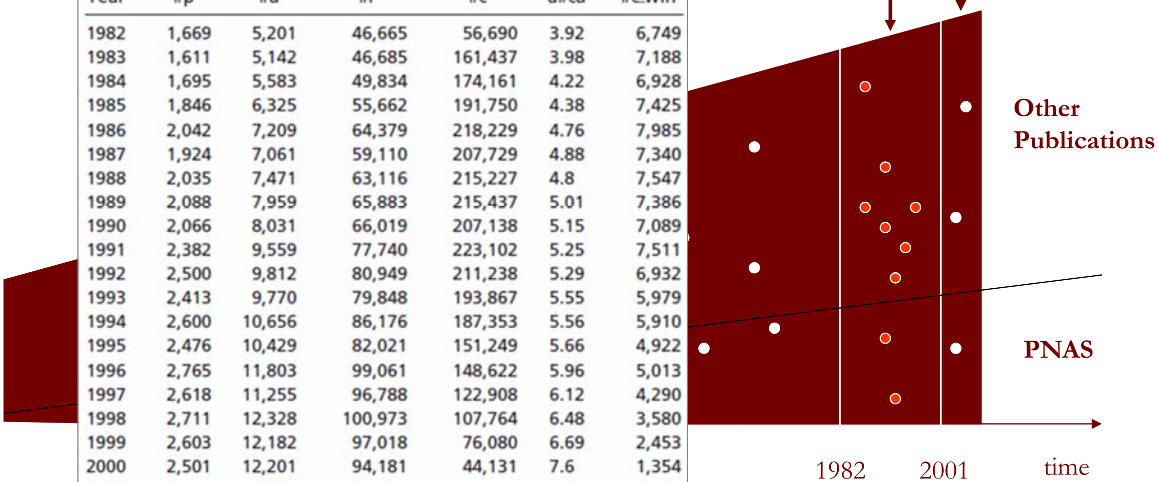


20-Year PNAS Dataset (1982-2001)

45,120 regular articles written by 105,915 unique authors.
114,000 citation references within the set and 472,000 co-author links.

Table 1. PNAS statistics in terms of total number of papers (#p), unique authors (#a), references (#r), citations received per paper (#c), number of coauthors per paper (a#ca), and the number of citations (#c_win) within the PNAS data set for each year

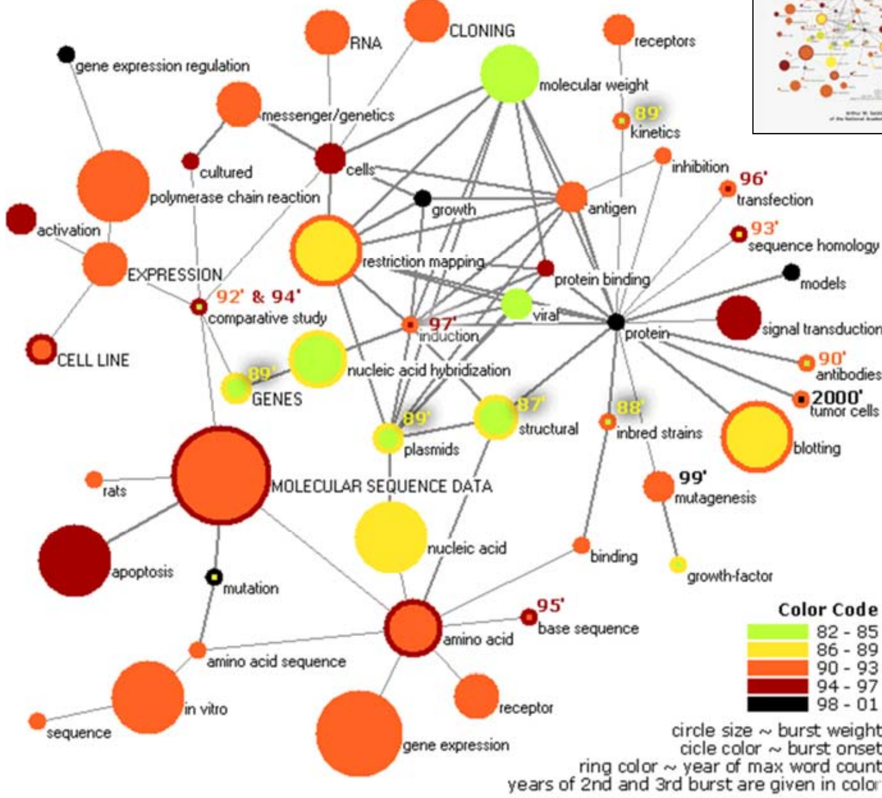
Year	#p	#a	#r	#c	a#ca	#c_win
1982	1,669	5,201	46,665	56,690	3.92	6,749
1983	1,611	5,142	46,685	161,437	3.98	7,188
1984	1,695	5,583	49,834	174,161	4.22	6,928
1985	1,846	6,325	55,662	191,750	4.38	7,425
1986	2,042	7,209	64,379	218,229	4.76	7,985
1987	1,924	7,061	59,110	207,729	4.88	7,340
1988	2,035	7,471	63,116	215,227	4.8	7,547
1989	2,088	7,959	65,883	215,437	5.01	7,386
1990	2,066	8,031	66,019	207,138	5.15	7,089
1991	2,382	9,559	77,740	223,102	5.25	7,511
1992	2,500	9,812	80,949	211,238	5.29	6,932
1993	2,413	9,770	79,848	193,867	5.55	5,979
1994	2,600	10,656	86,176	187,353	5.56	5,910
1995	2,476	10,429	82,021	151,249	5.66	4,922
1996	2,765	11,803	99,061	148,622	5.96	5,013
1997	2,618	11,255	96,788	122,908	6.12	4,290
1998	2,711	12,328	100,973	107,764	6.48	3,580
1999	2,603	12,182	97,018	76,080	6.69	2,453
2000	2,501	12,201	94,181	44,131	7.6	1,354
2001	2,575	13,038	97,450	16,357	8.4	422
Total	45,120		1,509,558	3,230,469		114,003



Mapping Topic Bursts

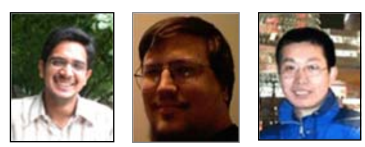
Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.

Mane & Börner. (2004) PNAS, 101(Suppl. 1): 5287-5290.



Information Diffusion Among Major U.S. Research Institutions

Börner, Katy, Penumarthy, Shashikant, Meiss, Mark & Ke, Weimao. (2006). Mapping the Diffusion of Information among Major U.S. Research Institutions. Scientometrics. Vol. 68(3), 415 - 426.

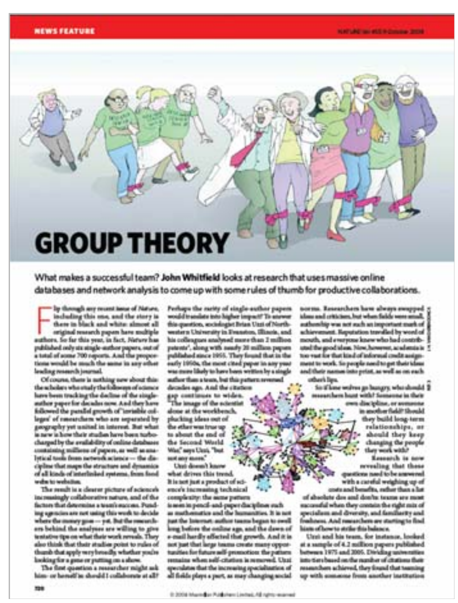


Questions:

1. Does space still matter in the Internet age, i.e., does one still have to study and work at major research institutions in order to have access to high quality data and expertise and to produce high quality research?
2. Does the Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research instructions?

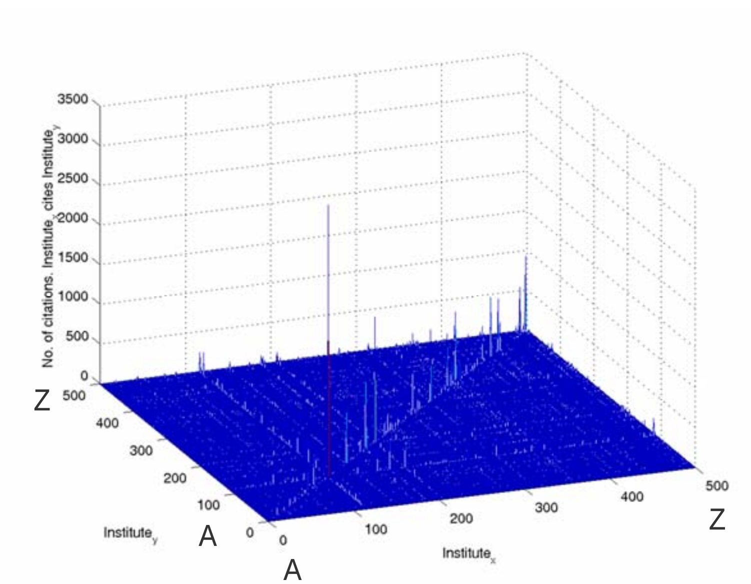
Contributions:

- Answer to Q1 is YES.
- Answer to Q2 is NO.
- Novel approach to analyzing the dual role of institutions as information producers and consumers and to study and visualize the diffusion of information among them.



Citation Matrix

Unsymmetrical direct citation linkage patterns among the top 500 institutions in US. High peak values in the diagonal reflect the high amount of self-citations for all institutions. Medium peak horizontal and vertical lines denote references from and citations to papers written at Harvard University.



Information Sources (Export) and Sinks (Import)

Calculate ratio of the number of references made by an institution divided by the sum of received citations and references made, multiplied by 100.

131 have a value between 0-40% acting mostly as information producers = information sources.

71 have a value between 60-100% and act mostly as information consumers – they reference a large number of papers but the number of citations they receive is comparably low = information sinks.

Geographic Location of Received Citations

ESRI's ArcGIS program was used to show the geographic distribution of the top 500 institutions using the Albers equal area projection.

U.S. states are color coded based on the population size in the year 2000. Lighter shades of green represent lower populations.

Overlaid are the top 500 institutions, each represented by a 'citation stick'. The color and height of the stick corresponds to the number of received citations (excluding self citations).

Five institutions produced papers that attracted more than 5,000 citations and are labeled.

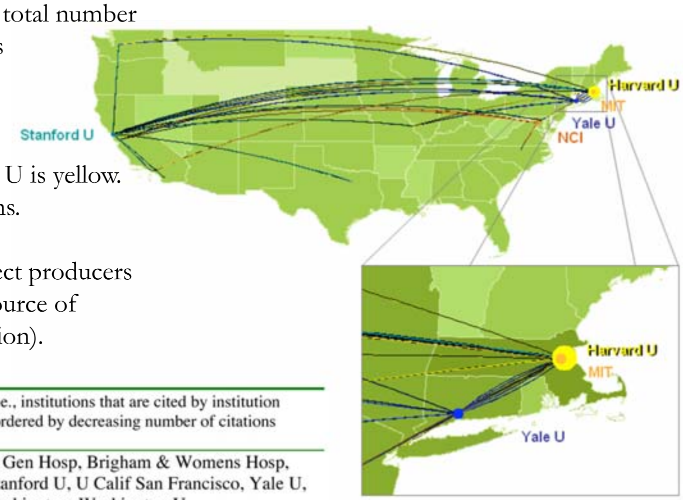
Harvard leads with 16,531 citations.

Information Flow Among the Top-5 Consumers and Their Top-10 Producers

U.S. states are color coded based on the total number of citations received by their institutions (excluding self citations).

Dots indicate the five producers.
Each has a different color, e.g., Harvard U is yellow.
Dot area size depicts number of citations.

Lines represent citations that interconnect producers and consumers shaded from colored (source of information) to white (sink of information).

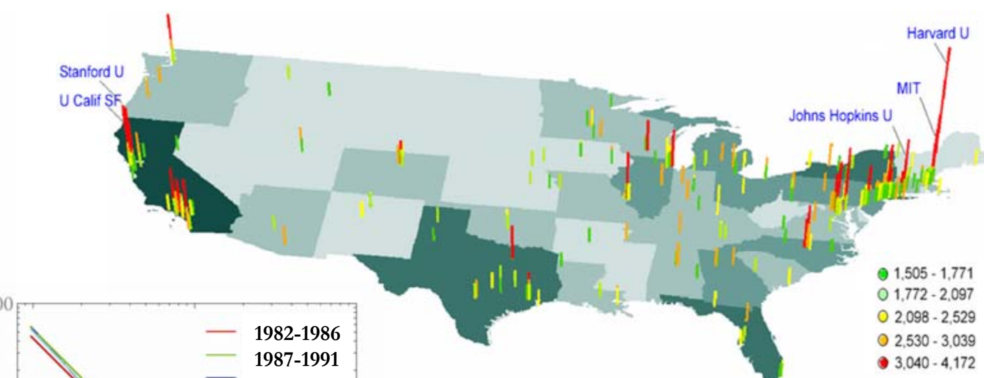


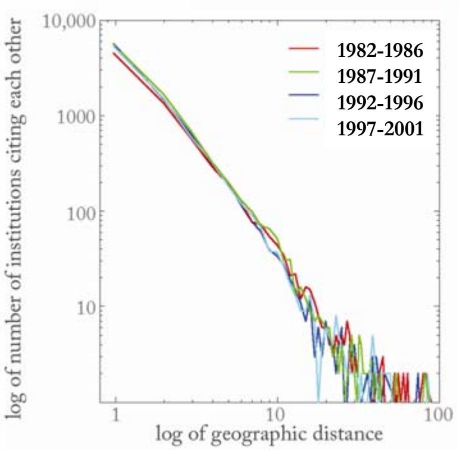
Consumers, i.e., citing institutions	# citations made	Top ten producers, i.e., institutions that are cited by institution listed in first column ordered by decreasing number of citations received.
Harvard U	13,552	MIT, Massachusetts Gen Hosp, Brigham & Womens Hosp, Johns Hopkins U, Stanford U, U Calif San Francisco, Yale U, Rockefeller U, U Washington, Washington U
U Calif SF	4,682	Harvard U, MIT, Stanford U, Johns Hopkins U, U Washington, Washington U, U Calif Berkeley, U Texas, U Calif SD, U Calif LA
MIT	4,655	Harvard U, Whitehead Inst Biomed Res, Johns Hopkins U, Stanford U, U Calif SF, Yale U, Rockefeller U, U Calif LA, Massachusetts Gen Hosp, U Calif Berkeley
NCI (zip: 20814)	4,519	Harvard U, NCI (zip: 20205), NCI (zip: 21701), MIT, Duke U, Johns Hopkins U, NIAID NICHHD, Stanford U, U Calif SF
Yale U	4,464	Harvard U, MIT, Stanford U, Rockefeller U, Johns Hopkins U, Washington U, U Calif SF, U Washington, NCI, Massachusetts Gen Hosp

Paper also shows top-5 producers and their top-10 consumers.

Changes in Citation Behavior Over Time

As time progresses and the amount of produced papers increases, space seems to matter more.
Authors are more likely to cite papers generated by authors at close-by institutions.





1982-1986: **1.94** (R^2=91.5%)
1987-1991: **2.11** (R^2=93.5%)
1992-1996: **2.01** (R^2=90.8%)
1997-2001: **2.01** (R^2=90.7%)

Modeling the Co-Evolving Author-Paper Networks

Börner, Katy, Maru, Jeegar & Goldstone, Robert. (2004). The Simultaneous Evolution of Author and Paper Networks. PNAS. Vol. 101(Suppl. 1), 5266-5273.

The TARL Model (Topics, Aging, and Recursive Linking) incorporates

- A partitioning of authors and papers into topics,
- Aging, i.e., a bias for authors to cite recent papers, and
- A tendency for authors to cite papers cited by papers that they have read resulting in a rich get richer effect.

The model attempts to capture the roles of authors and papers in the production, storage, and dissemination of knowledge.

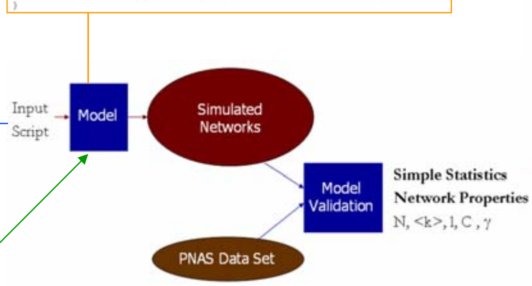
Model Assumptions

- Co-author and paper-citation networks co-evolve.
- Authors come and go.
- Papers are forever.
- Only authors that are 'alive' are able to co-author.
- All existing (but no future) papers can be cited.
- Information diffusion occurs directly via co-authorships and indirectly via the consumption of other authors' papers.

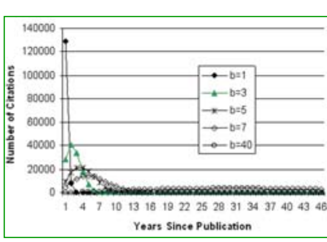
- Preferential attachment is modeled as an *emergent property* of the elementary, local networking activity of authors reading and citing papers, but also the references listed in papers.


```
--------------------------------------------
Model Parameters (0=without, 1=with)
--------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
--------------------------------------------
Model Initialization Values
--------------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1   # Levels References are Considered
```

```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select #_read_ papers from existing papers;
                get all references of read_ papers up to #_reference_path_length;
                for each new_paper_reference do {
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```



Aging function



Model Validation

The properties of the networks generated by this model are validated against the 20-year PNAS data set (1982-2001).

Table 3 Statistics for SIM data

Year	#p	#a	#r	#c	a#ca
1981	1624	3953	0	756	8.21
1982	1040	5200	31200	112161	4
1983	1118	5590	33540	21397	4
1984	1197	5985	35910	10224	4
1985	1275	6375	38250	6184	4
1986	1353	6765	40590	4687	4
1987	1432	7160	42960	3573	4
1988	1510	7550	45300	2816	4
1989	1589	7945	47670	2219	4
1990	1667	8335	50010	1853	4
1991	1745	8725	52350	1634	4
1992	1824	9120	54720	1431	4
1993	1902	9510	57060	1167	4
1994	1981	9905	59430	1040	4
1995	2059	10295	61770	767	4
1996	2137	10685	64110	632	4
1997	2216	11080	66480	522	4
1998	2294	11470	68820	400	4
1999	2373	11865	71190	265	4
2000	2451	12255	73530	125	4
2001	2529	12645	75870	0	4
Total	**37316**		**1070760**	**173853**	

Table 2. PNAS Statistics

Year	#p	#a	#r	#c	a#ca
1982	1669	5201	46665	156690	3.92
1983	1611	5142	46685	161437	3.98
1984	1695	5583	49834	174161	4.22
1985	1846	6325	55662	191750	4.38
1986	2042	7209	64379	218229	4.76
1987	1924	7061	59110	207729	4.88
1988	2035	7471	63116	215227	4.8
1989	2088	7959	65883	215437	5.01
1990	2066	8031	66019	207138	5.15
1991	2382	9559	77740	223102	5.25
1992	2500	9812	80949	211238	5.29
1993	2413	9770	79848	193867	5.55
1994	2600	10656	86176	187353	5.56
1995	2476	10429	82021	151249	5.66
1996	2765	11803	99061	148622	5.96
1997	2618	11255	96788	122908	6.12
1998	2711	12328	100973	107764	6.48
1999	2603	12182	97018	76080	6.69
2000	2501	12201	94181	44131	7.6
2001	2575	13038	97450	16357	8.4
Total	**45120**		**1509558**	**3230469**	

Process Model in Pseudocode & Input Parameters

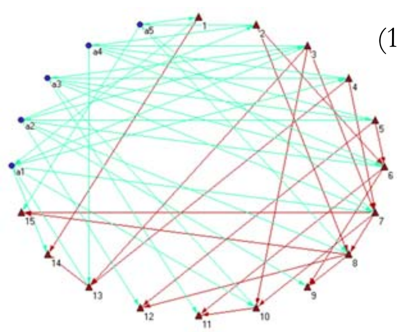
If no topics are considered then the number of topics is one, i.e., all papers and authors have the same topic. If no coauthors are considered then each paper has exactly one author. If the reference path length is 0 then no references are considered for citation.

```
-----------------------------------------
Model Parameters (0=without, 1=with)
-----------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
-----------------------------------------
Model Initialization Values
-----------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1  # Levels References are Considered
```

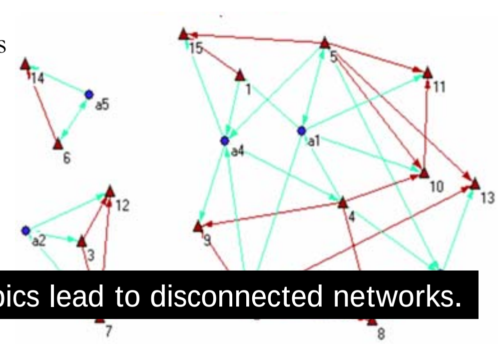
```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select #_read_ papers from existing papers;
                get all references of read_ papers up to #_reference_path_length;
                for each new_paper_reference do {
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```


The TARL Model: The Effect of Parameters

(0000)

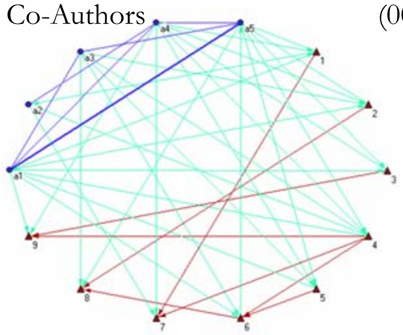


(1000) Topics

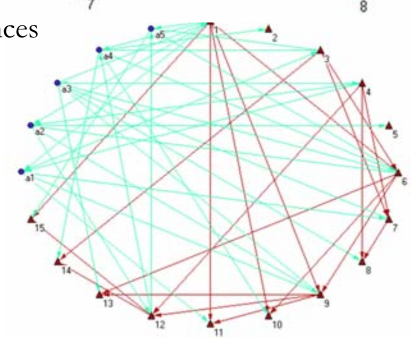


Topics lead to disconnected networks.

(0100) Co-Authors



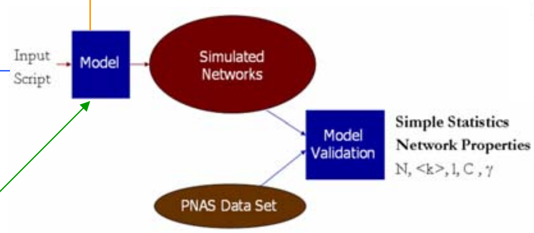
(0010) References



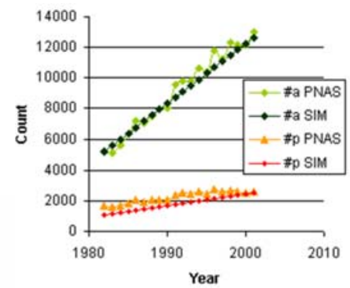
Co-authoring leads to fewer papers.

```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select #_read_ papers from existing papers;
                get all references of read_ papers up to #_reference_path_length;
                for each new_paper_reference do {
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```

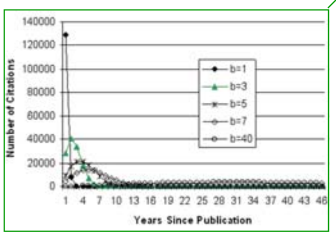
```
---------------------------------------------
Model Parameters (0=without, 1=with)
---------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
---------------------------------------------
Model Initialization Values
---------------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1  # Levels References are Considered
```



Counts for Papers and Authors



Aging function



Counts for Citations

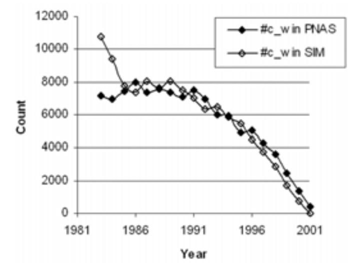
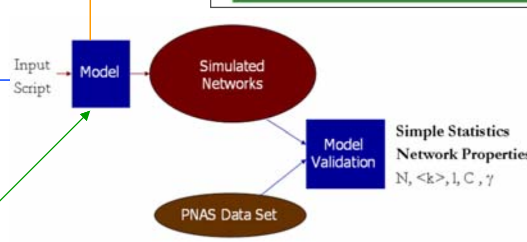


Table 2. Properties of co-author & paper citation networks comprising number of nodes n, average node degree <k>, path length l, cluster coefficient C, and power law exponent γ. Source references are given in the left column.

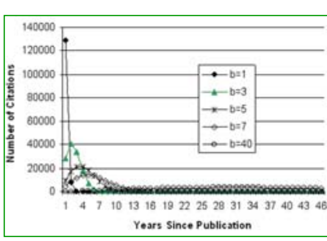
Network	n	<k>	l	C	γ	Reference
Co-authorship networks						
LANL	52,909	9.7	5.9	0.43	--	Newman, (2001a; 2001b; 2001c)
MEDLINE	1,520,251	18.1	4.6	0.066	--	
SPIRES	56,627	1.73	4.0	0.726	1.2	
NCSTRL	11,994	3.59	9.7	0.496	--	
Math.	70,975	3.9	9.5	0.59	2.5	Barabasi et al., (2002)
Neurosci.	209,293	11.5	6	0.76	2.1	
PNAS	105,915	8.97	5.89	0.399	2.54	
Paper-citation networks						
ISI	783,339	8.57	--	--	3	Redner, (1998)
PhysRev	24,296	14.5	--	--	3	
PNAS	45,120	3.53	--	0.081	2.29	
SIM	37,114	2.13	--	0.074	2.05	

Co-Author and Paper-Citation Network Properties

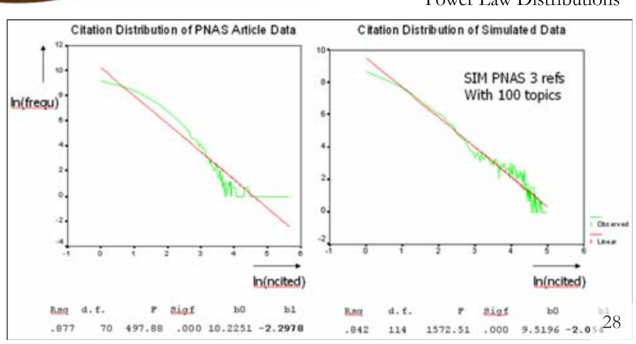
```
---------------------------------------------
Model Parameters (0=without, 1=with)
---------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
---------------------------------------------
Model Initialization Values
---------------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1  # Levels References are Considered
```

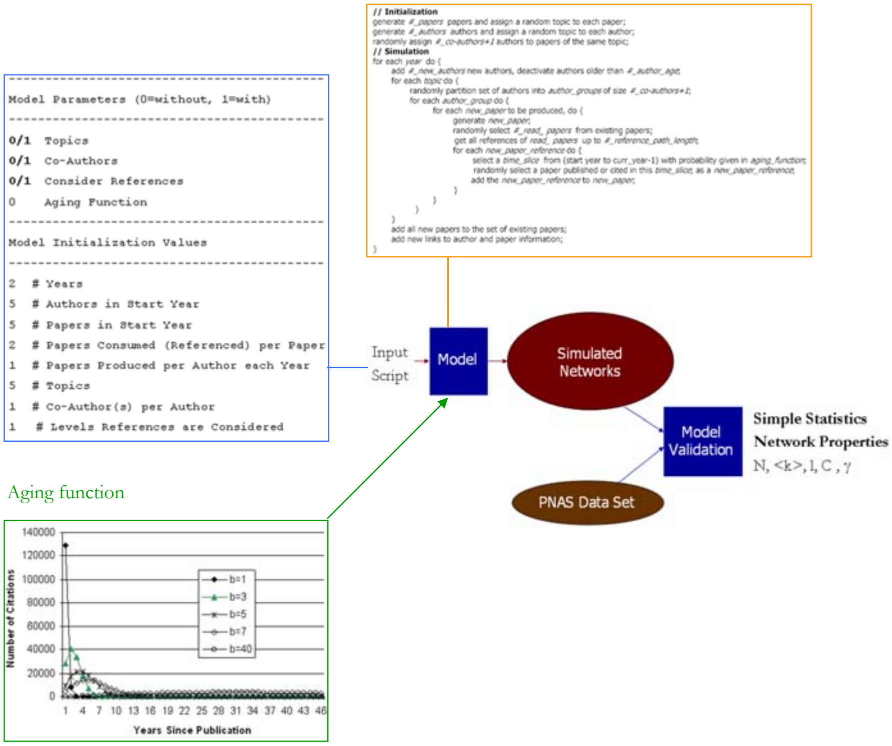


Aging function



Power Law Distributions





Topics: The number of topics is linearly correlated with the clustering coefficient of the resulting network: C= 0.000073 * #topics. Increasing the number of topics increases the power law exponent as authors are now restricted to cite papers in their own topics area.

Aging: With increasing b, and hence increasing the number of older papers cited as references, the clustering coefficient decreases. Papers are not only clustered by topic, but also in time, and as a community becomes increasingly nearsighted in terms of their citation practices, the degree of temporal clustering increases.

References/Recursive Linking: The length of the chain of paper citation links that is followed to select references for a new paper also influences the clustering coefficient. Temporal clustering is ameliorated by the practice of citing (and hopefully reading!) the papers that were the earlier inspirations for read papers.

References

Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). **Visualizing Knowledge Domains.** In Blaise Cronin (Ed.), *ARIST*, Medford, NJ: Information Today, 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., Volume 41, Chapter 12, pp. 537-607. http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf

Börner, Katy (2010) **Atlas of Science**. MIT Press. http://scimaps.org/atlas

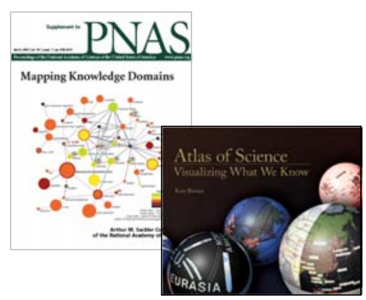
Scharnhorst, Andrea, Börner, Katy, van den Besselaar, Peter (2011) **Models of Science Dynamics**. Springer Verlag.

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*

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

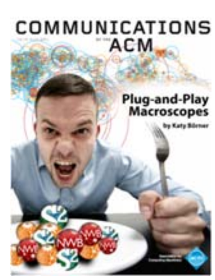
Information on how to host the exhibit or acquire a subset of the maps is at http://scimaps.org

Overview



1. **Data mining and visualization research** that aims to increase our scientific understanding of the structure and dynamics of science and technology.

2. **Novel approaches and services** that improve information access, researcher networking, and research management.



3. **Data services and plug-and-play macroscope tools** that commoditize data mining and visualization.

cyberinfrastructure for NETWORK SCIENCE CENTER
School of Library and Information Science | Indiana University Bloomington

Different Stakeholder Groups and Their Needs

Funding Agencies

- Need to monitor (long-term) money flow and research developments, identify areas for future development, stimulate new research areas, evaluate funding strategies for different programs, decide on project durations, funding patterns.

Scholars

- Want easy access to research results, relevant funding programs and their success rates, potential collaborators, competitors, related projects/publications *(research push)*.

Industry

- Is interested in fast and easy access to major results, experts, etc. Influences the direction of research by entering information on needed technologies *(industry-pull)*.

Advantages for Publishers

- Need easy to use interfaces to massive amounts of interlinked data. Need to communicate data provenance, quality, and context.

Society

- Needs easy access to scientific knowledge and expertise.

Scholars Have Different Roles/Needs

Researchers and Authors—need to select promising research topics, students, collaborators, and publication venues to increase their reputation. They benefit from a global view of competencies, reputation and connectivity of scholars; hot and cold research topics and bursts of activity, and funding available per research area.

Editors—have to determine editorial board members, assign papers to reviewers, and ultimately accept or reject papers. Editors need to know the position of their journals in the evolving world of science. They need to advertise their journals appropriately and attract high-quality submissions, which will in turn increase the journal's reputation.

Reviewers—read, critique, and suggest changes to help improve the quality of papers and funding proposals. They need to identify related works that should be cited or complementary skills that authors might consider when selecting project collaborators.

Teachers/Mentors—teach classes, train doctoral students, and supervise postdoctoral researchers. They need to identify key works, experts, and examples relevant to a topic area and teach them in the context of global science.

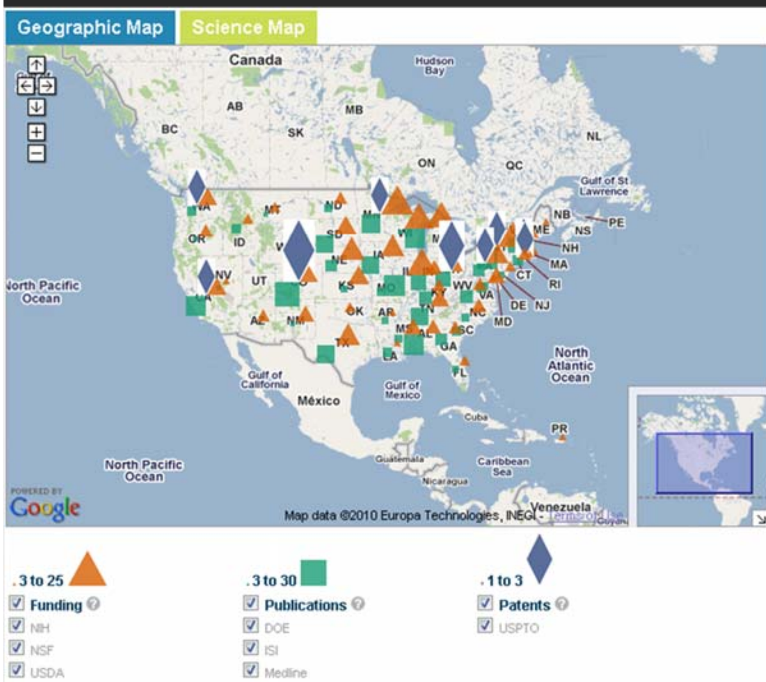
Inventors—create intellectual property and obtain patents, thus needing to navigate and make sense of research spaces as well as intellectual property spaces.

Investigators—scholars need funding to support students, hire staff, purchase equipment, or attend conferences. Here, research interests and proposals have to be matched with existing federal and commercial funding opportunities, possible industry collaborators and sponsors.

Team Leads and Science Administrators—many scholars direct multiple research projects simultaneously. Some have full-time staff, research scientists, and technicians in their laboratories and centers. Leaders need to evaluate performance and provide references for current or previous members; report the progress of different projects to funding agencies.

Mapping Sustainability Research

http://mapsustain.cns.iu.edu



Maps Detail **Data** About

Datasets

The dataset covers 13,528 records on "biomass" and "biofuel" research and technology from seven different publication, patent, funding datasets for the years 1901 to 2010.

Funding

National Institutes of Health (NIH) awards retrieved from the Scholarly Database (http://sdb.slis.indiana.edu) at Indiana University on 11/20/2010. Search query used was biomass OR biofuel OR "bio mass" OR "bio fuel" in the 'All Text' field.

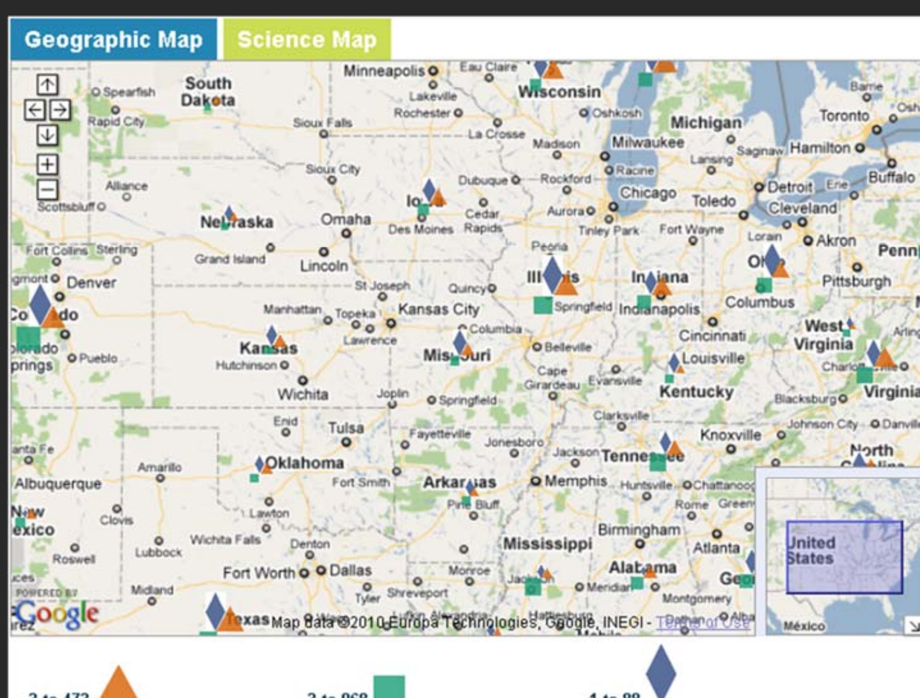
National Science Foundation (NSF) awards retrieved from the Scholarly Database (http://sdb.slis.indiana.edu) at Indiana University on 11/20/2010. Search query used was biomass OR biofuel OR "bio mass" OR "bio fuel" in the 'All Text' field.

US Department of Agriculture (USDA) awards made available by a staff member of the Office of Scientific and Technical Information from the US Department of Energy (DOE).

Publications

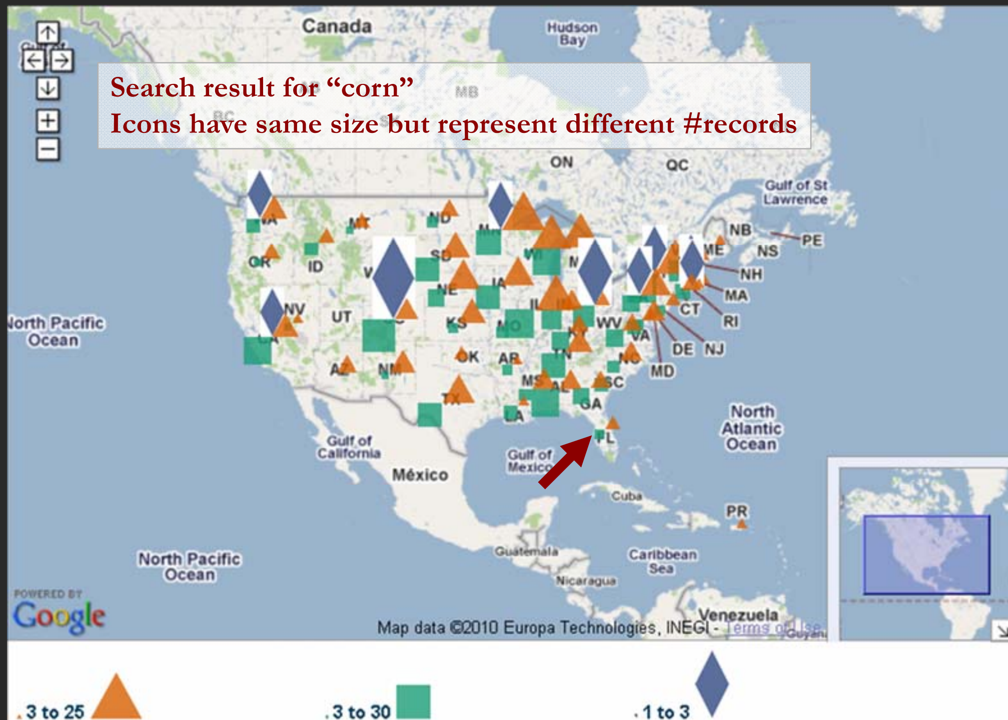
MEDLINE papers by the National Library of Medicine retrieved from the Scholarly Database (http://sdb.slis.indiana.edu) at Indiana University on 11/20/2010. Search query used was biomass OR biofuel OR "bio mass" OR "bio fuel" in the 'All Text' field.

The geographic map at **state level**.

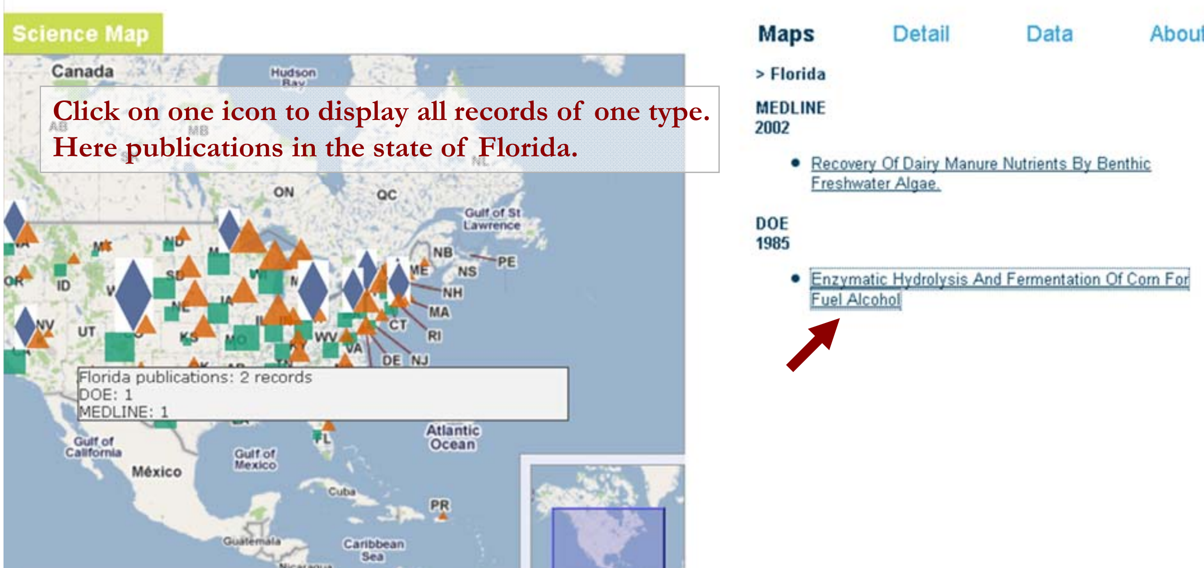


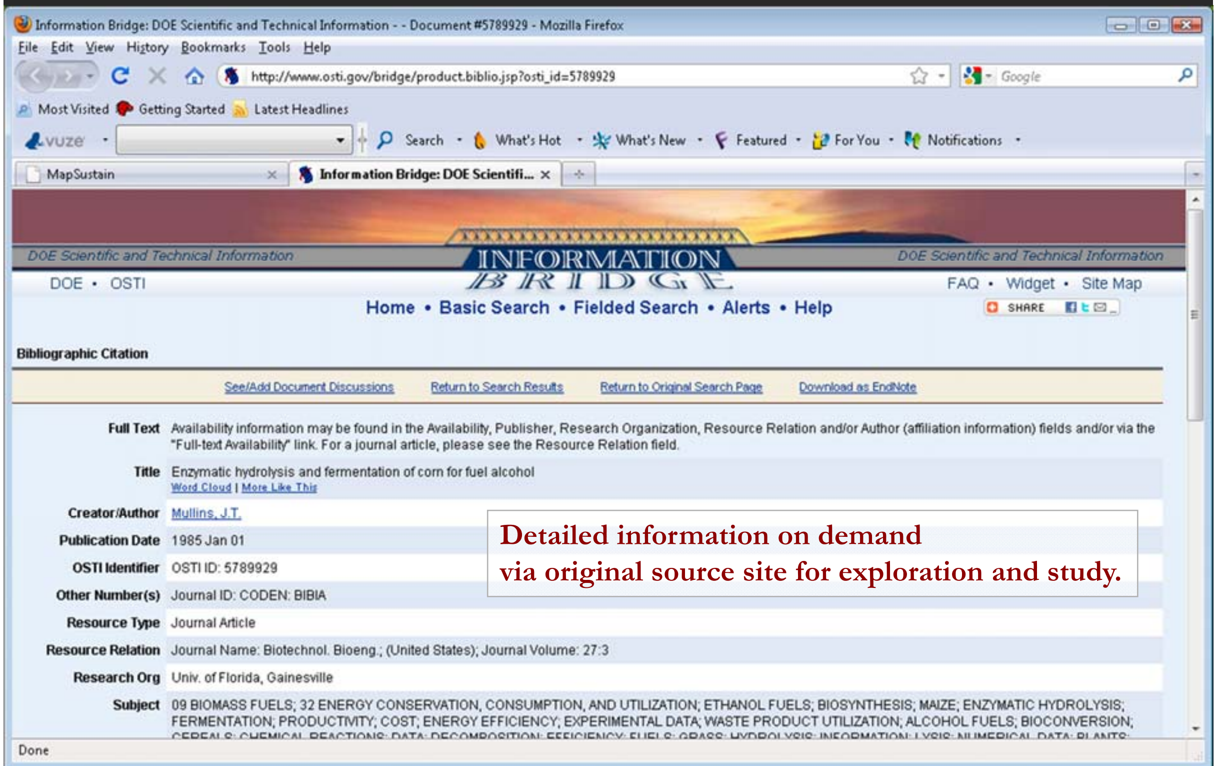
The geographic map at **city level**.

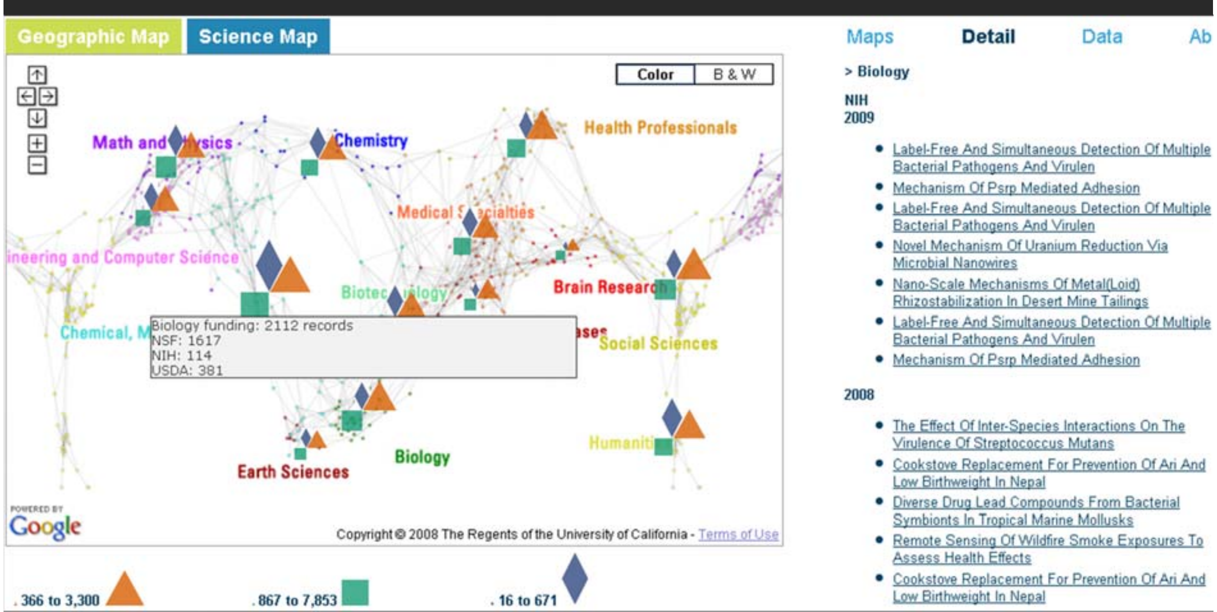
Search result for “corn”
Icons have same size but represent different #records



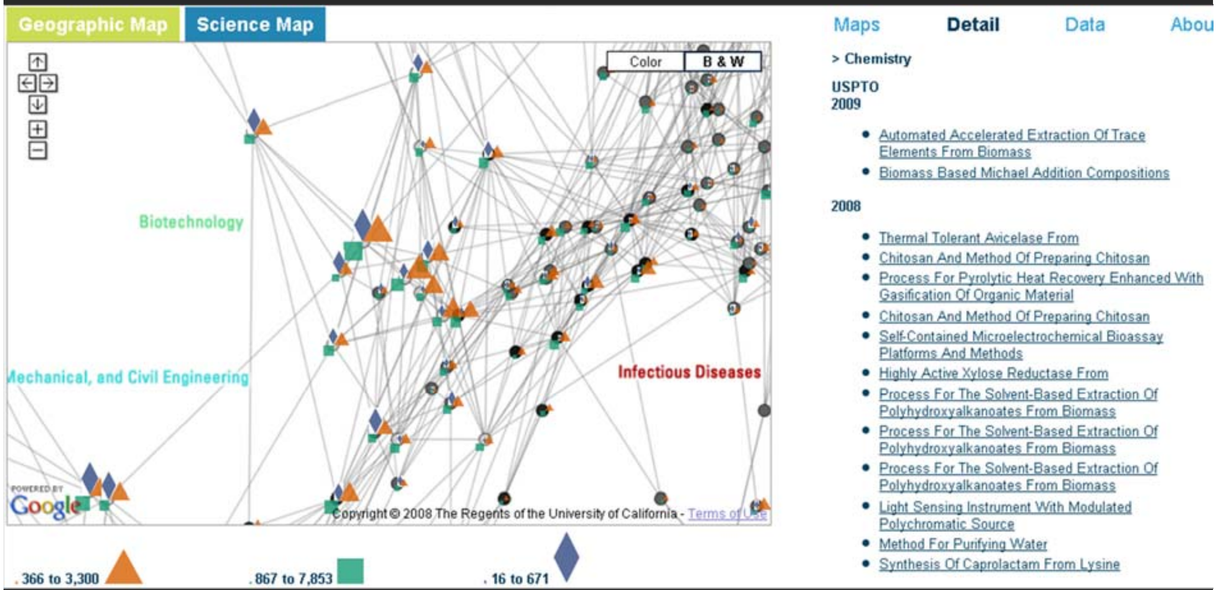
Click on one icon to display all records of one type.
Here publications in the state of Florida.



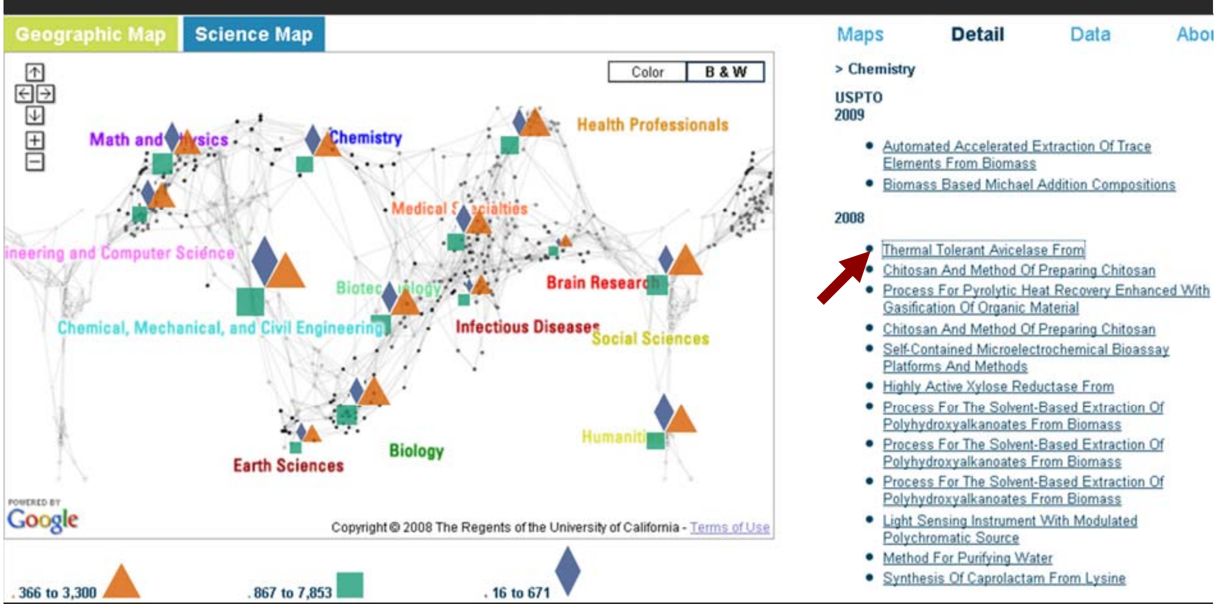


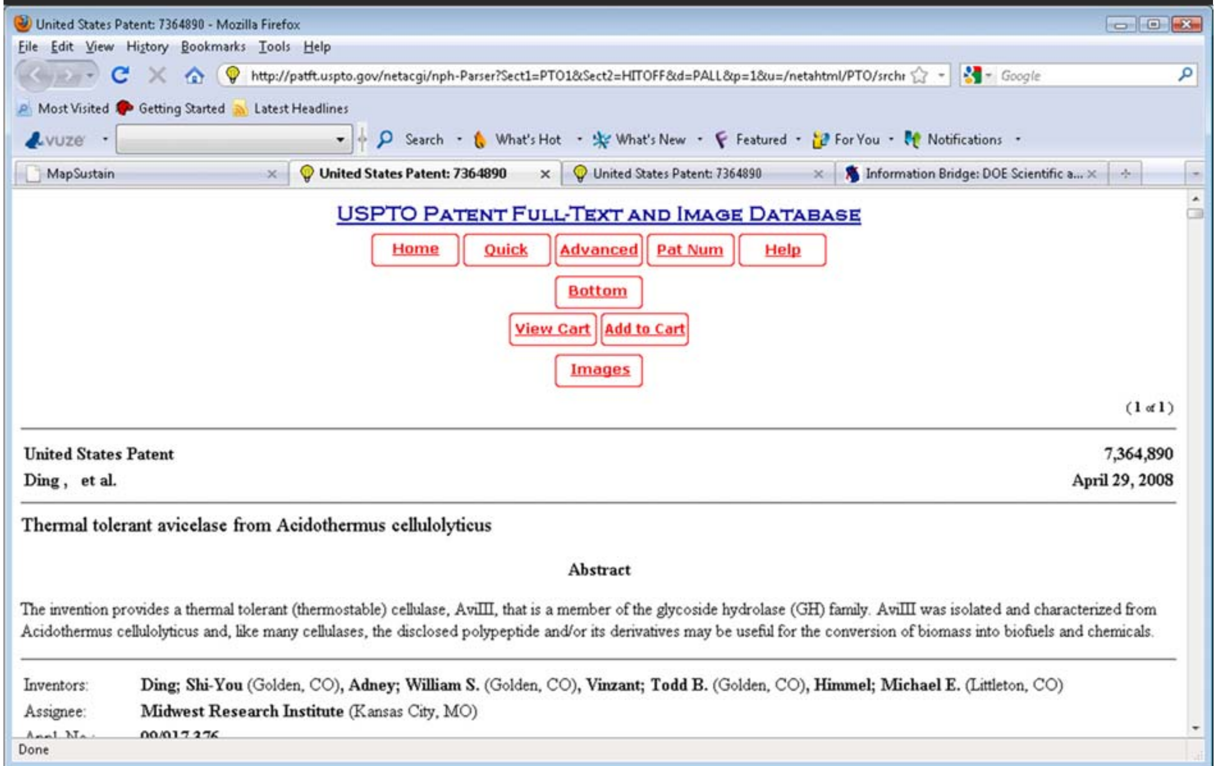


The science map at 13 top-level scientific **disciplines level**.

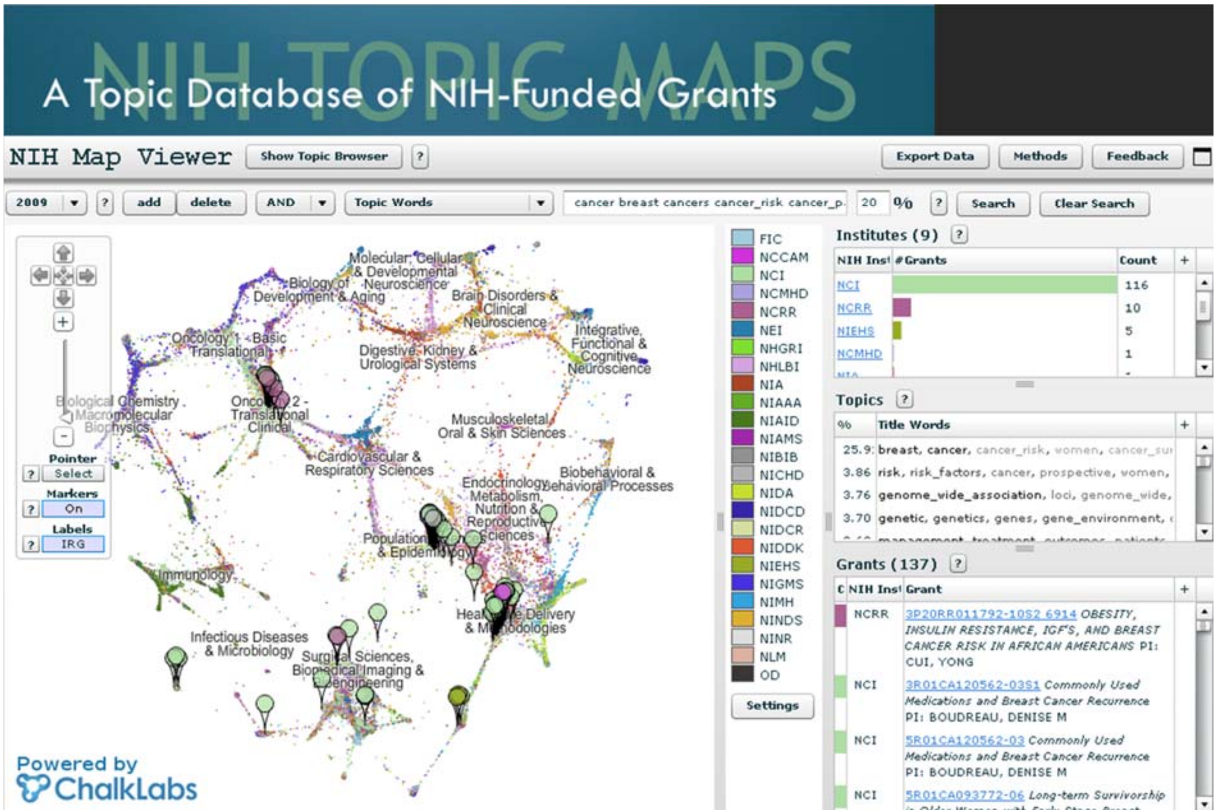


The science map at 554 **sub-disciplines level**.

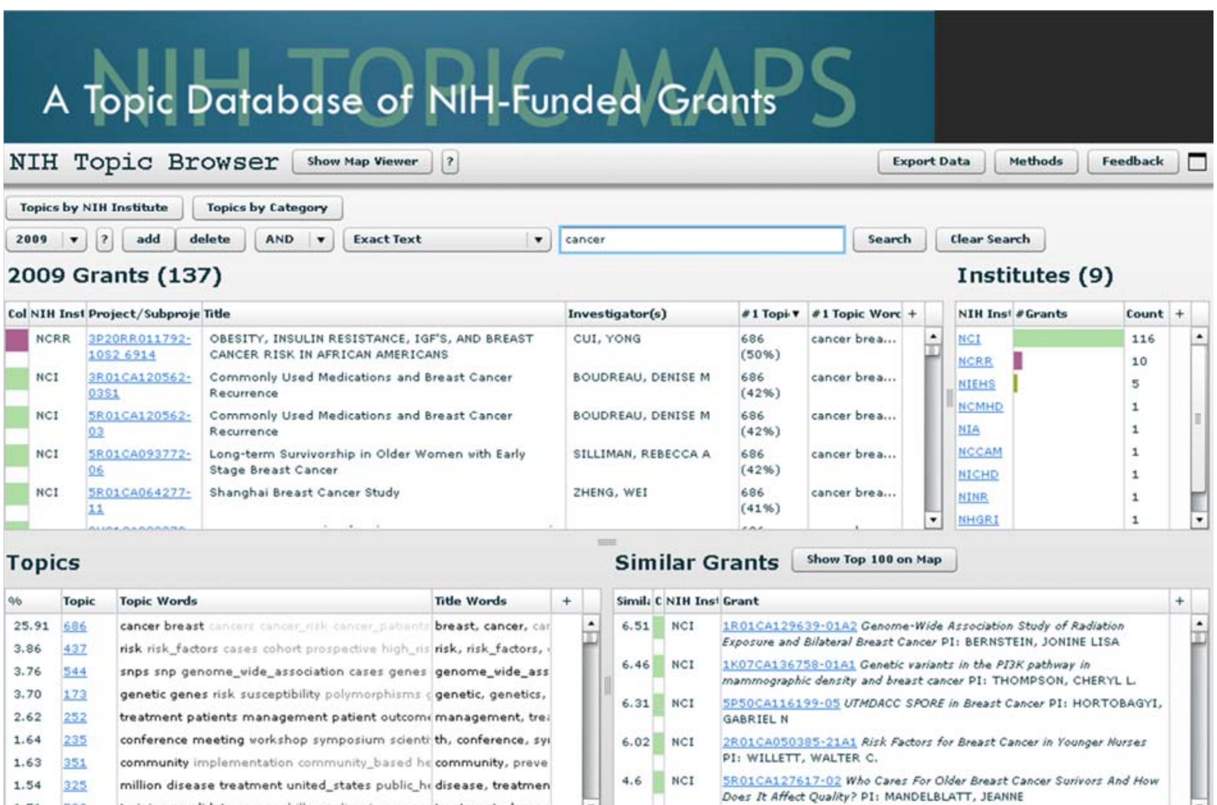




NIH Topic Maps



https://app.nihmaps.org

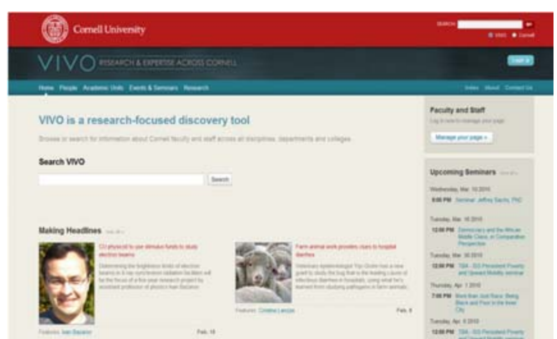


https://app.nihmaps.org

VIVO International Researcher Network

VIVO: A Semantic Approach to Creating a National Network of Researchers (http://vivoweb.org)

- Semantic web application and ontology editor originally developed at Cornell U.
- Integrates research and scholarship info from systems of record across institution(s).
- Facilitates research discovery and cross-disciplinary collaboration.
- Simplify reporting tasks, e.g., generate biosketch, department report.



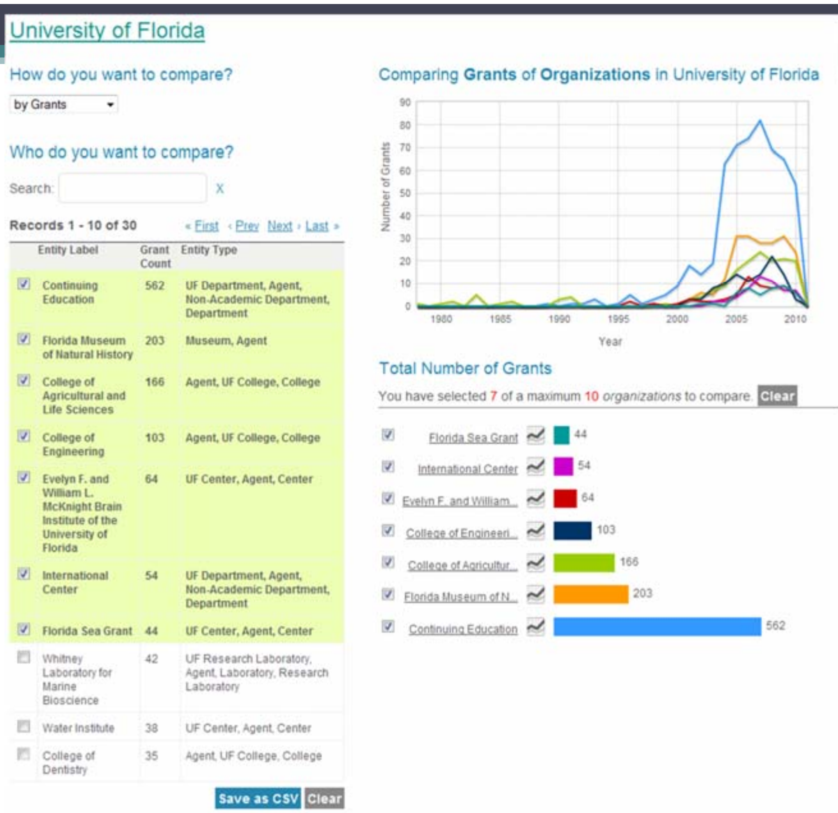
Funded by $12 million NIH award.

Cornell University: Dean Krafft (Cornell PI), Manolo Bevia, Jim Blake, Nick Cappadona, Brian Caruso, Jon Corson-Rikert, Elly Cramer, Medha Devare, John Fereira, Brian Lowe, Stella Mitchell, Holly Mistlebauer, Anup Sawant, Christopher Westling, Rebecca Younes. **University of Florida:** Mike Conlon (VIVO and UF PI), Cecilia Botero, Kerry Britt, Erin Brooks, Amy Buhler, Ellie Bushhousen, Chris Case, Valrie Davis, Nita Ferree, Chris Haines, Rae Jesano, Margeaux Johnson, Sara Kreinest, Yang Li, Paula Markes, Sara Russell Gonzalez, Alexander Rockwell, Nancy Schaefer, Michele R. Tennant, George Hack, Chris Barnes, Narayan Raum, Brenda Stevens, Alicia Turner, Stephen Williams. **Indiana University**: Katy Borner (IU PI), William Barnett, Shanshan Chen, Ying Ding, Russell Duhon, Jon Dunn, Micah Linnemeier, Nianli Ma, Robert McDonald, Barbara Ann O'Leary, Mark Price, Yuyin Sun, Alan Walsh, Brian Wheeler, Angela Zoss. **Ponce School of Medicine:** Richard Noel (Ponce PI), Ricardo Espada, Damaris Torres. **The Scripps Research Institute:** Gerald Joyce (Scripps PI), Greg Dunlap, Catherine Dunn, Brant Kelley, Paula King, Angela Murrell, Barbara Noble, Cary Thomas, Michaeleen Trimarchi. **Washington University, St. Louis**: Rakesh Nagarajan (WUSTL PI), Kristi L. Holmes, Sunita B. Koul, Leslie D. McIntosh. **Weill Cornell Medical College:** Curtis Cole (Weill PI), Paul Albert, Victor Brodsky, Adam Cheriff, Oscar Cruz, Dan Dickinson, Chris Huang, Itay Klaz, Peter Michelini, Grace Migliorisi, John Ruffing, Jason Specland, Tru Tran, Jesse Turner, Vinay Varughese.

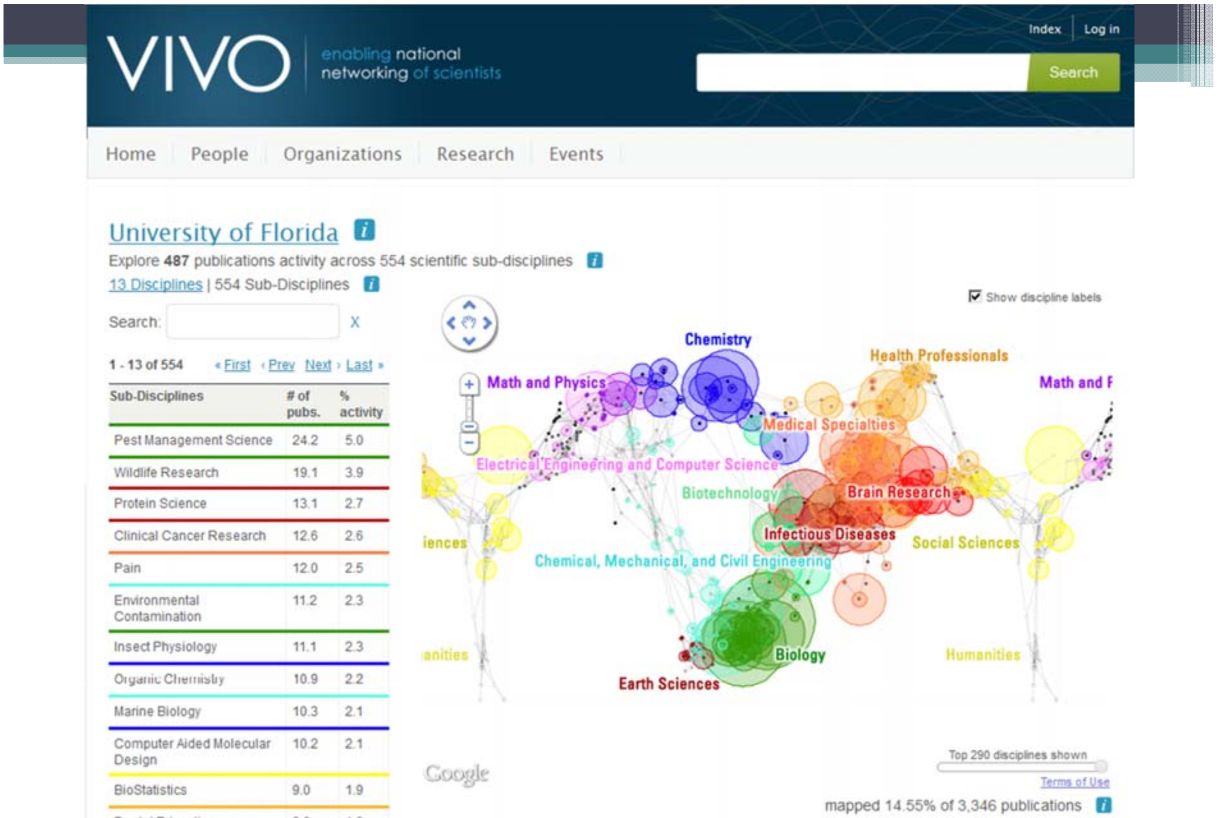



Clinical and Translational Science Institute
UNIVERSITY of FLORIDA

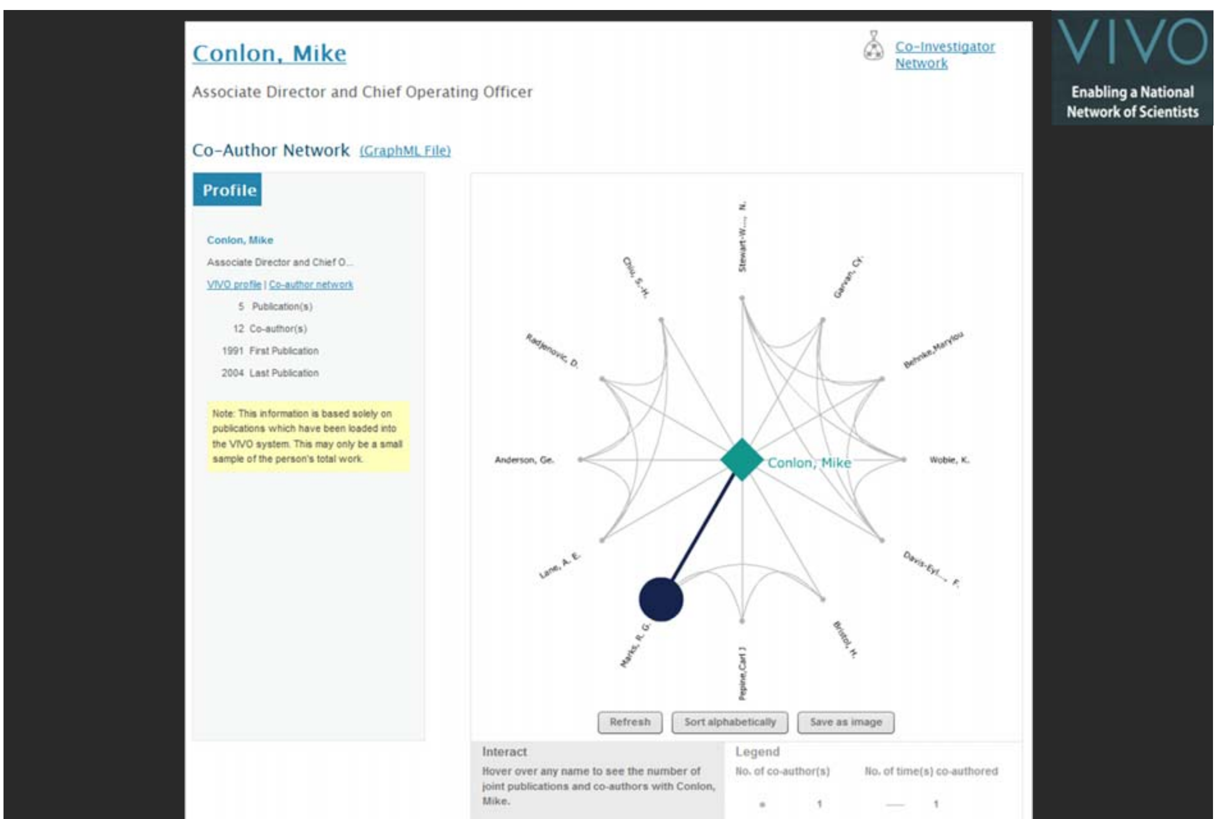
VIVO
Enabling a National Network of Scientists



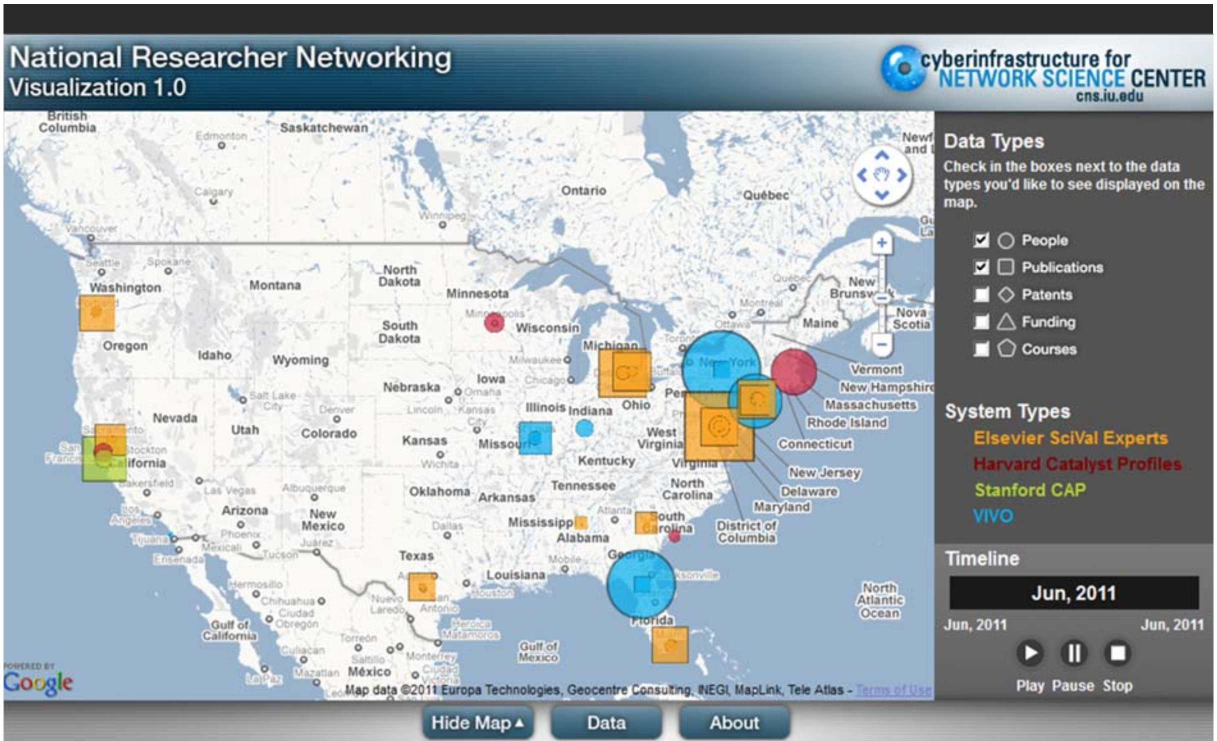
Temporal Analysis (When) Temporal visualizations of the number of papers/funding award at the institution, school, department, and people level



Topical Analysis (What) Science map overlays will show where a person, department, or university publishes most in the world of science. (in work)



Network Analysis (With Whom?) Who is co-authoring, co-investigating, co-inventing with whom? What teams are most productive in what projects?



http://nrn.cns.iu.edu

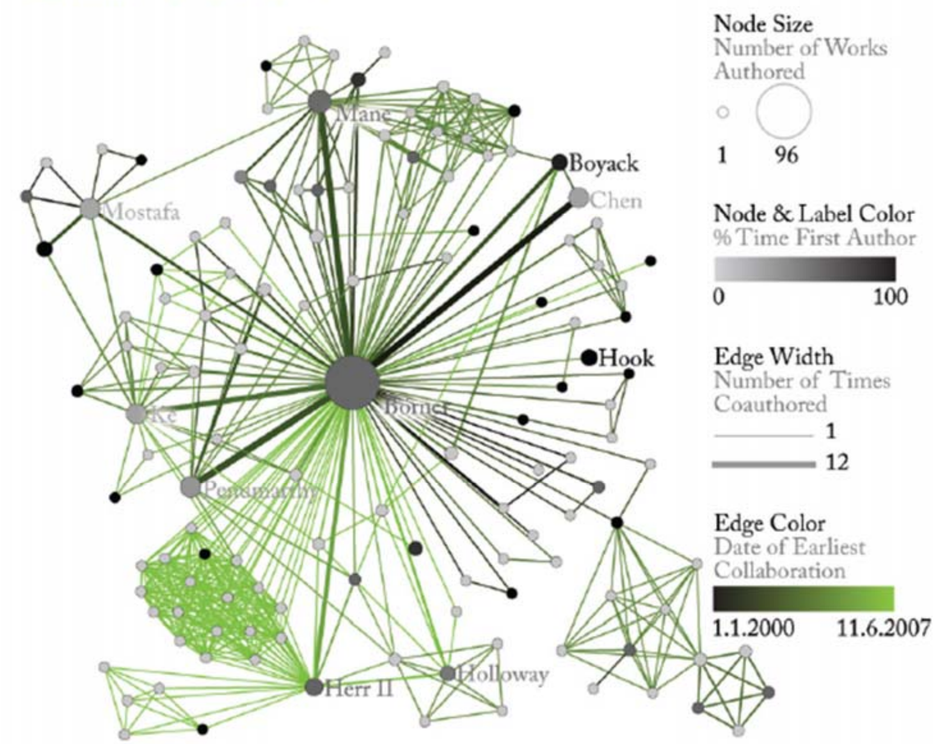
Geospatial Analysis (Where) Where is what science performed by whom? Science is global and needs to be studied globally.

VIVO
Enabling a National Network of Scientists

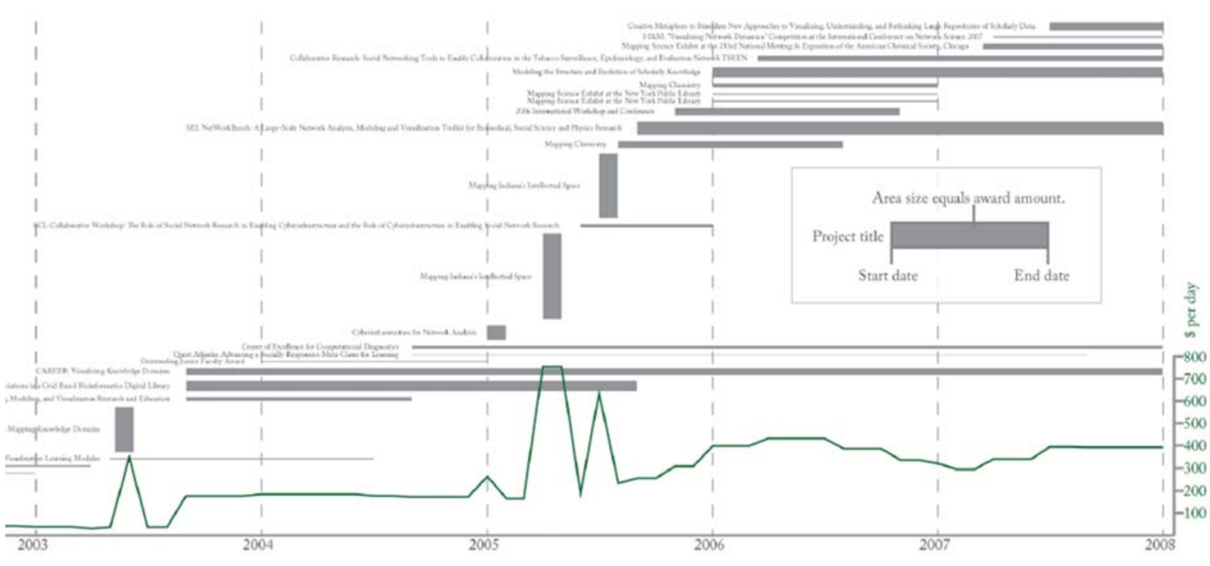
VIVO On-The-Go

Overview, Interactivity, Details on Demand come to commonly used devices and environments

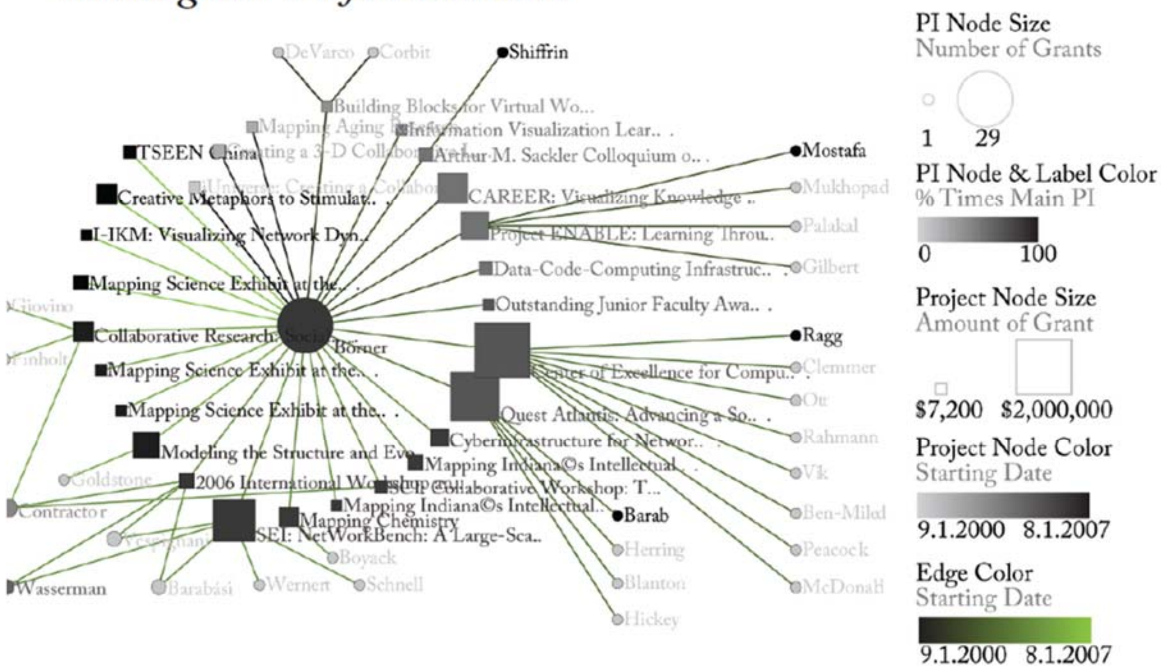
Coauthor Network



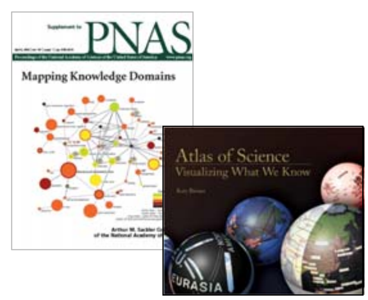
Project Timeline



Investigator-Project Network

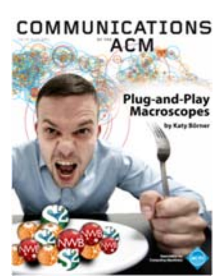


Overview



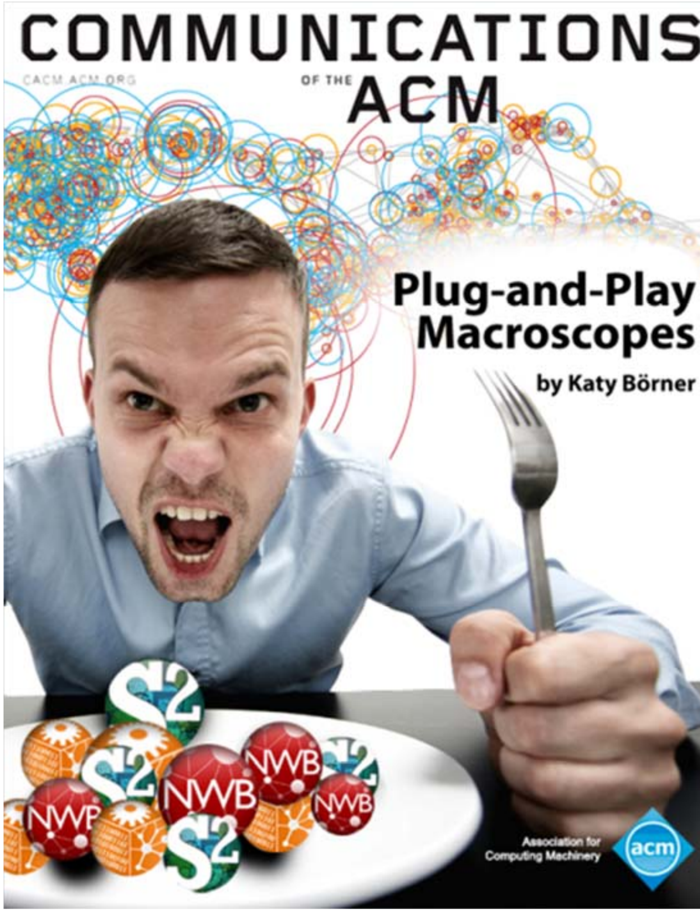
1. **Data mining and visualization research** that aims to increase our scientific understanding of the structure and dynamics of science and technology.

2. **Novel approaches and services** that improve information access, researcher networking, and research management.



3. **Data services and plug-and-play macroscope tools** that commoditize data mining and visualization.

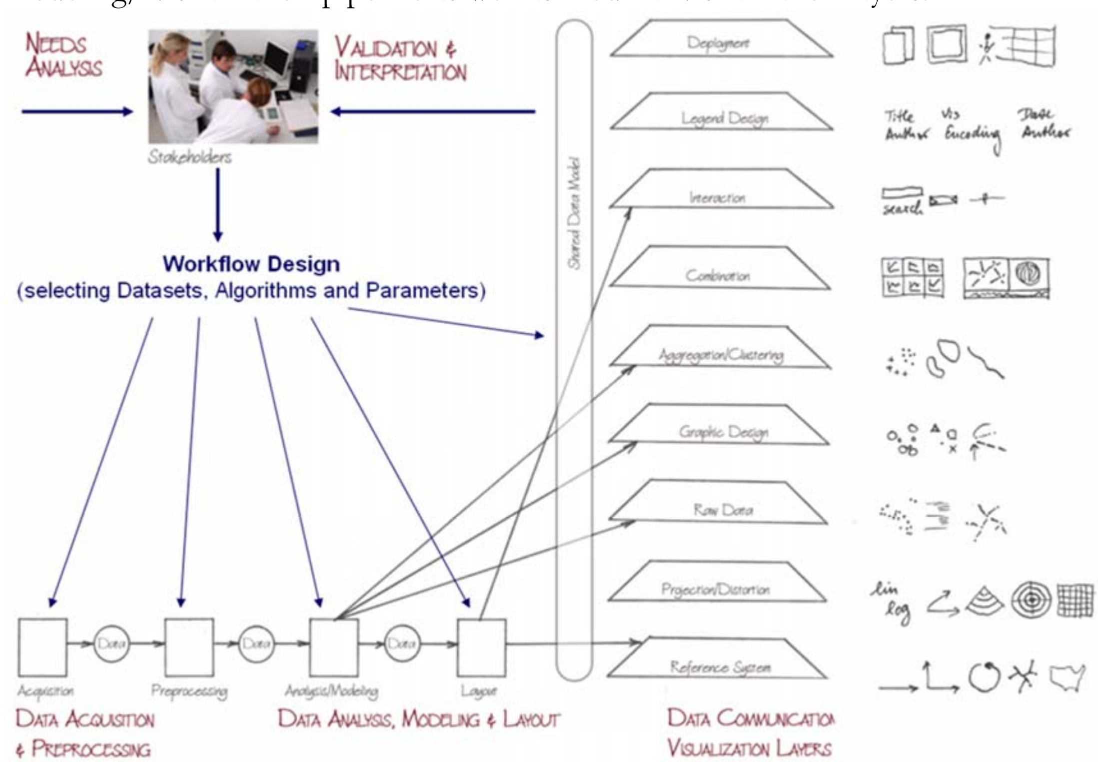
cyberinfrastructure for NETWORK SCIENCE CENTER
School of Library and Information Science | Indiana University Bloomington



Börner, Katy. (March 2011). Plug-and-Play Macroscopes. *Communications of the ACM, 54(3), 60-69.*

Video and paper are at
http://www.scivee.tv/node/27704

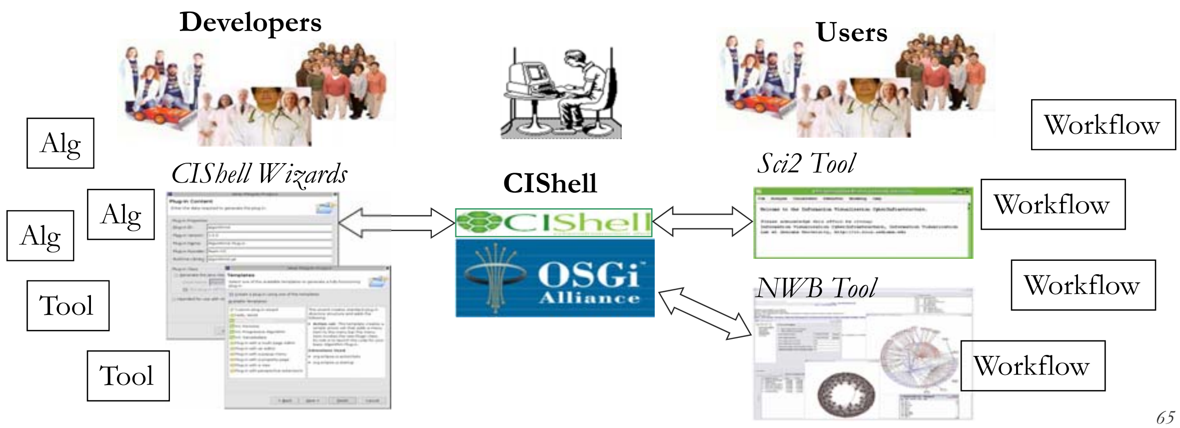
Needs-Driven Workflow Design using a modular data acquisition/analysis/ modeling/ visualization pipeline as well as modular visualization layers.



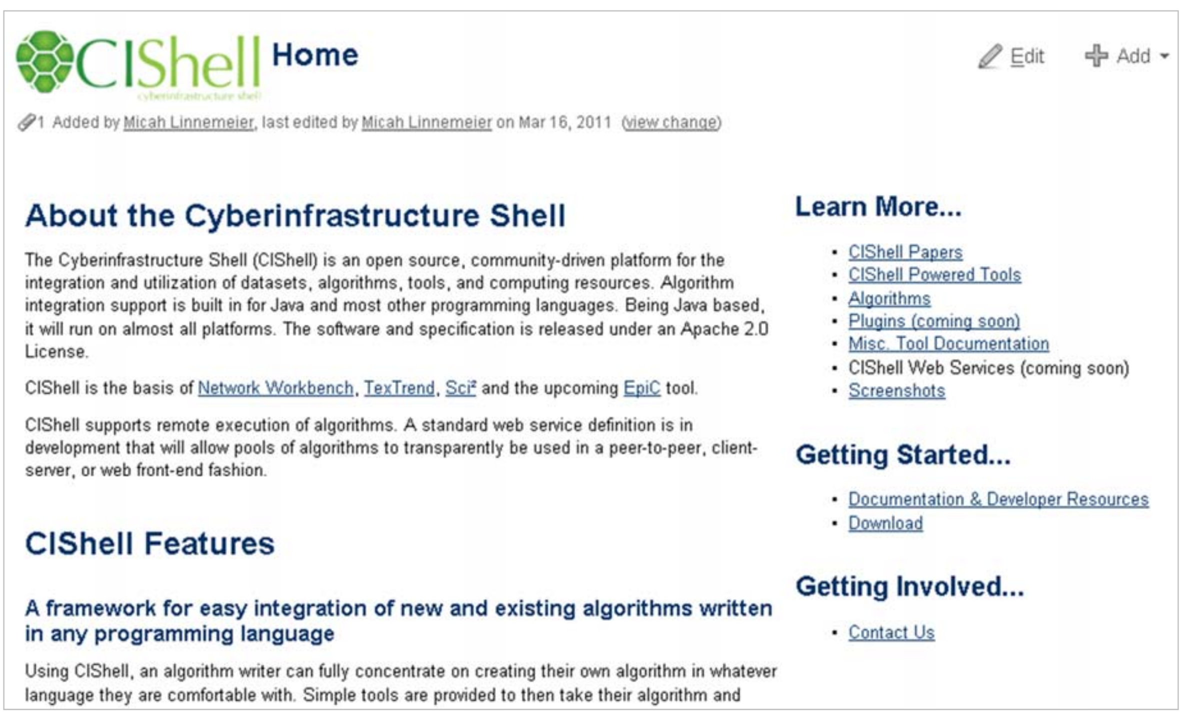
Börner, Katy (2010) Atlas of Science. MIT Press.

OSGi & CIShell

- CIShell (http://cishell.org) is an open source software specification for the integration and utilization of datasets, algorithms, and tools.
- It extends the Open Services Gateway Initiative (OSGi) (http://osgi.org), a standardized, component oriented, computing environment for networked services widely used in industry since more than 10 years.
- Specifically, CIShell provides "sockets" into which existing and new datasets, algorithms, and tools can be plugged using a wizard-driven process.



CIShell Developer Guide
(*http://cishell.wiki.cns.iu.edu*)



Home

Edit Add

1 Added by Micah Linnemeier, last edited by Micah Linnemeier on Mar 16, 2011 (view change)

About the Cyberinfrastructure Shell

The Cyberinfrastructure Shell (CIShell) is an open source, community-driven platform for the integration and utilization of datasets, algorithms, tools, and computing resources. Algorithm integration support is built in for Java and most other programming languages. Being Java based, it will run on almost all platforms. The software and specification is released under an Apache 2.0 License.

CIShell is the basis of Network Workbench, TexTrend, Sci² and the upcoming EpiC tool.

CIShell supports remote execution of algorithms. A standard web service definition is in development that will allow pools of algorithms to transparently be used in a peer-to-peer, client-server, or web front-end fashion.

CIShell Features

A framework for easy integration of new and existing algorithms written in any programming language

Using CIShell, an algorithm writer can fully concentrate on creating their own algorithm in whatever language they are comfortable with. Simple tools are provided to then take their algorithm and

Learn More...

- CIShell Papers
- CIShell Powered Tools
- Algorithms
- Plugins (coming soon)
- Misc. Tool Documentation
- CIShell Web Services (coming soon)
- Screenshots

Getting Started...

- Documentation & Developer Resources
- Download

Getting Involved...

- Contact Us

CIShell Portal (http://cishell.org)

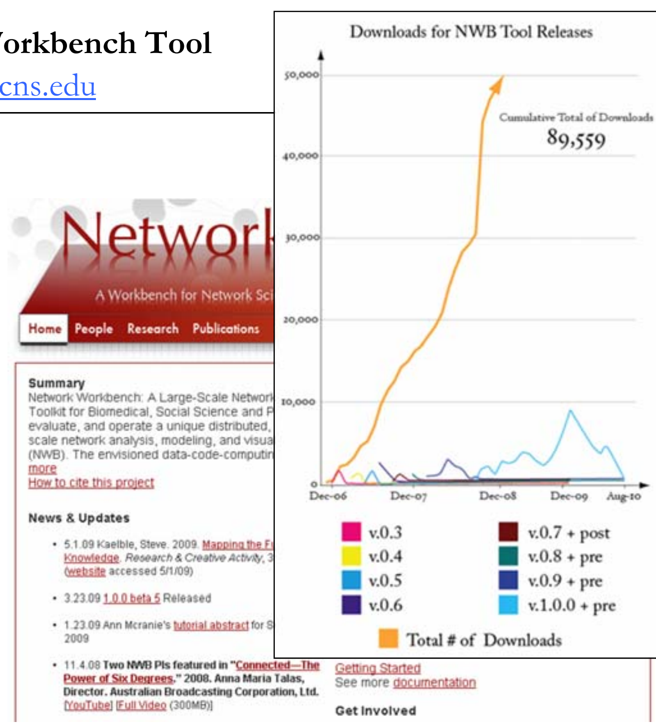
Network Workbench Tool
http://nwb.cns.edu

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 February 2009, the tool provides more 169 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

More than 50 of these plugins can be applied or were specifically designed for S&T studies.

It has been downloaded more than 65,000 times since December 2006.



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.

Computational Proteomics

What relationships exist between protein targets of all drugs and all disease-gene products in the human protein–protein interaction network?

Yildriim, Muhammed A., Kwan-Il Goh, Michael E. Cusick, Albert-László Barabási, and Marc Vidal. (2007). Drug-target Network. Nature Biotechnology 25 no. 10: 1119-1126.

Figure 2 Drug–target network (DT network). The DT network is generated by using the known associations between FDA-approved drugs and their target proteins. Circles and rectangles correspond to drugs and target proteins, respectively. A link is placed between a drug node and a target node if the protein is a known target of that drug. The area of the drug (protein) node is proportional to the number of targets that the drug has (the number of drugs targeting the protein). Color codes are given in the legend. Drug nodes (circles) are colored according to their Anatomical Therapeutic Chemical Classification, and the target proteins (rectangular boxes) are colored according to their cellular component obtained from the Gene Ontology database.

Computational Economics

Does the type of product that a country exports matter for subsequent economic performance?

C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann (2007) The Product Space Conditions the Development of Nations. Science 317, 482 (2007).

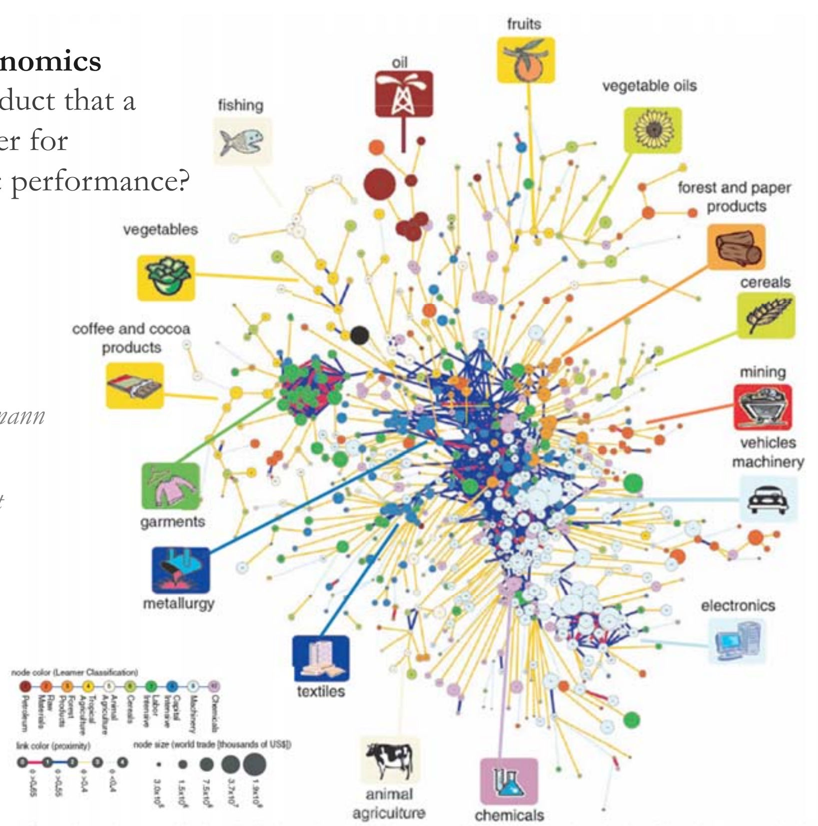
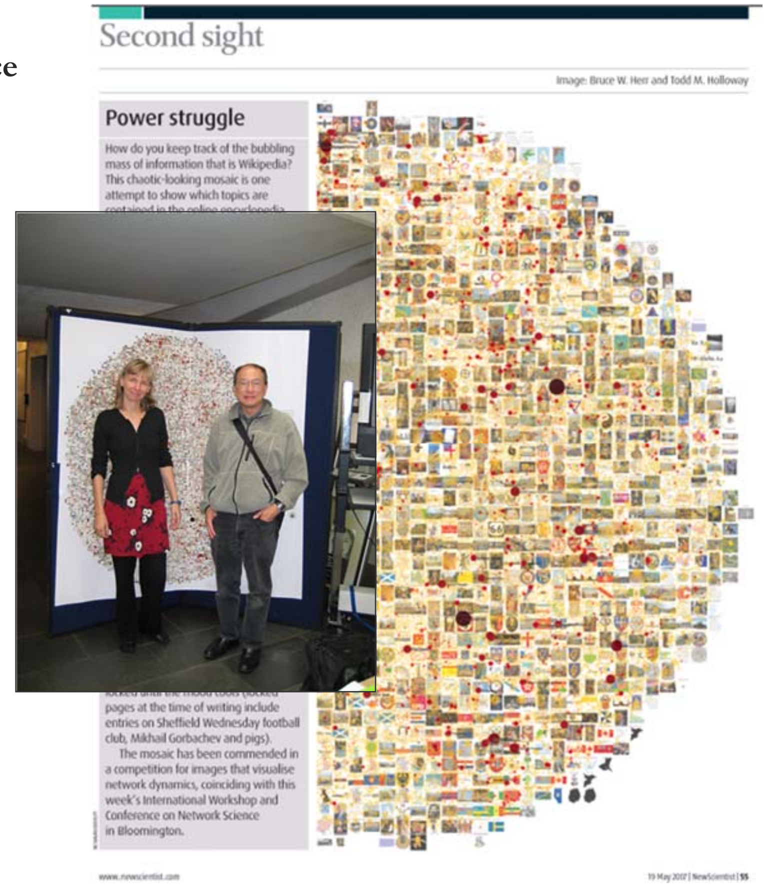


Fig. 1. The product space. **(A)** Hierarchically clustered proximity (φ) matrix representing the 775 SITC-4 product classes exported in the 1998–2000 period. **(B)** Network representation of the product space. Links are color coded with their proximity value. The sizes of the nodes are proportional to world trade, and their colors are chosen according to the classification introduced by Leamer.

Computational Social Science
Studying large scale social networks such as Wikipedia

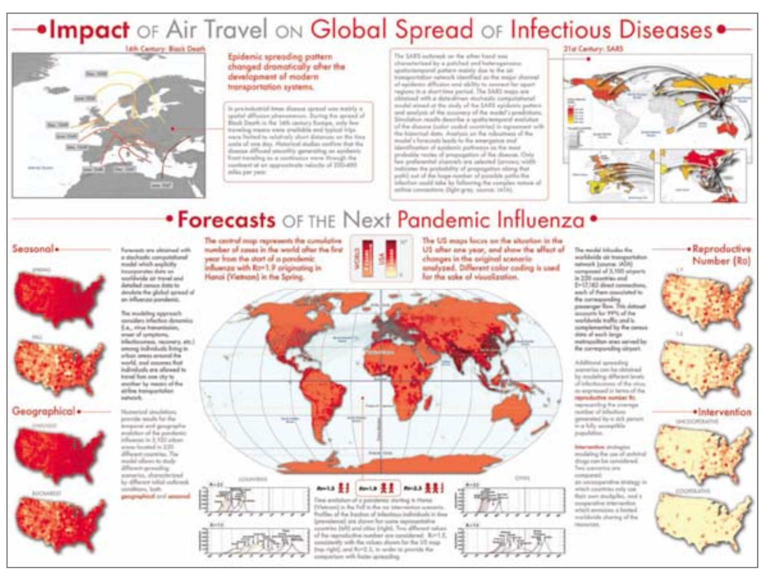
Second Sight: An Emergent Mosaic of Wikipedian Activity,
The NewScientist, May 19, 2007



Computational Epidemics
Forecasting (and preventing the effects of) the next pandemic.

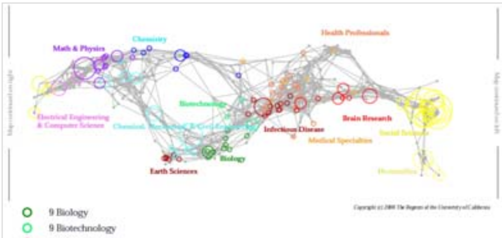
Epidemic Modeling in Complex realities, V. Colizza, A. Barrat, M. Barthelemy, A.Vespignani, Comptes Rendus Biologie, 330, 364-374 (2007).
Reaction-diffusion processes and metapopulation models in heterogeneous networks, V.Colizza, R. Pastor-Satorras, A.Vespignani, Nature Physics 3, 276-282 (2007).

Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions, V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, A.Vespignani, PloS-Medicine 4, e13, 95-110 (2007).

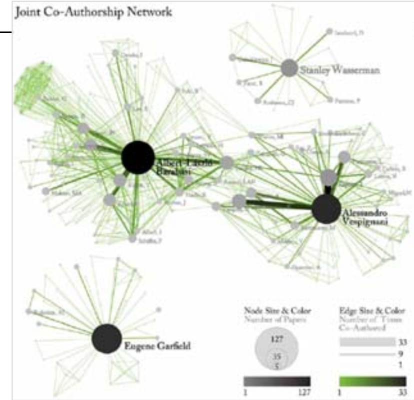


Sci² Tool – "Open Code for S&T Assessment"

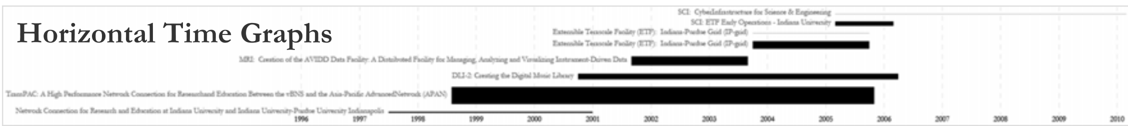
OSGi/CIShell powered tool with NWB plugins and many new scientometrics and visualizations plugins.



Sci Maps

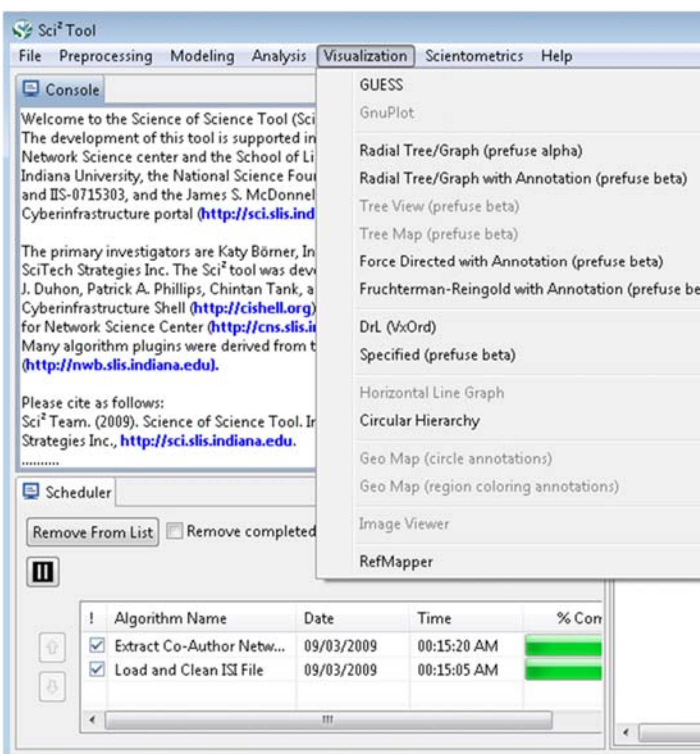


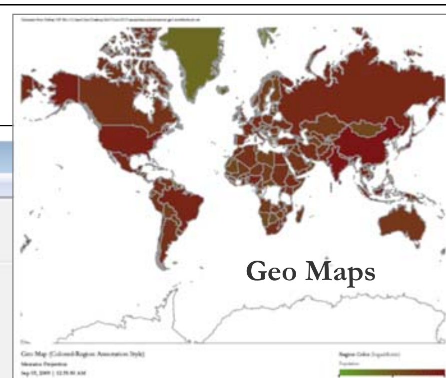
GUESS Network Vis

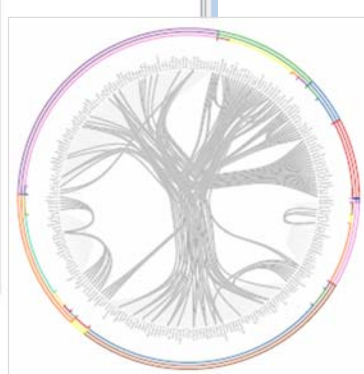
Horizontal Time Graphs


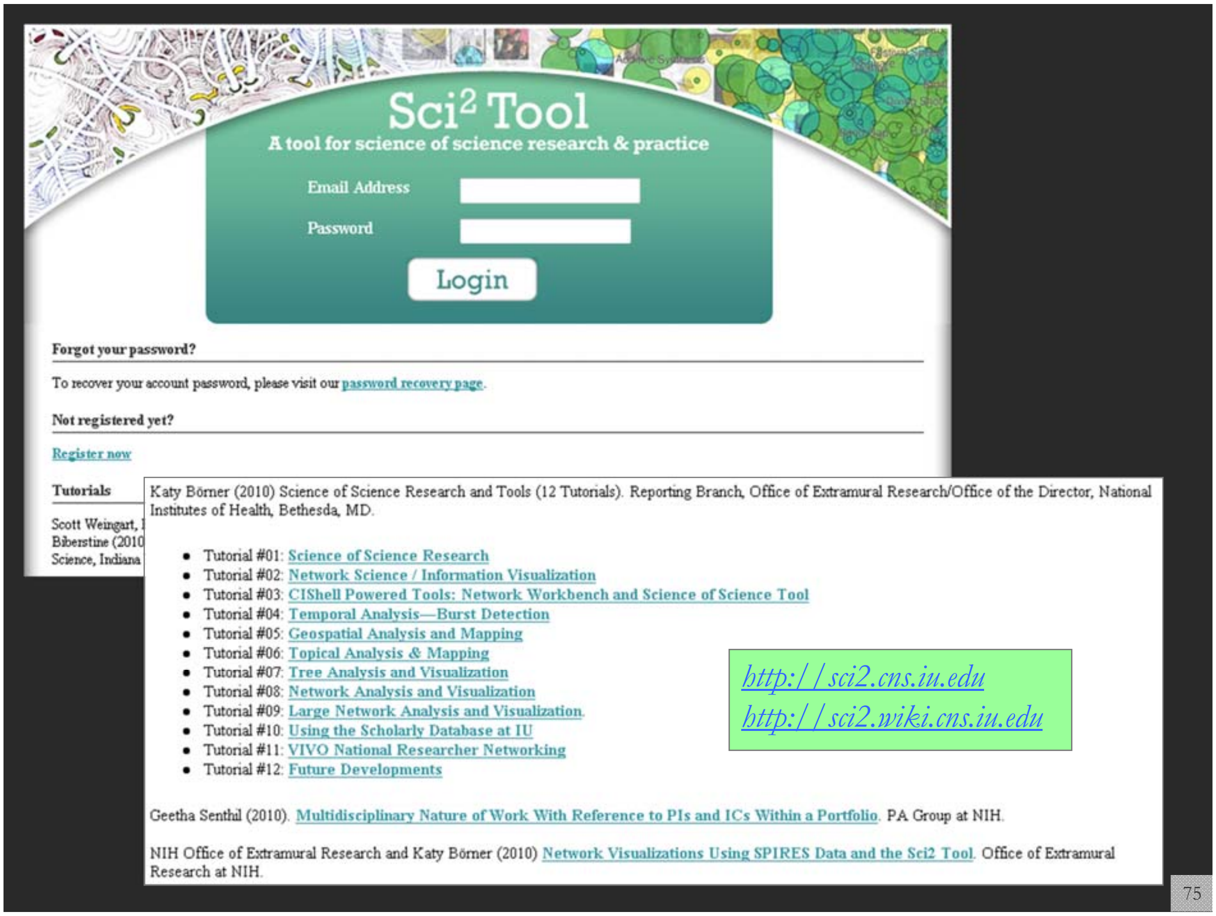
Börner, Katy, Huang, Weixia (Bonnie), Linnemeier, Micah, Duhon, Russell Jackson, Phillips, Patrick, Ma, Nianli, Zoss, Angela, Guo, Hanning & Price, Mark. (2009). Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Scholarly Database and the Network Workbench Tool. Proceedings of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

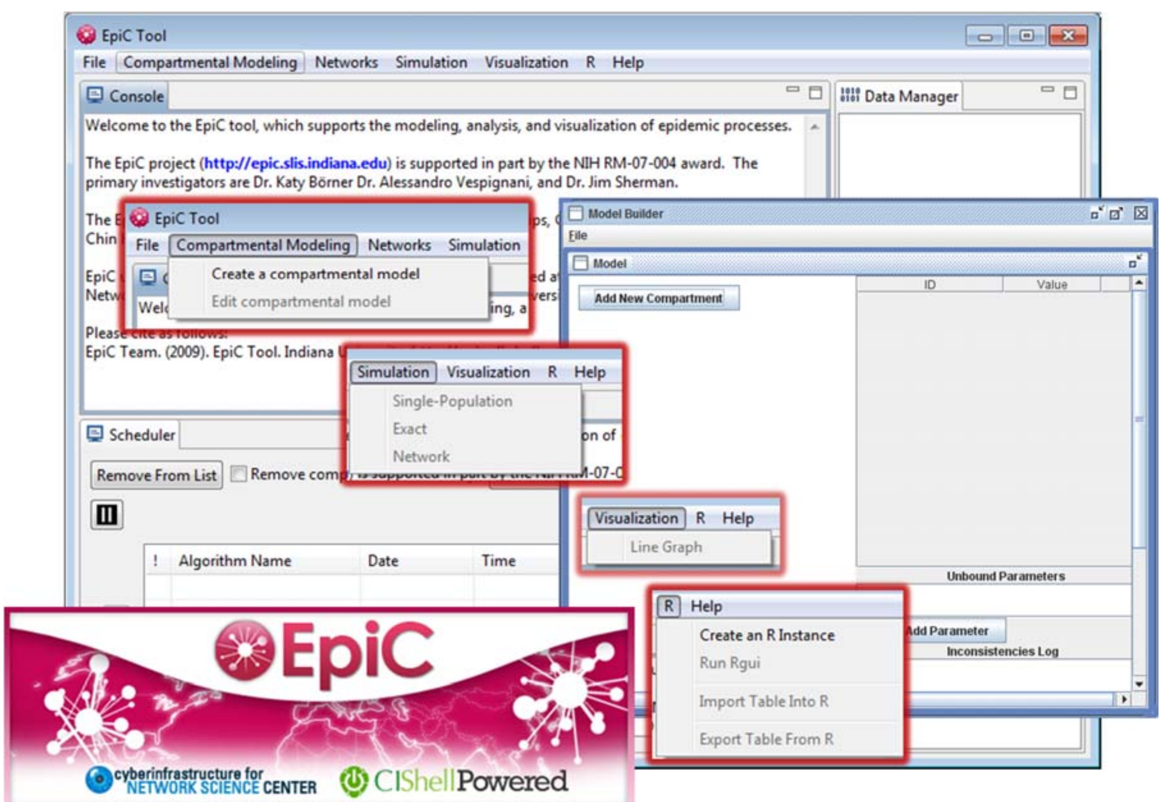
Sci² Tool Vis cont.




Geo Maps


Circular Hierarchy


Sci2 Tool
A tool for science of science research & practice
Email Address
Password
Login
Forgot your password?
To recover your account password, please visit our password recovery page.
Not registered yet?
Register now
Tutorials
Katy Börner (2010) Science of Science Research and Tools (12 Tutorials). Reporting Branch, Office of Extramural Research/Office of the Director, National Institutes of Health, Bethesda, MD.
Tutorial #01: Science of Science Research
Tutorial #02: Network Science / Information Visualization
Tutorial #03: CIShell Powered Tools: Network Workbench and Science of Science Tool
Tutorial #04: Temporal Analysis—Burst Detection
Tutorial #05: Geospatial Analysis and Mapping
Tutorial #06: Topical Analysis & Mapping
Tutorial #07: Tree Analysis and Visualization
Tutorial #08: Network Analysis and Visualization
Tutorial #09: Large Network Analysis and Visualization.
Tutorial #10: Using the Scholarly Database at IU
Tutorial #11: VIVO National Researcher Networking
Tutorial #12: Future Developments
Geetha Senthil (2010). Multidisciplinary Nature of Work With Reference to PIs and ICs Within a Portfolio. PA Group at NIH.
NIH Office of Extramural Research and Katy Börner (2010) Network Visualizations Using SPIRES Data and the Sci2 Tool. Office of Extramural Research at NIH.
http://sci2.cns.iu.edu
http://sci2.wiki.cns.iu.edu
75


EpiC Tool
File Compartmental Modeling Networks Simulation Visualization R Help
Console
Welcome to the EpiC tool, which supports the modeling, analysis, and visualization of epidemic processes.
The EpiC project (http://epic.slis.indiana.edu) is supported in part by the NIH RM-07-004 award. The primary investigators are Dr. Katy Börner Dr. Alessandro Vespignani, and Dr. Jim Sherman.
EpiC Team. (2009). EpiC Tool. Indiana
Data Manager
Create a compartmental model
Edit compartmental model
Simulation Visualization R Help
Single-Population
Exact
Network
Model Builder
Model
Add New Compartment
ID
Value
Visualization R Help
Line Graph
Scheduler
Remove From List
Algorithm Name
Date
Time
R Help
Create an R Instance
Run Rgui
Import Table Into R
Export Table From R
Unbound Parameters
Add Parameter
Inconsistencies Log
EpiC
cyberinfrastructure for NETWORK SCIENCE CENTER
CIShell Powered

OSGi/CIShell Adoption

A number of other projects recently adopted OSGi and/or CIShell:

USA

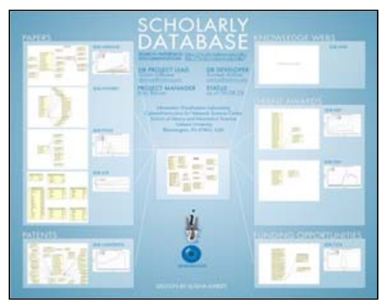
- *Cytoscape* (*http://cytoscape.org*) Led by Trey Ideker at the University of California, San Diego is an open source bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data (Shannon et al., 2002).
- *MAEviz* (*https://wiki.ncsa.uiuc.edu/display/MAE/Home*) Managed by Jong Lee at NCSA is an open-source, extensible software platform which supports seismic risk assessment based on the Mid-America Earthquake (MAE) Center research.

Europe

- *Taverna Workbench* (*http://taverna.org.uk*) Developed by the myGrid team (http://mygrid.org.uk) led by Carol Goble at the University of Manchester, U.K. is a free software tool for designing and executing workflows (Hull et al., 2006). Taverna allows users to integrate many different software tools, including over 30,000 web services.
- *TEXTrend* (*http://textrend.org*) Led by George Kampis at Eötvös Loránd University, Budapest, Hungary supports natural language processing (NLP), classification/mining, and graph algorithms for the analysis of business and governmental text corpuses with an inherently temporal component.
- *DynaNets* (*http://www.dynanets.org*) Coordinated by Peter M.A. Sloot at the University of Amsterdam, The Netherlands develops algorithms to study evolving networks.
- *SISOB* (*http://sisob.lcc.uma.es*) An Observatory for Science in Society Based in Social Models.

As the functionality of OSGi-based software frameworks improves and the number and diversity of dataset and algorithm plugins increases, the capabilities of custom tools will expand.

Computational Scientometrics Cyberinfrastructures



Scholarly Database: 25 million scholarly records
http://sdb.cns.iu.edu

James S. McDonnell Foundation

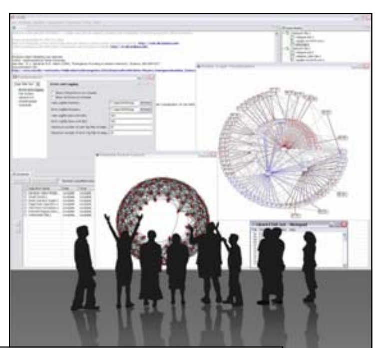
VIVO Research Networking
http://vivoweb.org

Information Visualization Cyberinfrastructure
http://iv.cns.iu.edu

Network Workbench Tool & Community Wiki
http://nwb.cns.iu.edu

Science of Science (Sci^2) Tool
http://sci2.cns.iu.edu

Epidemics Tool & Marketplace
Forthcoming



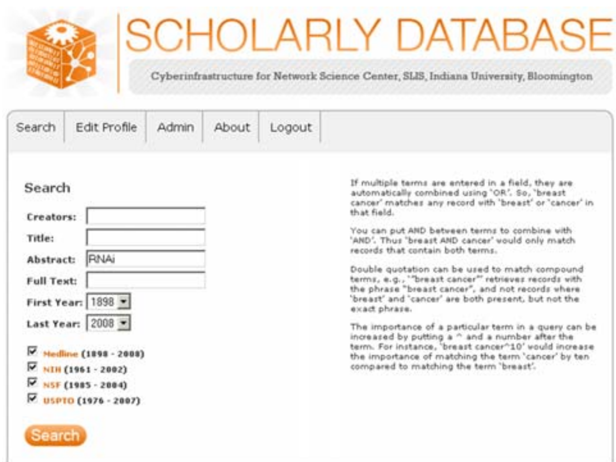
OSGi Alliance

CIShell cyberinfrastructure shell

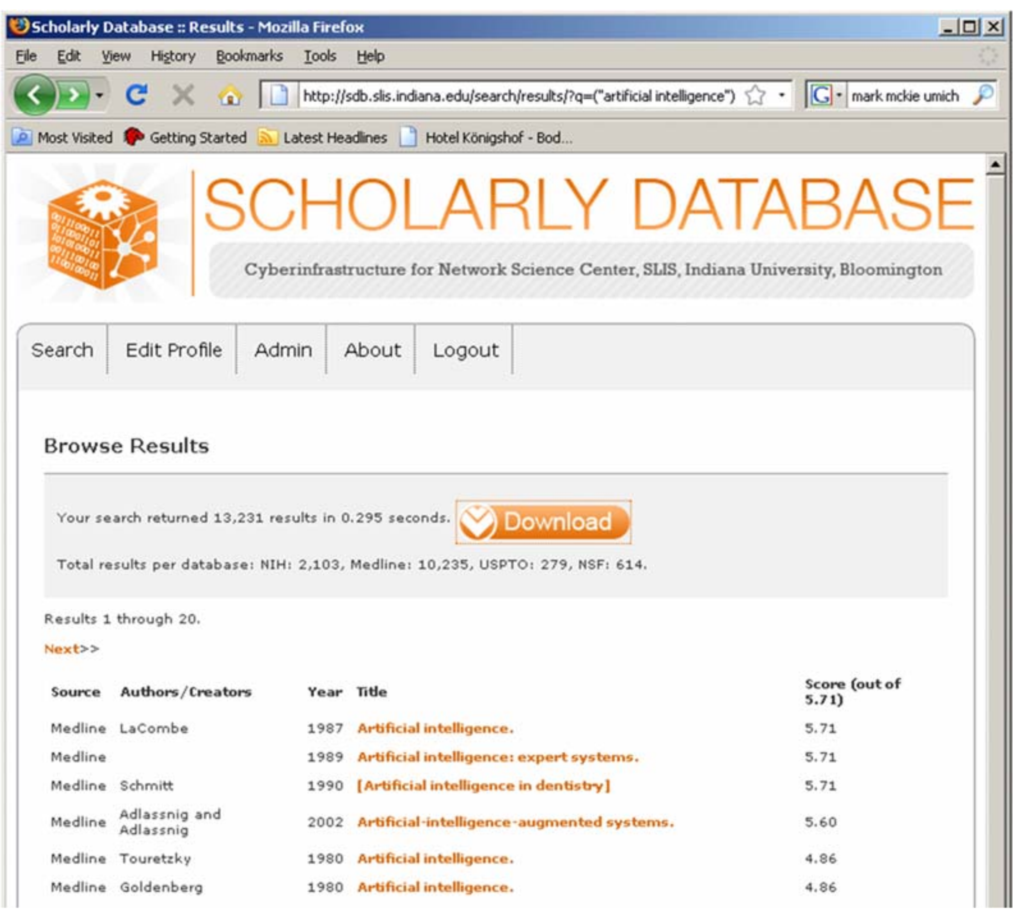
Scholarly Database at Indiana University

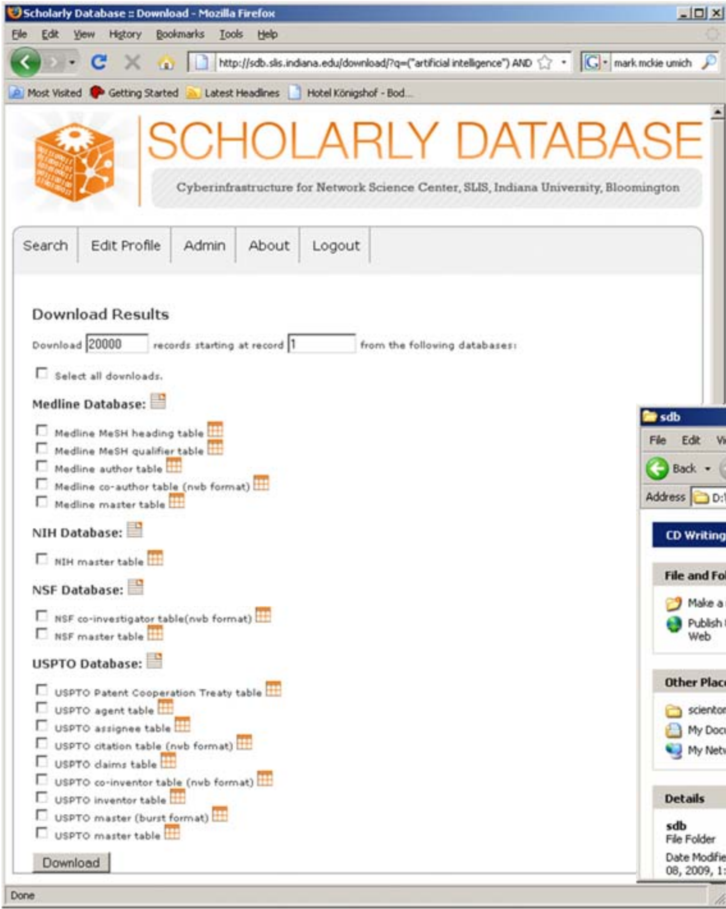
http://sdb.wiki.cns.iu.edu

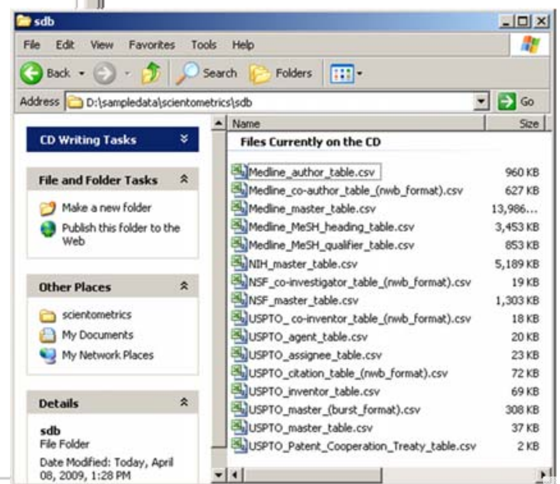
Supports federated search of 25 million publication, patent, grant records.
Results can be downloaded as data dump and (evolving) co-author, paper-citation networks.



Register for free access at http://sdb.cns.iu.edu







Since March 2009:
Users can download networks:
- Co-author
- Co-investigator
- Co-inventor
- Patent citation

and tables for
burst analysis in NWB.

CTSI Accelerating Science Core

The core provides consulting, data mining, and visualization of information on the current practice of science to accelerate science and competitive research using a network science and science mapping approach.

Findings from theory-based research on the formation of productive teams, the identification of trends and emerging ideas, and the effective communication of complex results to diverse stakeholders are used to optimize science itself.

The Accelerating Science Core provides integrative analyses of relationships in support of institutes, programs, and projects, interested to accelerate the translation of scientific results to the improvement of human health.

CTSI Accelerating Science Core—Services Offered

Evaluation & Monitoring: Impact and/or strength analysis for a lab, center (e.g., NSF STCs or NIH CTSAs), institution, or region in order to evaluate, plan, or implement research efforts. Relevant data must be provided.
$2000-$6000*

Data Compilation: The construction of a custom data set (e.g., all papers, patents, grants for a certain institution or area of research) using the Scholarly Database (http://sdb.cns.iu.edu).
$3000*

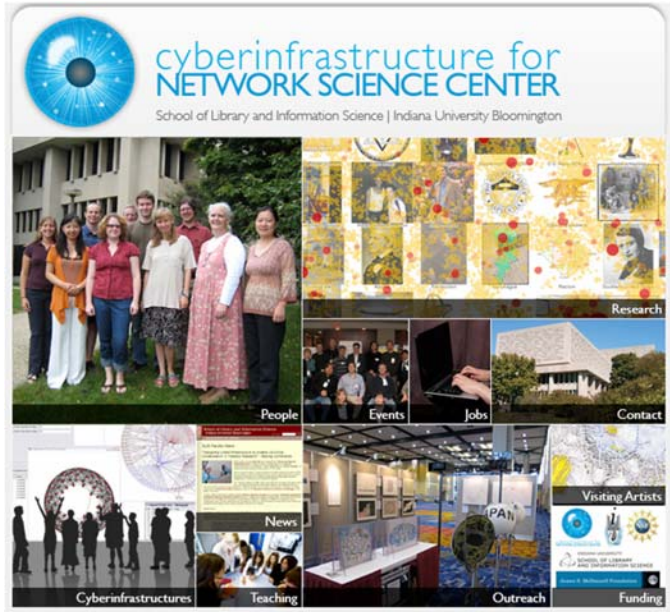
Visual Interface to Community Data: Setting up an online interactive interface similar to http://mapsustain.cns.iu.edu (relevant data must be provided to the Core).
$6000*

Accelerating Science Core—Services Offered

VIVO Researcher Networking: Design and implementation of interactive custom VIVO visualizations (see http://vivoweb.org for more info on VIVO and http://vivo.iu.edu/vis/map-of-science/BL-ARSC for a map of science visualization).
$5000*

Training and Consulting in Data Mining and Visualization:
Introduction of advanced data mining and visualization tools: 2-hour tutorials or 4-hour tutorials that also feature exemplary analyses of client data. See http://sci2.cns.iu.edu/user/documentation.php for sample slides.
$2000-$5000*

**Prices vary according to scope of work. Please contact us to request an estimate.*



All papers, maps, tools, talks, press are linked from http://cns.iu.edu

CNS Facebook: http://www.facebook.com/cnscenter
Mapping Science Exhibit Facebook: http://www.facebook.com/mappingscience