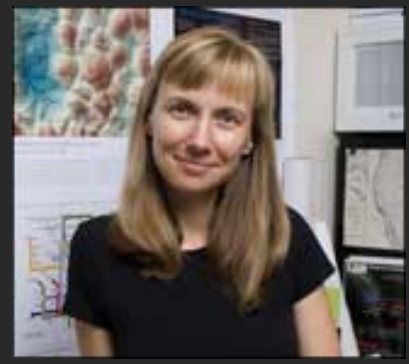
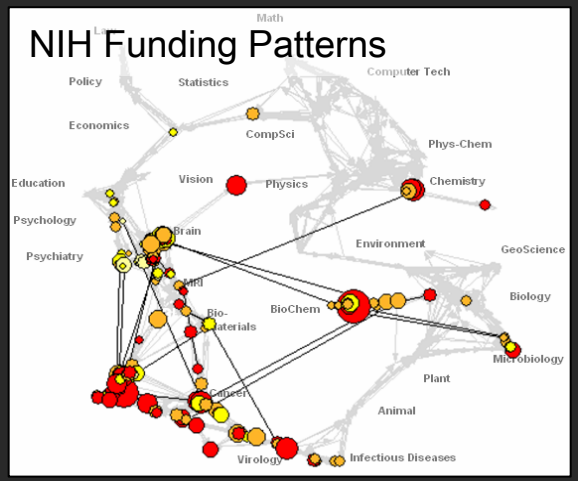
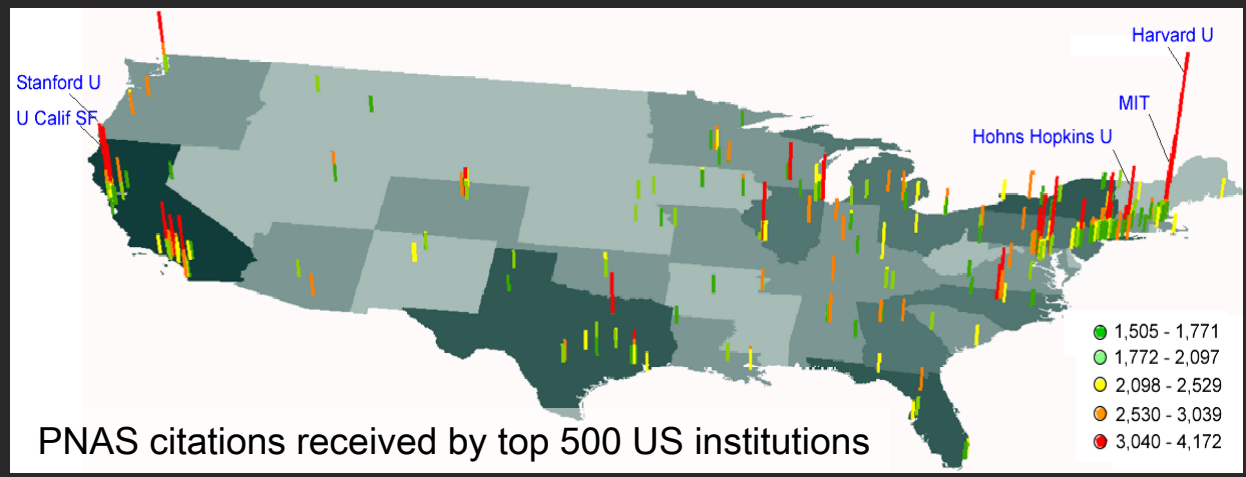
# Mapping the Structure and Evolution of Science



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

Symposium on Knowledge Discovery and Knowledge Management Tools at NIH
Natcher Conference Center, Bethesda
February 6, 2006.



# Towards Effective Knowledge Discovery and Knowledge Management Tools for NIH

**The Talk has Three Parts:**

1. Mapping Aging Research = **NIH Needs**
2. Improved Scholarly Data Collection = **NIH ToDo**
3. Computational Scientometrics = **NIH KM Tool**

# Case Study: Mapping Aging Research

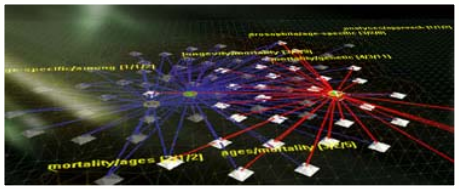
Part of this work was previously presented at the NIH Planning Workshop: "Assessments of Behavioral and Social Science Research Vitality" on Nov 9th, 2001.

**Publication:**

Boyack, Kevin W. and Börner, Katy. (2003). Indicator-Assisted Evaluation and Funding of Research: Visualizing the Influence of Grants on the Number and Citation Counts of Research Papers, *Journal of the American Society of Information Science and Technology*, Special Topic Issue on Visualizing Scientific Paradigms, 54(5):447-461.

**Funding Support:**

Börner, Katy & Kevin W. Boyack. Mapping Aging Research. NIH Evaluation Express award on "Investigating Methodologies for Assessing Research Excellence in the Social and Behavioral Sciences" to Dr. Richard Suzman, BSR, NIA, Aug.-Dec. 2001

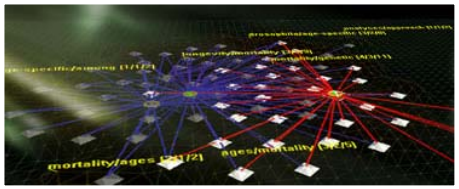


# Mapping Aging Research: Client Needs

## NIA's **Most Actionable Questions**

1. What are the major research areas, experts, institutions, regions, grants, publications, journals in aging research?
2. Which areas are most insular?
3. What are the main connections for each of our areas?
4. What is the relative speed of areas? Which areas are the most dynamic/static?
5. What new research areas are evolving?
6. Citation rate of grantees compared with other S&B&Life scientists?
7. How influential have NAS reports been?
8. How does NIA funding influence the number and quality of publications?
9. Impact of NIA funded research on other fields as measured by press coverage, Congressional presentations, etc.?
10. How can we measure progress?**

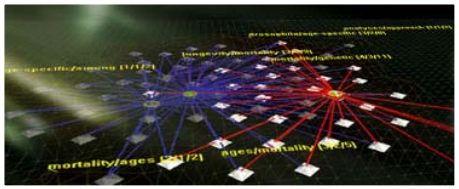
** Needs clarification.



# Mapping Aging Research: Client Needs

NIAs Indices of an **Individual Researcher's Productivity**

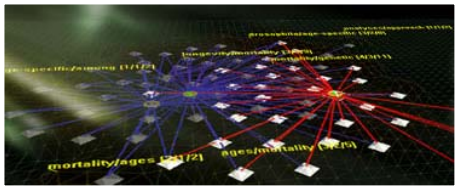
- Success/Discovery stories.
- Publications in top science journals such as *Science, Nature, PNAS*.
- Most significant publications in disciplinary journals or NIA volumes and publications.
- Citations of papers.
- Recent major honors received by funded investigators such as NAS membership, Nobel prizes, etc.
- Media coverage of grant research.
- Recent public policy activities by funded investigators. Impact on policy and practice.
- Export of ideas to other fields.



# Mapping Aging Research: Client Needs

NIAs Indices of the **Vitality of Research Field**

- Average age of researchers in a field.
- Employment of post docs in the field.
- Employment of post docs in the private sector.
- Rapidity of publication of new results in leading journals in the field.
- Speed of promotion and tenure.
- Concentrations at Centers of Excellence; size of project teams over time; number of projects with multiple institutions.



# Mapping Aging Research: Datasets Used

## Publication Data

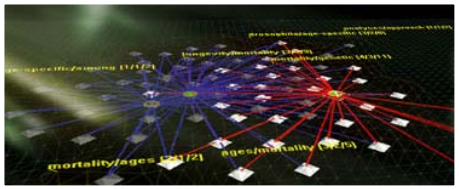
- Medline publication data (Query for 'aging' in title, abstract or MeSH, 1996-2000) resulting in 28,700+ articles. (Query for selected 37 journals with >= 10 articles represented in NIA publication data) resulted in 101,800+ articles.

  Concerns: No citation links. Medline does not cover all relevant journals such as Journal of Marriage and the Family, Research on Aging, American Economic Review, Journal of Cross-cultural Gerontology, Journal of Family Issues, Aging Neuropsychology & Cognition.

- ISI publication data incl. citation links. 2,296 papers by 32 merit scholars, 26,880 papers that cited those 2,296, and the 14,828 ISI-indexed papers cited in (or by) the 2,296 for 1981-June 2001.

  Concerns: For the demonstration work, we received no abstracts or keywords.

**General Concerns: No data on books, media coverage, awards, etc.**



# Mapping Aging Research: Datasets Used

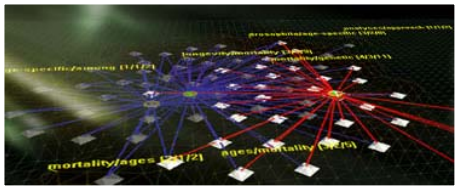
## NIA/BSR Data

- NIA grant data (PI names, institutions, years, award amounts, titles, abstracts, MeSH terms etc. for 1975-2001) resulting in 33,448 (sub)grants

  Concerns: Different name formats, changing BAS codes. Most grants are multi-year grants with several subcontracts, changing names & titles, etc.

- BSR accomplishment reports (CFA, DPE, HCO, PHB, PSP) resulting in 4,500+ publications

  Concerns: grant – publication linkage is not unique, grants by non-PI's are often missing, many duplicates.



# Mapping Aging Research: Datasets Used

**Carolyn M. Aldwin** **University of California -- Davis**
**AG13006** **Mental and Physical Health Trajectories in Adulthood**

**Bibliography**


Spiro, A. III, Aldwin, C. M., Ward, K. D., & Mroczek, D. K. (1995). Personality and the incidence of hypertension among older men: Longitudinal findings from the Normative Aging Study. *Health Psychology, 14,* 563-569.

Aldwin, C. M., Ober, B. A., & Levenson, M. R. (1996). The psychology of aging. In D. B. Reuben, T. T. Yoshikawa, R. W. Besdine, & B. R. Popecki (Eds.). *Geriatrics Review Syllabus (GRS-III)* (pp. 24-29). New York: American Geriatrics Society.

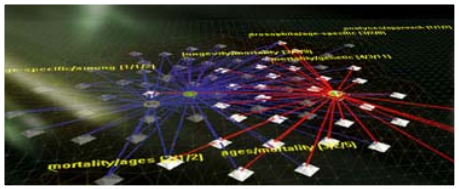

**NIA Behavioral and Social Research Program**
**DEMOGRAPHY AND POPULATION EPIDEMIOLOGY UNIT**
**Partial Grantee Bibliography**

***Publications, 1992-present, in top journals in science (Science, Nature, PNAS) or medicine (NEJM, Lancet, JAMA, BMJ, Cell).***


Bruce M. Alberts, Francisco J. Ayala, David Botstein, Ellen Frank, Edward W. Holmes, Ronald D. Lee, Eduardo, R. Macagno, Philippa Marrack, Suzanne Oparil, Stuart H. Orkin, Arthur H. Rubenstein, Carolyn W. Slayman, P., Frederick Sparling, Larry R. Squire, Peter H. von Hippel, and Keith R. Yamamoto. "Grant Review Process. Proposed Changes for NIH?s Center for Scientific Review" Science 1999 July 30; 285: 666-667. (in Policy Forum)

Bosma H, Marmot MG, Hemingway, H et al. 1997. Low Job Control and Risk of Coronary Heart Disease in Whitehall II (Prospective Cohort) Study. BMJ 314(7080): 558-65.

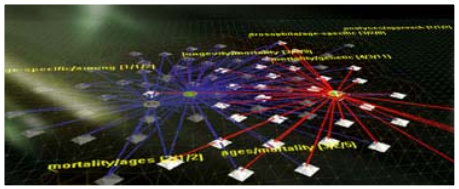

**Figure 2. Example extracts (PSP and DPE, respectively) from BSR accomplishment reports.**



# Mapping Aging Research: Data Analysis
## Top Journals

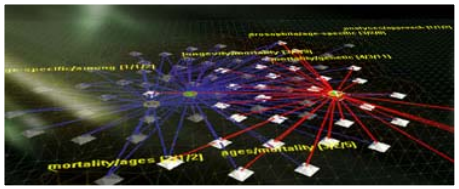
**Table 2. Top journals represented in the BSR publication data (* 3-year impact factor)**

| | Number | Medline | ISI | IF9599 |
|---|---|---|---|---|
| Gerontologist | 165 | yes | yes | 1.73 |
| Journals of Gerontology Series B - Psychological Sciences and Social Sciences | 160 | yes | yes | 1.33 |
| Psychology and Aging | 102 | yes | yes | 2.03 |
| Journal of the American Geriatrics Society | 81 | yes | yes | 2.66 |
| Journals of Gerontology Series A - Biological Sciences and Medical Sciences | 75 | yes | yes | 1.02 |
| Journal of Marriage and the Family | 48 | no | yes | 1.43 |
| American Journal of Public Health | 46 | yes | yes | 3.23 |
| Research on Aging | 40 | no | yes | 0.61 |
| American Journal of Epidemiology | 40 | yes | yes | 3.86 |
| Journal of Aging and Health | 37 | yes | yes | 0.85* |
| Demography | 33 | yes | yes | 1.70 |
| American Economic Review | 30 | no | yes | 1.77 |
| Health Psychology | 29 | yes | yes | 2.67 |
| JAMA - Journal of the American Medical Association | 28 | yes | yes | 9.44 |
| International Journal of Aging & Human Development | 23 | yes | yes | 0.48 |
| Experimental Aging Research | 21 | yes | yes | 0.60 |
| Social Science & Medicine | 21 | yes | yes | 1.30 |
| Medical Care | 21 | yes | yes | 2.24 |
| Journal of Family Issues | 20 | no | yes | 0.86 |
| Journal of Clinical Epidemiology | 20 | yes | yes | 1.80 |
| Aging Neuropsychology and Cognition | 19 | no | yes | 0.86* |
| Journal of Human Resources | 19 | no | yes | 1.17 |
| Annals of Behavioral Medicine | 19 | yes | yes | 1.47* |
| Journal of Health And Social Behavior | 19 | yes | yes | 2.53 |
| Neurology | 18 | yes | yes | 4.80 |
| Journal of Applied Gerontology | 17 | no | yes | 0.38 |
| Psychosomatic Medicine | 17 | yes | yes | 2.94 |
| Archives of Internal Medicine | 17 | yes | yes | 5.14 |
| Science | 16 | yes | yes | 23.84 |
| Journal of General Internal Medicine | 16 | yes | yes | 1.80 |



# Mapping Aging Research: Data Analysis

## # NIA Grants and Average Grant Amounts per Year

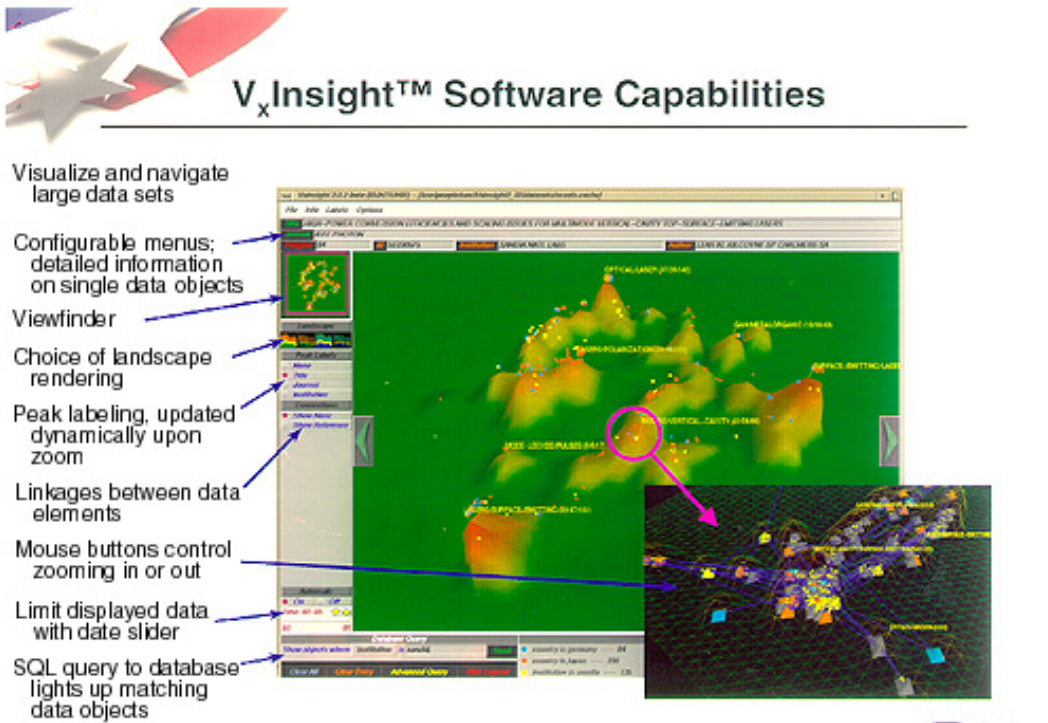


# Mapping Aging Research: Visualization

Subsequent slides show visualizations of

- BSR grants and publications.
- Show impact of funding on number of publications over time.
- Grant input and publication output for U Michigan and Duke U.
- Author supplied linkage patterns for one grant and its publications.

VxInsight is used for interactive visualization.




Sandia National Laboratories

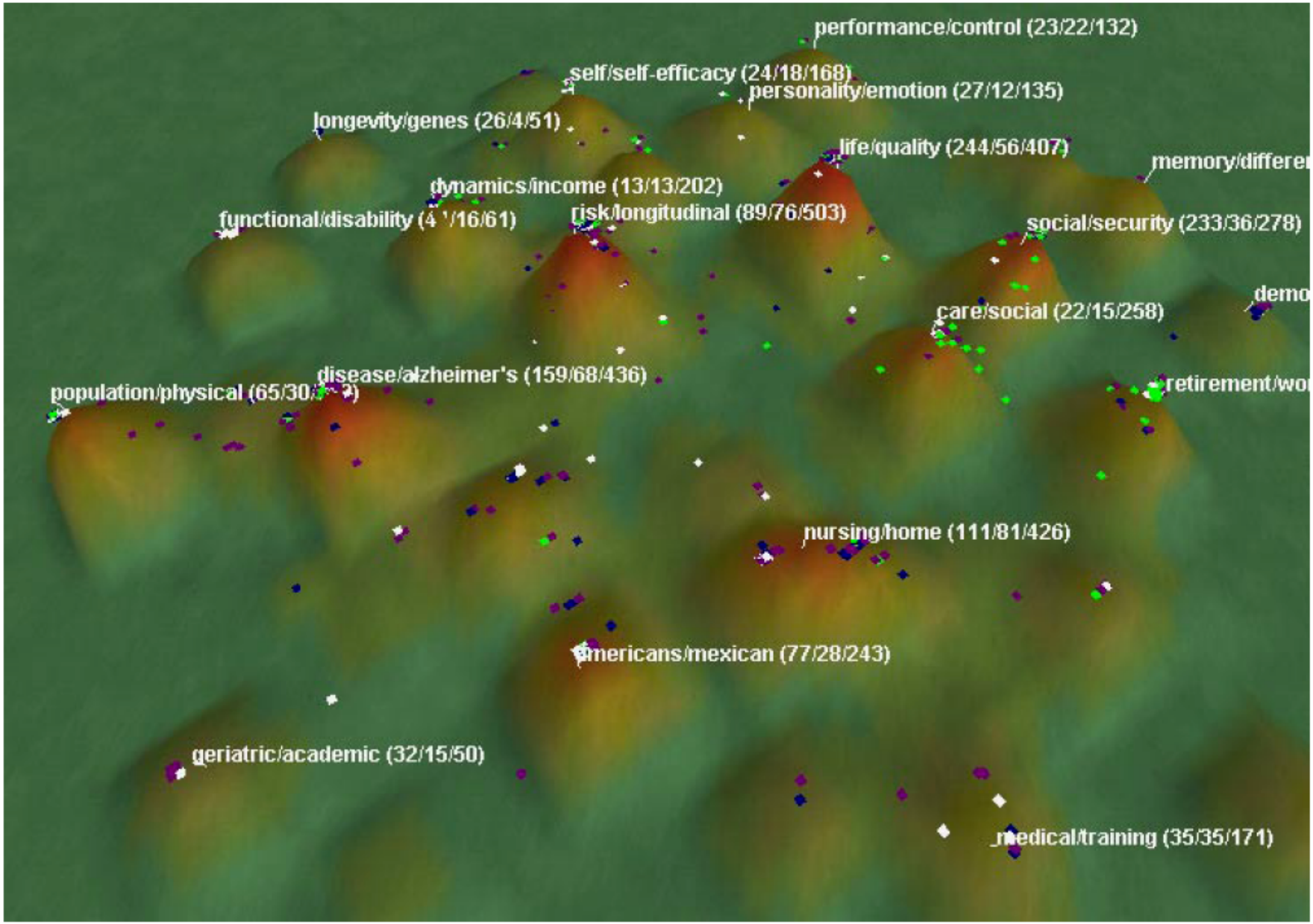
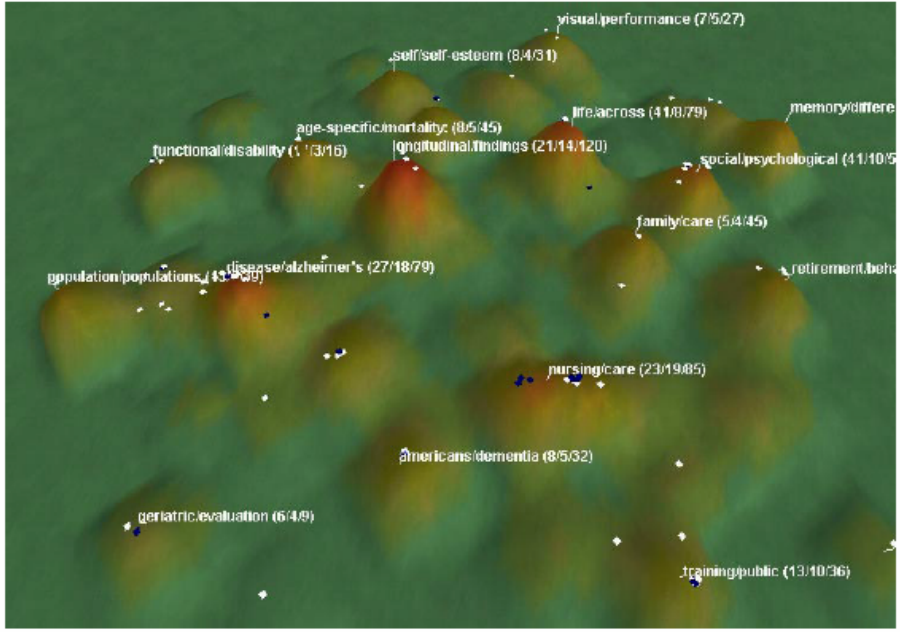
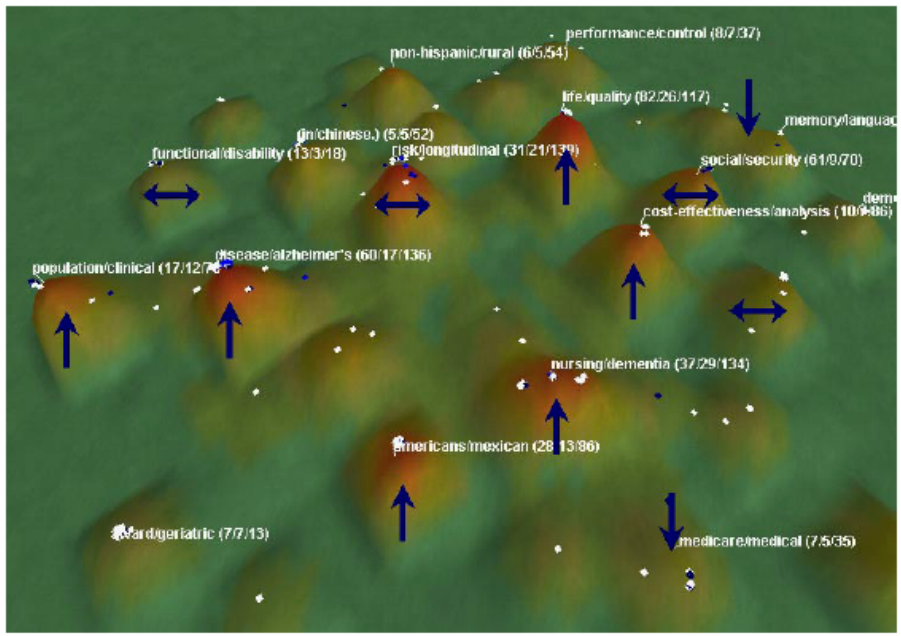



**Figure 3. Map of BSR grants and publications using MeSH words as a basis for clustering. The height of each peak is proportional to the number of documents in the peak. Labels show the two most common words in the titles of the documents in each cluster (peak). Query markers for papers with certain words in either the title or MeSH words are shown as colored dots: disability – white, retirement – light gray, economics – dark gray, demography(ics) - black.**




1995-1996




1999-2000

**Figure 4. BSR map in two different time periods showing the impact of funding on the number of articles (size of peaks). Small grants are shown by light colored markers on the terrain. Large grants are shown by dark colored markers. Significant changes in the number of publications over time by peak are shown by dark arrows.**

**Table 4. Grant and citation data for 30 institutions funded by NIA.**

| | Grants | | Awardee papers | | Citing papers | |
|---|---|---|---|---|---|---|
| | 1993-1997 | | 1995-1999 | | 1995-1999 | |
| Institution | # Grant Records | $M | # Papers | Cites / paper | # Papers | Cites / paper |
| University of Michigan | 373 | 89.9 | 81 | 11.73 | 254 | 1.76 |
| Duke University | 314 | 69.1 | 71 | 13.04 | 198 | 1.79 |
| Johns Hopkins University | 252 | 63.3 | 2 | 21.00 | 189 | 1.40 |
| University of California at San Diego | 225 | 55.1 | 4 | 5.25 | 82 | 1.23 |
| University of Washington | 267 | 53.7 | 8 | 7.63 | 161 | 1.47 |
| University of California at San Francisco | 218 | 49.2 | 29 | 10.76 | 184 | 1.58 |
| University of Southern California | 230 | 48.8 | 19 | 10.05 | 113 | 1.88 |
| University of Texas (system) | 302 | 43.9 | 8 | 12.63 | 207 | 1.45 |
| Case Western Reserve University | 266 | 40.8 | 22 | 4.77 | 90 | 1.67 |
| Washington University | 251 | 36.7 | 3 | 8.00 | 83 | 1.67 |
| University of California at Los Angeles | 165 | 31.7 | 10 | 9.70 | 228 | 1.61 |
| Massachusetts General Hospital | 169 | 31.4 | | | 46 | 1.17 |
| University of Pennsylvania | 161 | 30.2 | 23 | 9.83 | 122 | 1.44 |
| University of Pittsburgh | 152 | 29.3 | 6 | 11.50 | 135 | 1.54 |
| Columbia University | 138 | 29.0 | 5 | 1.60 | 121 | 1.57 |
| Rush Presbyterian – St. Luke's Medical Center | 119 | 27.8 | | | 33 | 1.97 |
| University of Kentucky | 174 | 27.6 | | | 37 | 1.41 |
| Boston University | 146 | 27.2 | 23 | 5.17 | 123 | 1.59 |
| New York University | 165 | 27.0 | | | 43 | 1.47 |
| Harvard University | 154 | 25.5 | 19 | 13.16 | 321 | 1.60 |
| Indiana University | 105 | 25.1 | 30 | 12.43 | 80 | 1.61 |
| Mayo Clinics & Mayo Foundation | 101 | 24.8 | | | 30 | 1.27 |
| University of Maryland | 158 | 24.6 | 2 | 2.00 | 90 | 1.70 |
| Mt. Sinai School of Medicine | 150 | 23.6 | | | 71 | 1.34 |
| University of Wisconsin | 108 | 21.9 | 3 | 19.67 | 133 | 1.65 |
| University of Colorado | 148 | 21.7 | 3 | 26.33 | 67 | 1.76 |
| Pennsylvania State University | 99 | 21.2 | 12 | 4.67 | 115 | 1.54 |
| Stanford University | 134 | 21.0 | 39 | 14.36 | 180 | 1.63 |
| University of Alabama | 121 | 19.0 | 12 | 6.33 | 95 | 1.44 |
| Brigham & Women's Hospital | 97 | 17.8 | | | 53 | 1.64 |
| | | | | | | |
| **TOTALS** / *AVERAGES* | **5462** | *35.3* | **434** | *10.85* | **3684** | *1.58* |

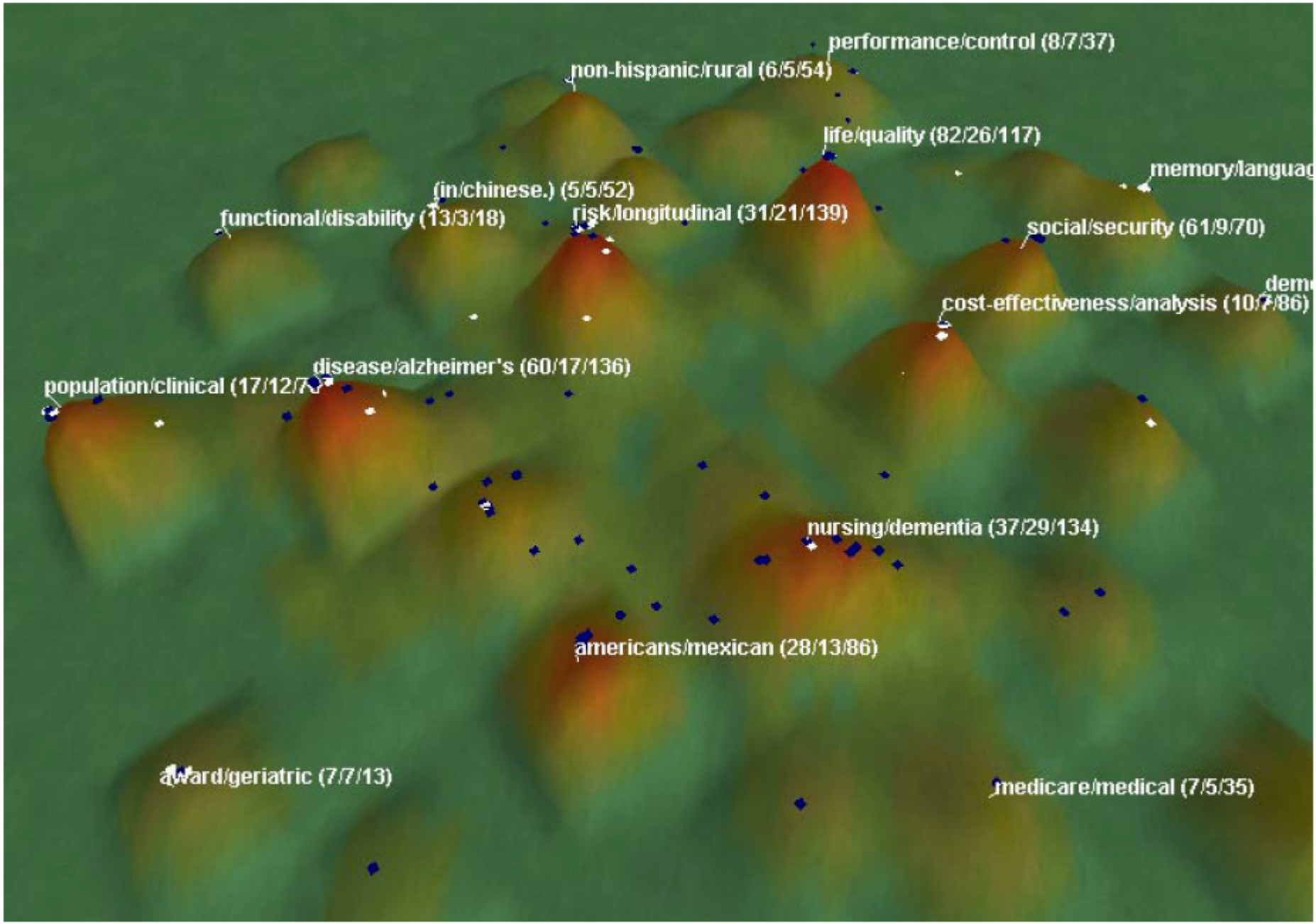
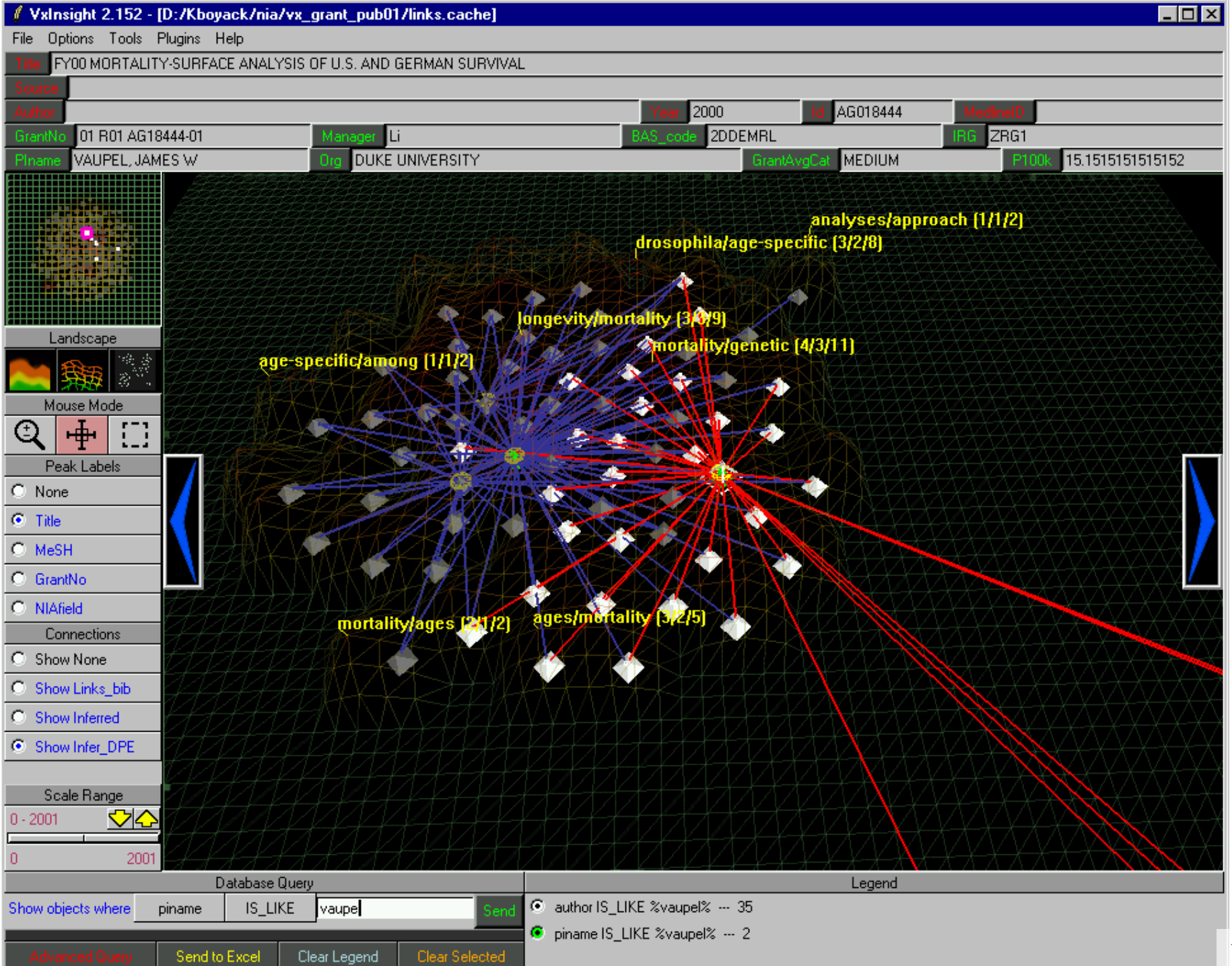
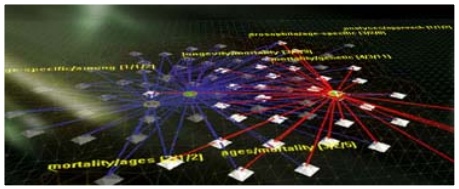



**Figure 5. NIA grants awarded to and BSR-related publication output of the University of Michigan (dark markers) and Duke University (white markers) in the BSR domain during the 1999-2000 time period.**


VxInsight 2.152 - [D:/Kboyack/nia/vx_grant_pub01/links.cache]
File Options Tools Plugins Help
Title FY00 MORTALITY-SURFACE ANALYSIS OF U.S. AND GERMAN SURVIVAL
Source
Author
Year 2000
Id AG018444
MedlineID
GrantNo 01 R01 AG18444-01
Manager Li
BAS_code 2DDEMRL
IRG ZRG1
PIname VAUPEL, JAMES W
Org DUKE UNIVERSITY
GrantAvgCat MEDIUM
P100k 15.1515151515152
Landscape
Mouse Mode
Peak Labels
None
Title
MeSH
GrantNo
NIAfield
Connections
Show None
Show Links_bib
Show Inferred
Show Infer_DPE
Scale Range
0 - 2001
0
2001
analyses/approach [1/1/2]
drosophila/age-specific [3/2/8]
longevity/mortality [3/3/9]
mortality/genetic [4/3/11]
age-specific/among [1/1/2]
mortality/ages [2/1/2]
ages/mortality [3/2/5]
Database Query
Show objects where
piname
IS_LIKE
vaupe
Send
Advanced Query
Send to Excel
Clear Legend
Clear Selected
Legend
author IS_LIKE %vaupel% --- 35
piname IS_LIKE %vaupel% --- 2

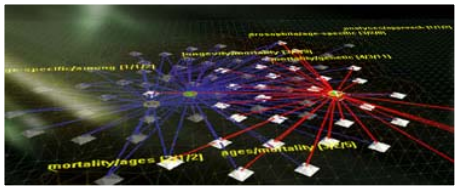


## Recommendations

**Create a clean and maintainable 'Scholarly Database' database for NIA grants & publications as a basis for the application of scientometric methods.**

- NIA enters new program announcements, grants, etc.
- Researchers enter relevant data (papers, patents, media coverage, policy changes, honors, etc.), short "new result(s)/impact headline" in standard form entries.
- Publishers input publications, e.g., Medline records, impact factors, etc.

**Make data and results of scientometric analysis publicly available.**



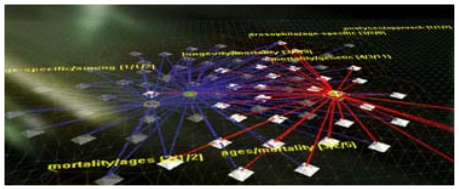
## Recommendations cont.

**This Scholarly Database & Associated KM Tools would**

- Enable cross-disciplinary information access & transfer between different research areas.
- Reveal unproductive duplication, unrealized complementarity, gaps & opportunities, and overlapping topics.
- Help to facilitate (cross-disciplinary) collaborations and to establish research priorities.
- Provide opportunities for each scientist to think systematically about the state of their field.
- Ideally help achieve consensus on what should be funded.

**Advantages for BSR/NIA**

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



## Recommendations cont.

### Advantages for Researchers

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

### Advantages for Industry

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

### Advantages for Publishers

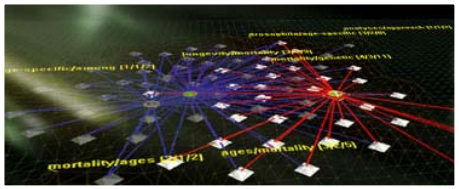
- Unique interface to their data.
- Publicly funded development of databases and their interlinkage.

### For Society

- Dramatically improved access to scientific knowledge and expertise.

# Improved Collection, Integration, Analysis and Management of

# Scholarly Knowledge & Expertise



# The Challenge: Data Collection & Integration

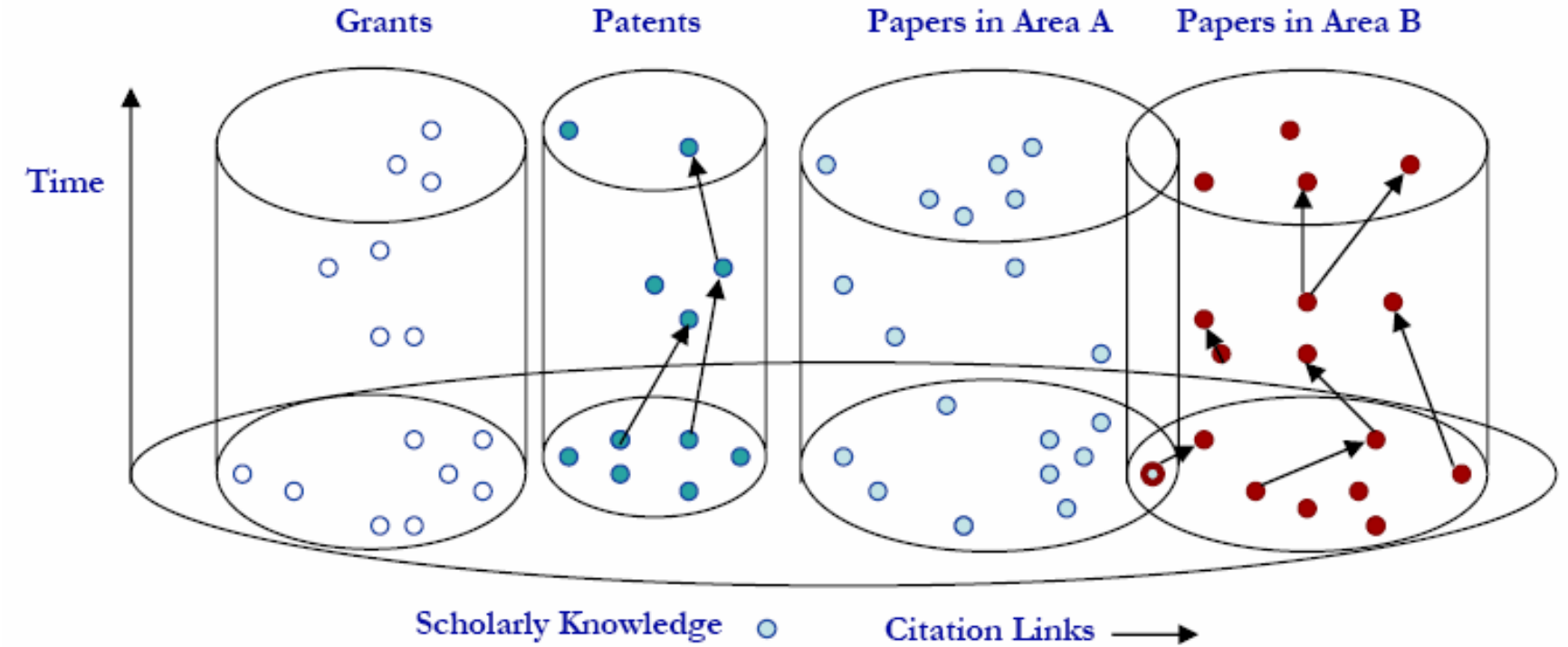
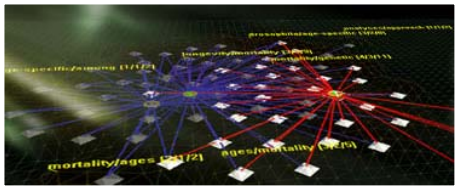



**Figure 1:** The interoperability and cross linkage problem. Many but not all of today's scholarly datasets, e.g., papers, patents, grants, are stored and made available so that 'vertical' citation linkages can be traversed. There are very few instances in which datasets of different origin and/or type are 'horizontally' interlinked.

*Börner, K. (2006) Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In Vladimir Geroimenko & Chaomei Chen (eds.) Visualizing the Semantic Web, Springer Verlag, 2nd Edition, chapter 11, pp. 183-198.*



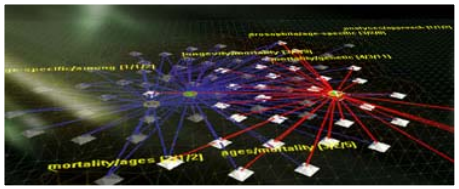
# Text Mining & Data Mining Tools

**Work well** if the records are written in similar styles, using similar formatting and conventions, are of similar length, etc.

**Work less well** if applied to interdisciplinary or multi-lingual datasets because

- Words, e.g., 'prototype', have very different meanings in computer science, biology, psychology, architecture, etc.
- Paper titles are frequently used to demonstrate the creativity of authors, e.g., "All you ever wanted to know about x", "A unifying theory of x".
- Author supplied keywords are useful to identify an author but not to find similar papers. Note: Controlled vocabularies/thesauri work well.

Humans might simply be too different and too creative to produce proper raw material that can be analyzed using existing text mining and data mining algorithms.



## Link Traversal & Link Mining

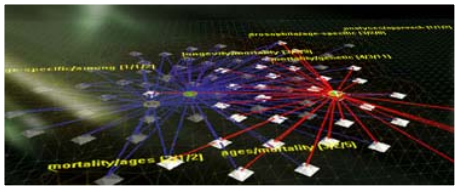
**Link Search:** Google

**(Citation) Link Traversal in One Database:** *Thomson Scientific, Google Scholar,* and *CiteSeer* already support citation link traversal. The *Proceedings of the National Academy of Sciences of the United States of* America (PNAS) online interface <http://www.pnas.org>. provides citation maps that shows articles citing or being cited by a selected article.

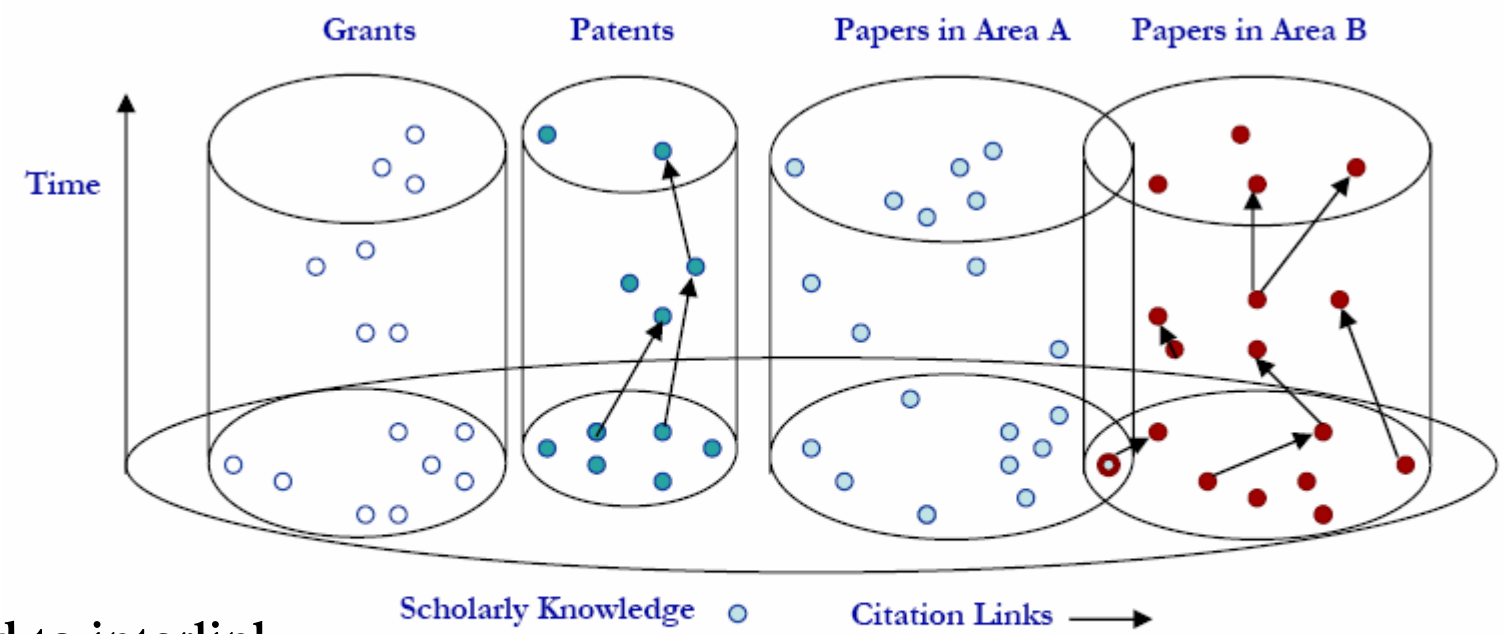
**(Citation) Link Traversal Across Databases:** The *Library Without Walls* project <http://library.lanl.gov/lww/> at the *Los Alamos National Laboratory* interlinks major publication databases and supporting citation based search across different holdings that have citation linkages.

**(Co-Author) Link Traversal:** Some digital libraries such as the citation indexes published by *Thomson Scientific*, *DBLP Bibliography Server* <http://www.sigmod.org/dblp/db/>, and *ACM Digital Library* <http://portal.acm.org> provide information on co-authorships. Services comprise a listing of all papers by an author, a listing of all co-authors for one author, and co-author link traversal.

Link traversal and link mining might be a viable alternative.



# The Challenge:
# Interlink $ Input & Publication/Patent Citation Output

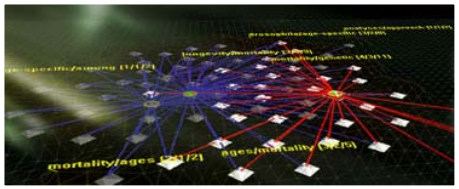



**Need to interlink**

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

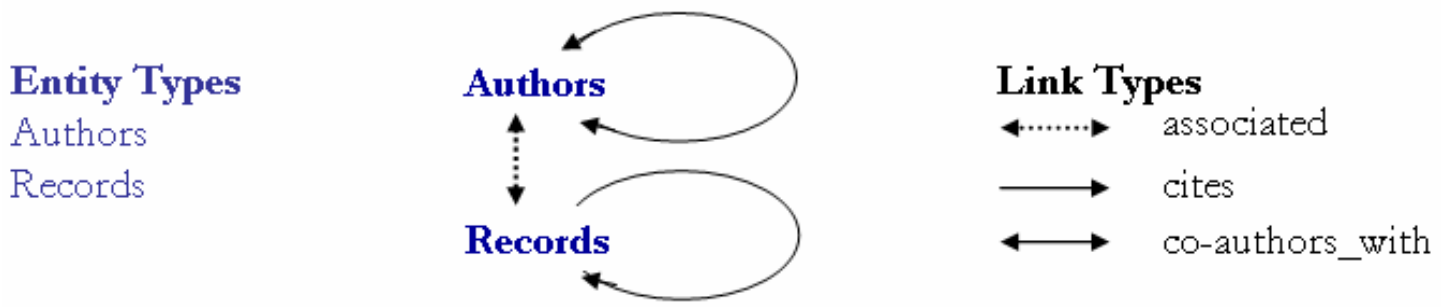
**Use resulting networks to**

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



# Improved Representation of Scholarly Knowledge

**Entity and link types:**

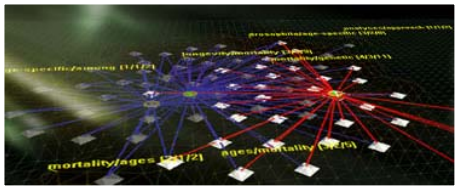



**Attributes:**

- Records often have a publication date, a publication type (e.g., journal paper, book, patents, grant, etc.), topics (e.g., keywords or classifications assigned by authors and/or publishers).
- Authors have an address with information on affiliation and geo-location.

**Derived attributes:**

- Because authors and records are associated, the geo-location(s) and affiliation(s) of an author can be attributed to the authors' papers.
- Similarly, the publication date, publication type and topic(s) can be associated with a paper's author(s).



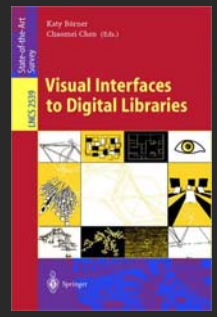
## Link Analysis

**Statistics:**

- Number of papers, grants, co-authorships, citation (over time) per author.
- Bursts of activity (#citations, #$, #patents, #collaborators, etc.).
- Changes of topics and geo-locations for authors and their institutions over time.

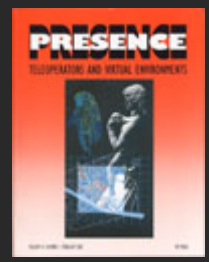
**Visualizations:**

- Geospatial and topical distribution of funding input & research output.
- Structure and evolution of research topics.
- Evolving research areas (e.g., based on young yet highly cited papers).
- Diffusion of information, people, $s over geospatial and topic space.





# Computational Scientometrics: Mapping the Structure and Evolution of Science



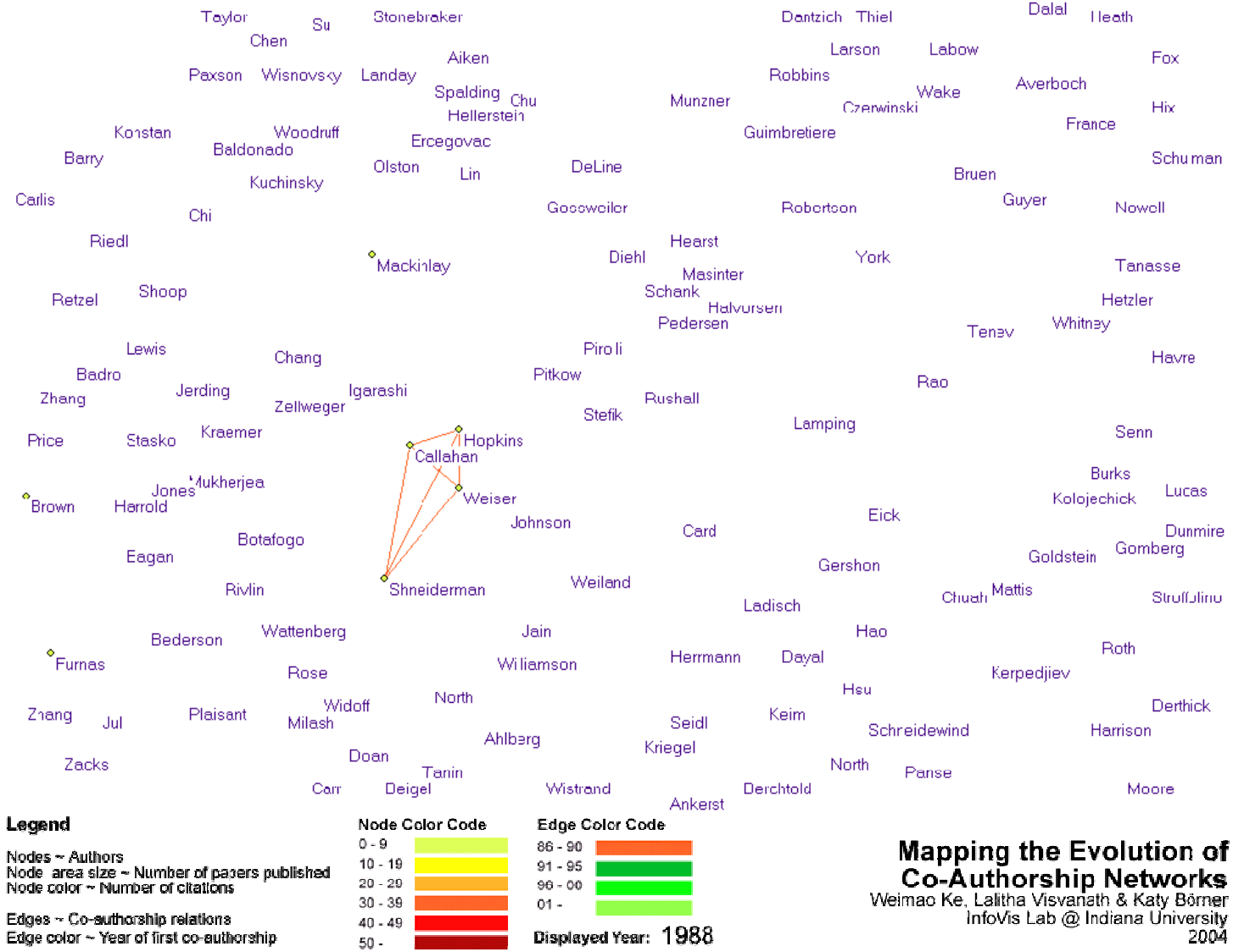



- *Shiffrin, Richard M. and Börner, Katy (Eds.) (2004). Mapping Knowledge Domains. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl_1).*
- *Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), Annual Review of Information Science & Technology, Volume 37, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 5, pp. 179-255.*

# Part (I):
# Mapping Science Locally

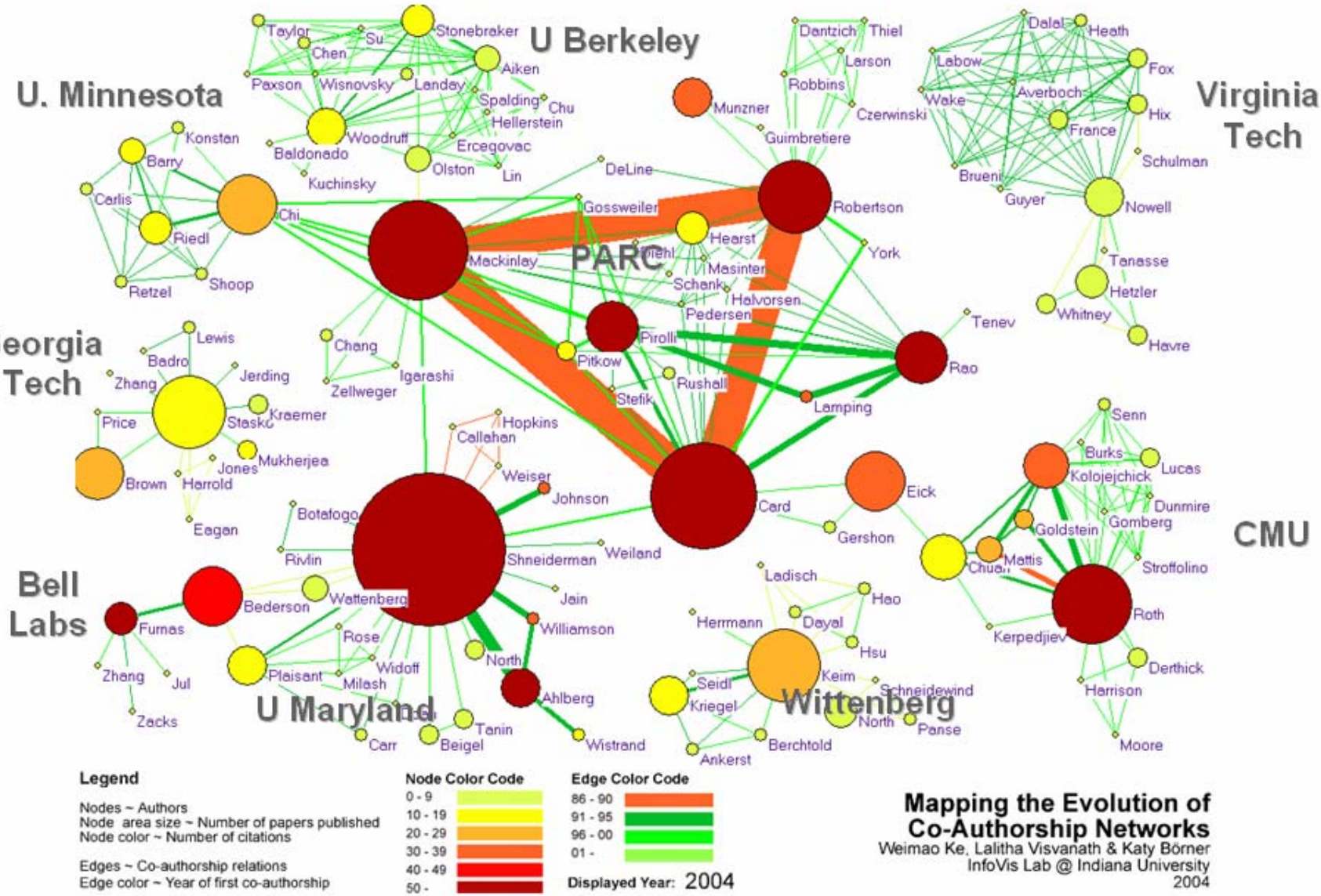
# Mapping the Evolution of Co-Authorship Networks

*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

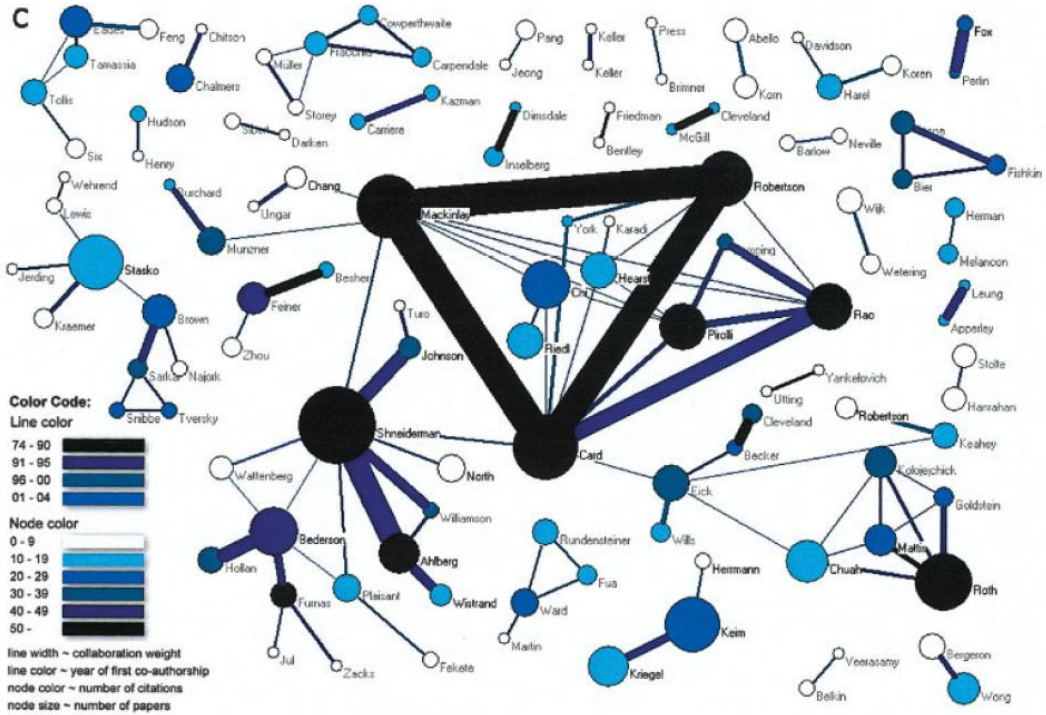
*Börner, Dall'Asta, Ke & Vespignani (2005) Complexity, 10(4):58-67.*

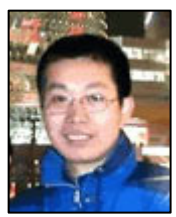
**Research question:**

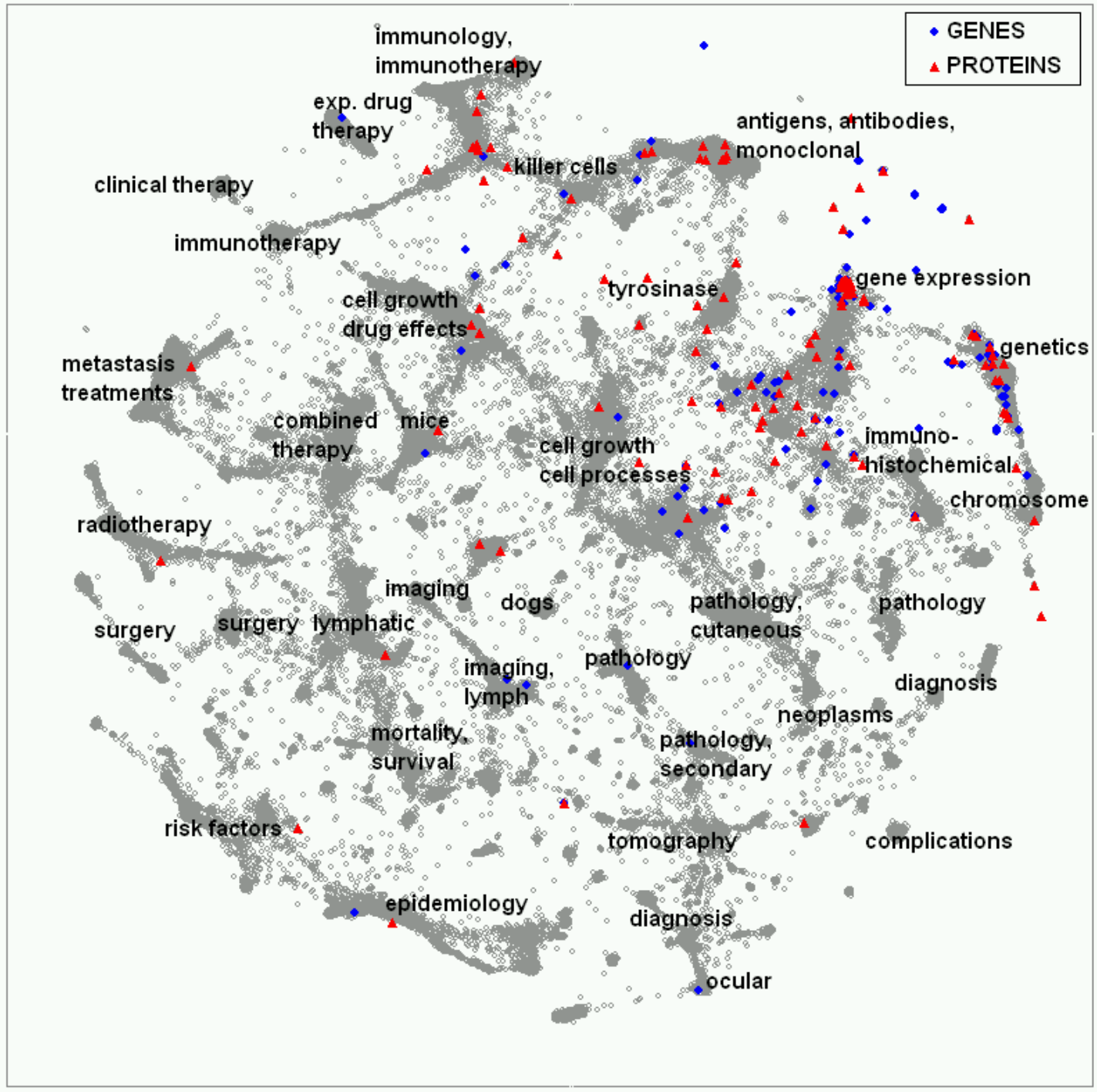
- Is science driven by prolific single experts or by high-impact co-authorship teams?

**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 Medline Papers, Genes, and Proteins Related to Melanoma Research

*Boyack, Mane & Börner. (2004) IV Conference, pp. 965-971.*

# 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.*

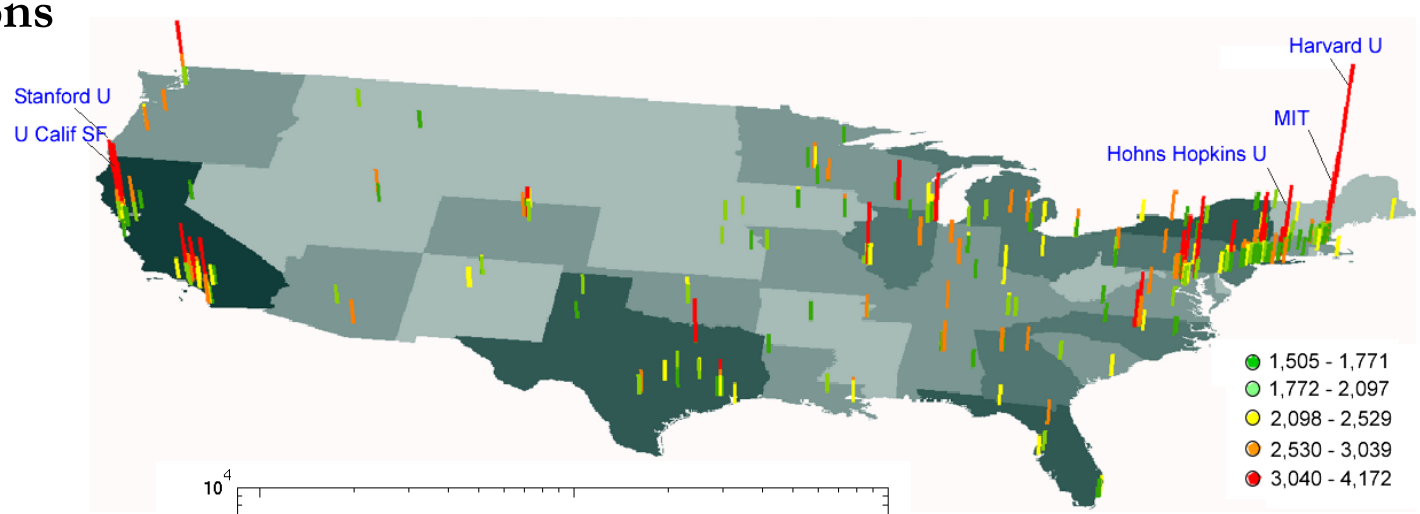

RNA
CLONING
receptors
gene expression regulation
molecular weight
messenger/genetics
89' kinetics
cultured
cells
inhibition
96' transfection
polymerase chain reaction
growth
antigen
activation
93' sequence homology
EXPRESSION
restriction mapping
protein binding
models
92' & 94' comparative study
viral
97' induction
protein
signal transduction
CELL LINE
nucleic acid hybridization
90' antibodies
89' GENES
2000' tumor cells
89' plasmids
87' structural
88' inbred strains
blotting
rats
MOLECULAR SEQUENCE DATA
99' mutagenesis
nucleic acid
binding
apoptosis
mutation
growth-factor
95' base sequence
amino acid
amino acid sequence
in vitro
receptor
sequence
gene expression

**Color Code**
82 - 85
86 - 89
90 - 93
94 - 97
98 - 01

circle size ~ burst weight
cicle color ~ burst onset
ring color ~ year of max word count
years of 2nd and 3rd burst are given in color

# Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions

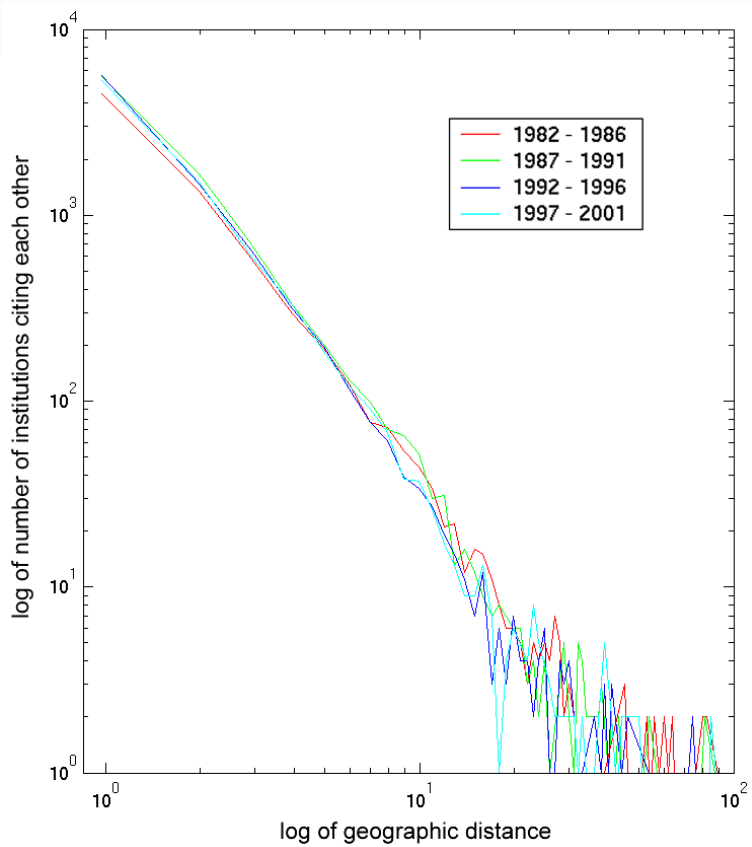
*Börner & Penumarthy. (2005)*




Does Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research instructions?

Analysis of top 500 most highly cited U.S. institutions.

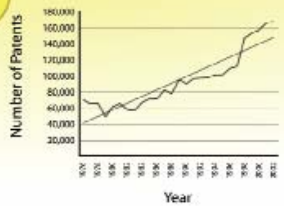
Each institution is assumed to produce and consume information.




$\gamma_{82\text{-}86} = 1.94$ ($R^2$=91.5%)
$\gamma_{87\text{-}91} = 2.11$ ($R^2$=93.5%)
$\gamma_{92\text{-}96} = 2.01$ ($R^2$=90.8%)
$\gamma_{97\text{-}01} = 2.01$ ($R^2$=90.7%)

# Examining the Evolution and Distribution of Patent Classifications

## 1 Patents Granted Over the Last 20 Years




Top Classes 1978 - 1982

| Class | Title | Patents |
|---|---|---|
| 260 | Chemistry of Carbon Compounds | 13,349 |
| 424 | Drug, Bio-Affecting and Body Treating Compositions | 9,621 |
| 252 | Compositions | 5,868 |
| 128 | Surgery | 5,404 |
| 428 | Stock Material or Miscellaneous Articles | 4,984 |
| 73 | Measuring and Testing | 4,895 |
| 123 | Internal-Combustion Engines | 4,142 |
| 340 | Communications: Electrical | 4,043 |
| 364 | Electrical Computers and Data Processing Systems | 3,780 |
| 29 | Metal Working | 3,664 |
| | Total | 60,705 |

Top Classes 1998 - 2002

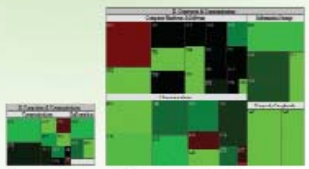
| Class | Title | Patents |
|---|---|---|
| 514 | Drug, Bio-Affecting and Body Treating Compositions | 18,778 |
| 438 | Semiconductor Device Manufacturing Process | 17,775 |
| 435 | Chemistry, Molecular Biology and Microbiology | 17,424 |
| 424 | Drug, Bio-Affecting and Body Treating Compositions | 13,697 |
| 428 | Stock Material or Miscellaneous Articles | 13,316 |
| 257 | Active Solid-State Devices (e.g., Transistors, Solid-state Diodes | 12,024 |
| 395 | Information Processing System Organization | 9,955 |
| 345 | Computer Graphics Processing, Operator Interface Processing, and Selective Visual Display Systems | 9,510 |
| 359 | Optical: Systems and Elements | 9,191 |
| 365 | Static Information Storage and Retrieval | 8,392 |
| | Total | 140,010 |

In the United States, each patent gets assigned to one out of more than 450 classes covering broad application domains. An examination of the size and growth of patent classes provides insight about patenting trends.

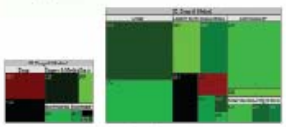
Treemaps, a space filling technique developed in the HCI Lab at the University of Maryland, are used to communicate major results. Treemaps represent a tree structure as nested rectangles with each rectangle representing a node. A rectangular area is first allocated to hold the representation of the tree, and this area is then subdivided into a set of rectangles that represent the top level of the tree. This process continues recursively on the resulting rectangles to represent each lower level of the tree. The parent-child relationship is indicated by enclosing the child rectangle by its parent rectangle. Typically, the size of each rectangle corresponds to the size of the node. Additional information about a node, e.g., its age or value, can be represented by the color of the respective rectangle.

## Fast Growth Domains
1983 - 1987 / 1998 - 2002

Electrical and Electronic



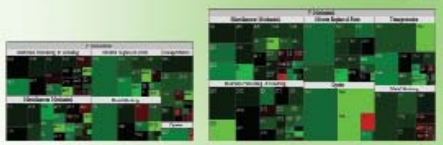
Computers and Communications



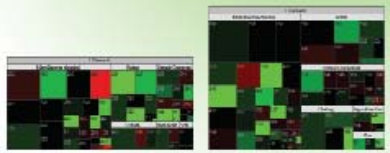
Drugs and Medical

## Slow Growth Domains
1983 - 1987 / 1998 - 2002

## 2



Mechanical

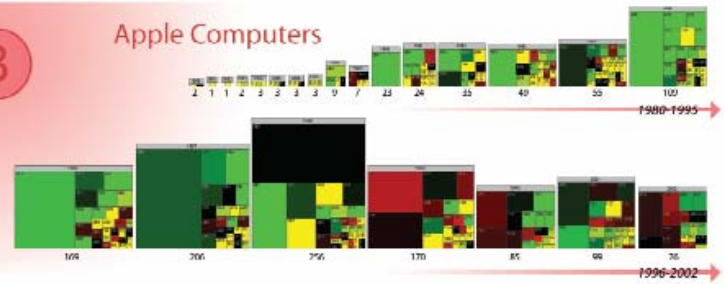


Chemical

Others

Shown is a comparison of the patent class space for 1983 to 1987 and 1998 to 2002. There is a predominance of growth in the 1998 to 2002 patent space, which correlates to the increase in patent grants during this period. By comparing the growth in categories, one can distinguish between domains that have been receiving a larger amount of patent grants.

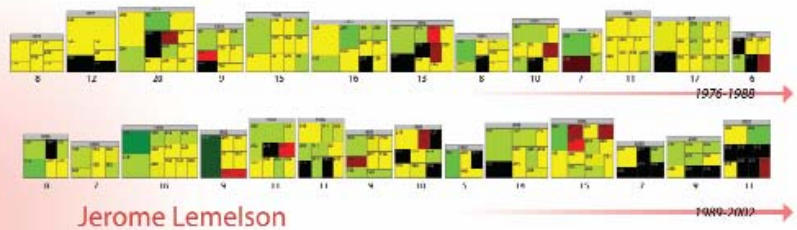
## 3 Apple Computers




Depicted above is how Apple Computers' portfolio has changed in yearly increments from 1980 to 2002.

Lemelson's patent holdings below show a more even distribution over multiple classes. No class dominates over a majority of the years for granted patents; instead they are distributed more broadly over the intellectual space.

**Legend:**
Green = Increase in number of patent grants in particular class.
Red = Decrease in number of patent grants in particular class.
Yellow = No patents granted in that class in the past five years.
Size = Number of patent grants in a particular class.

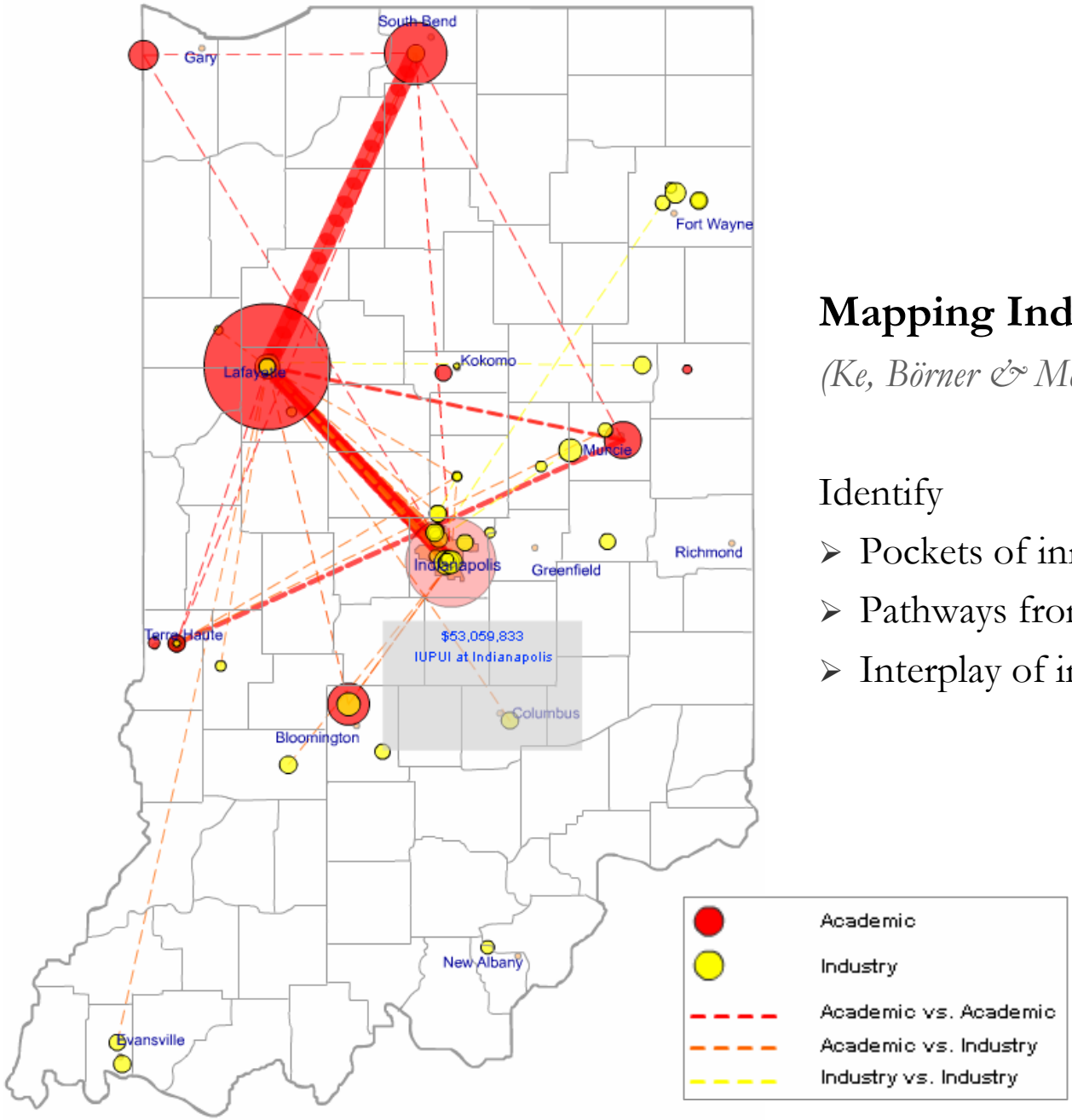


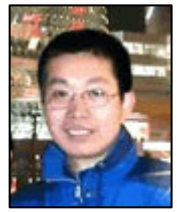

Jerome Lemelson

Kutz, Daniel O. Examining the Evolution and Distribution of Patent Classifications. Accepted for the Information Visualization Conference, London, UK, July 2004.

The material is based upon work supported bythe National Science Foundation under Grant No. IIS-0238261.

For more Information, contact Katy Borner at katy@indiana.edu.



## Mapping Indiana's Intellectual Space

*(Ke, Börner & Mei, 2005)*

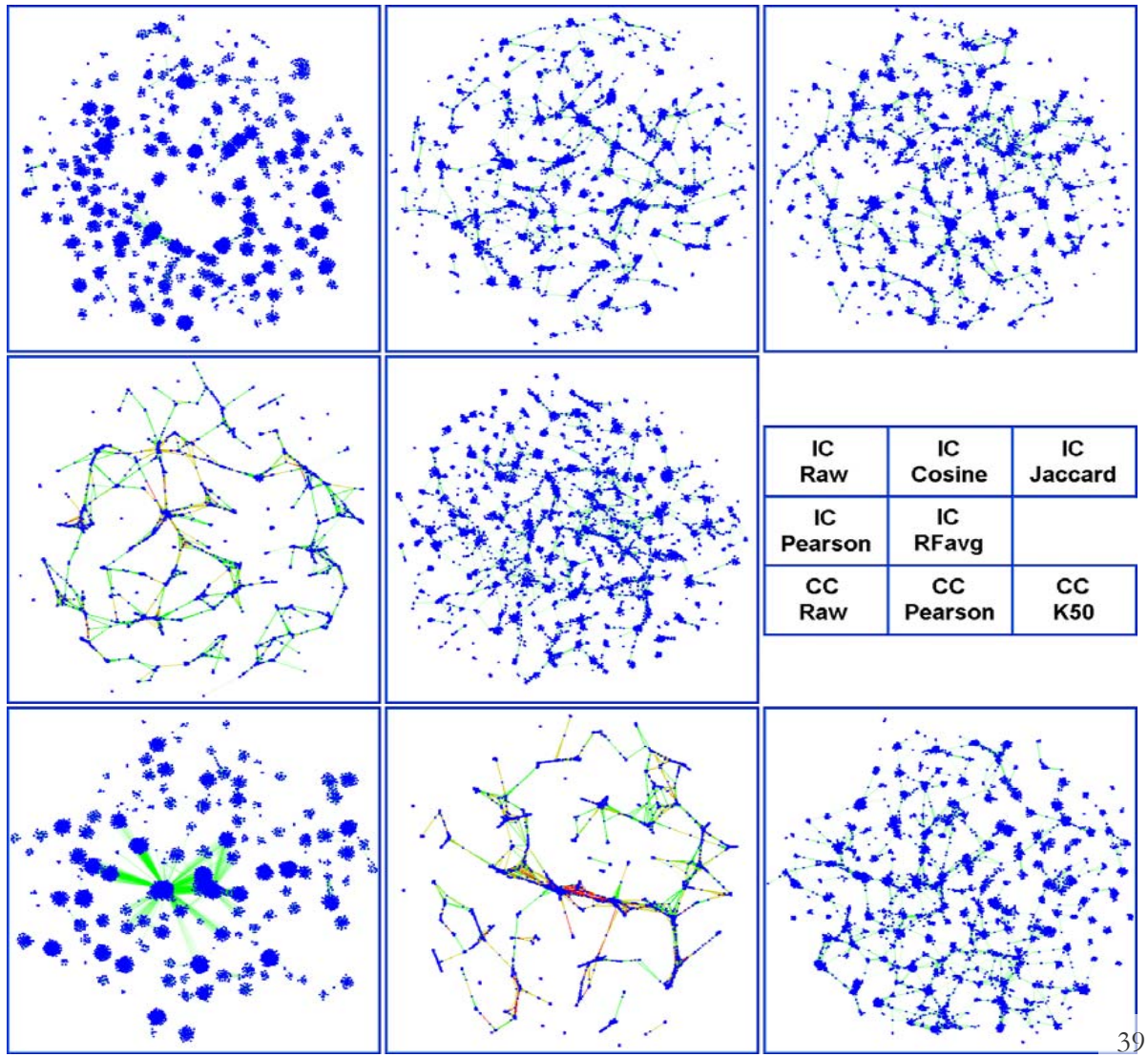
Identify

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

# Part (II):
# Mapping Science Globally

# Comparison of Similarity Metrics

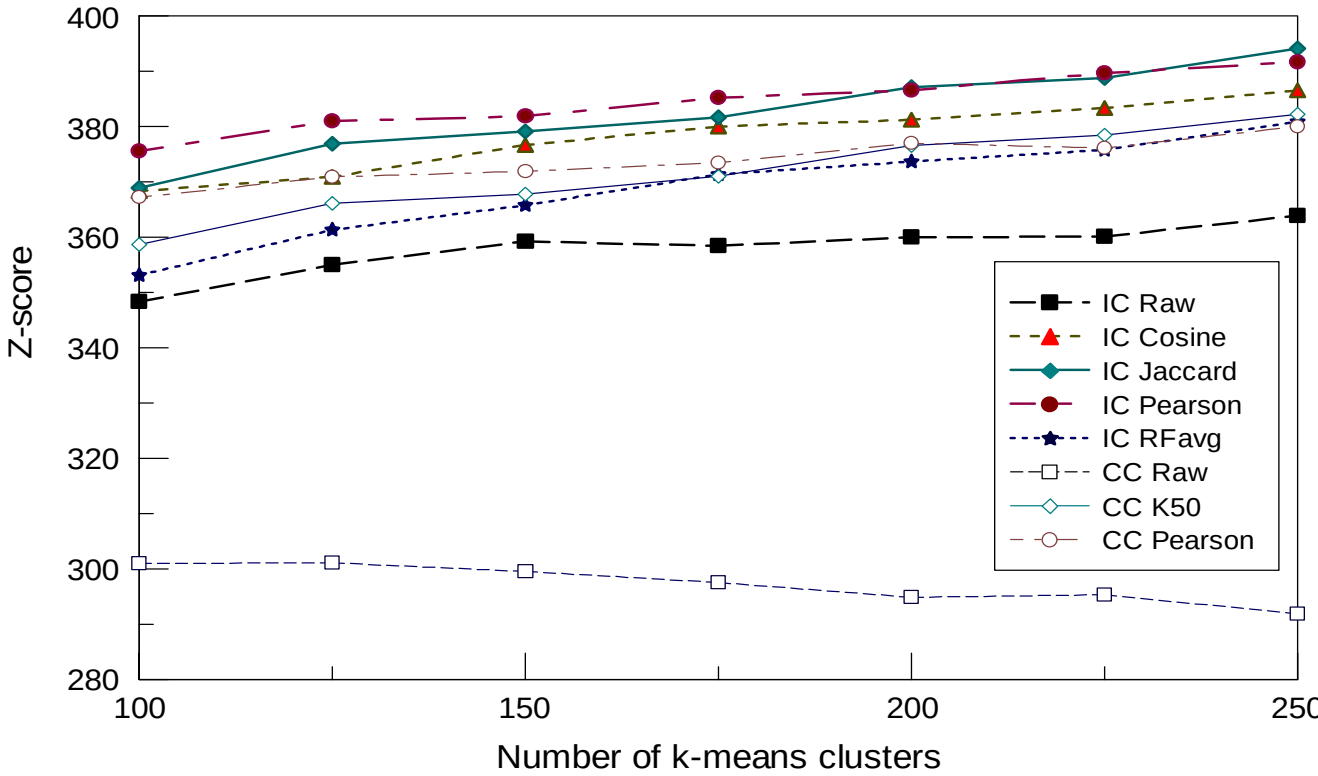
- ISI file year 2000, SCI and SSCI: 7,121 journals.
- Different similarity metrics
  - Inter-citation (raw counts, cosine, modified cosine, Jaccard, RF, Pearson)
  - Co-citation (raw counts, cosine, modified cosine, Pearson)
- Maps were compared based on
  - regional accuracy,
  - the scalability of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




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

## Selecting the similarity measure with the best regional accuracy

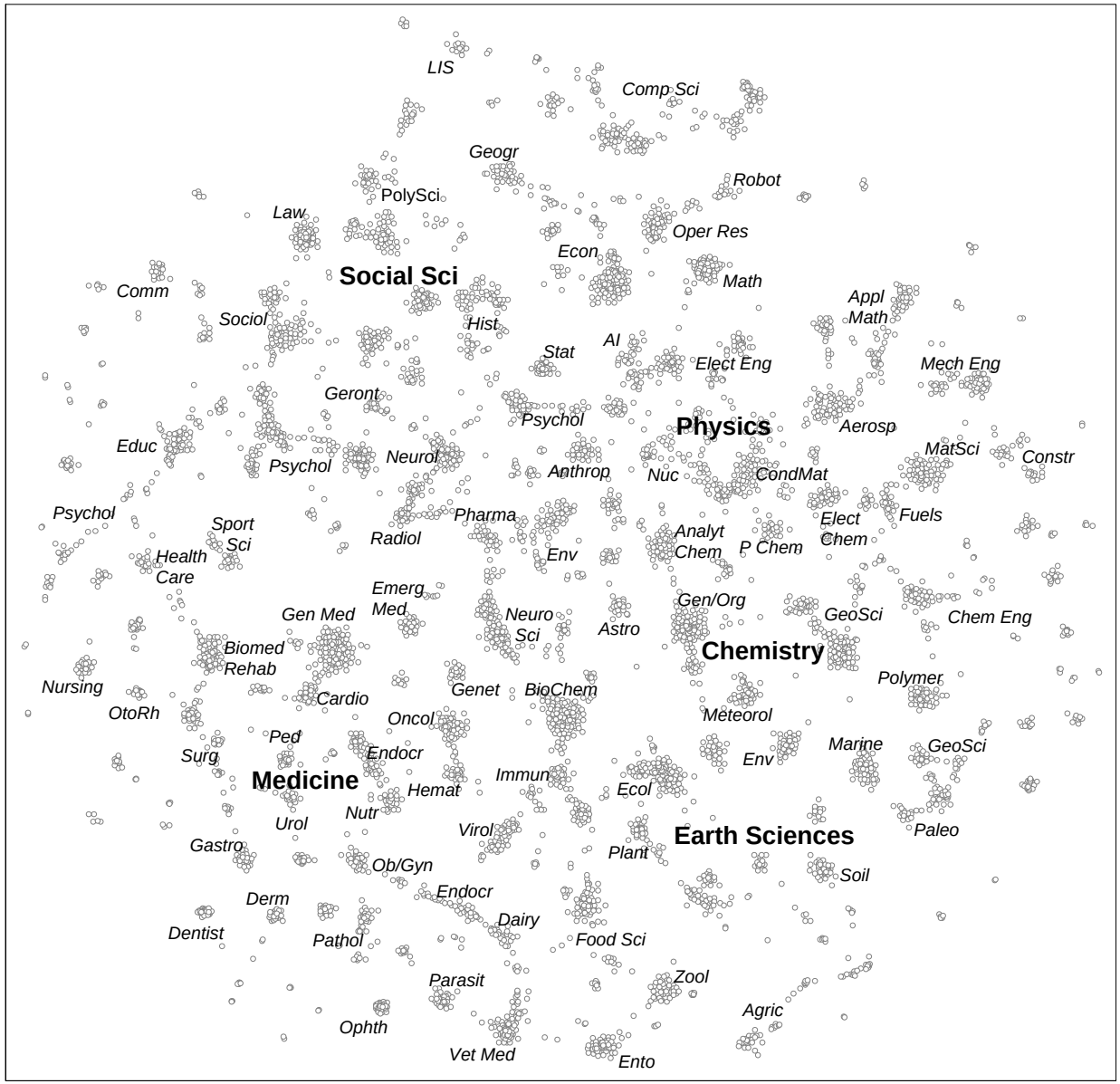
- For each similarity measure, the VxOrd layout was subjected to k-means clustering using different numbers of clusters.
- Resulting cluster/category memberships were compared to actual category memberships using entropy/mutual information method by Gibbons & Roth, 2002.
- Increasing Z-score indicates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

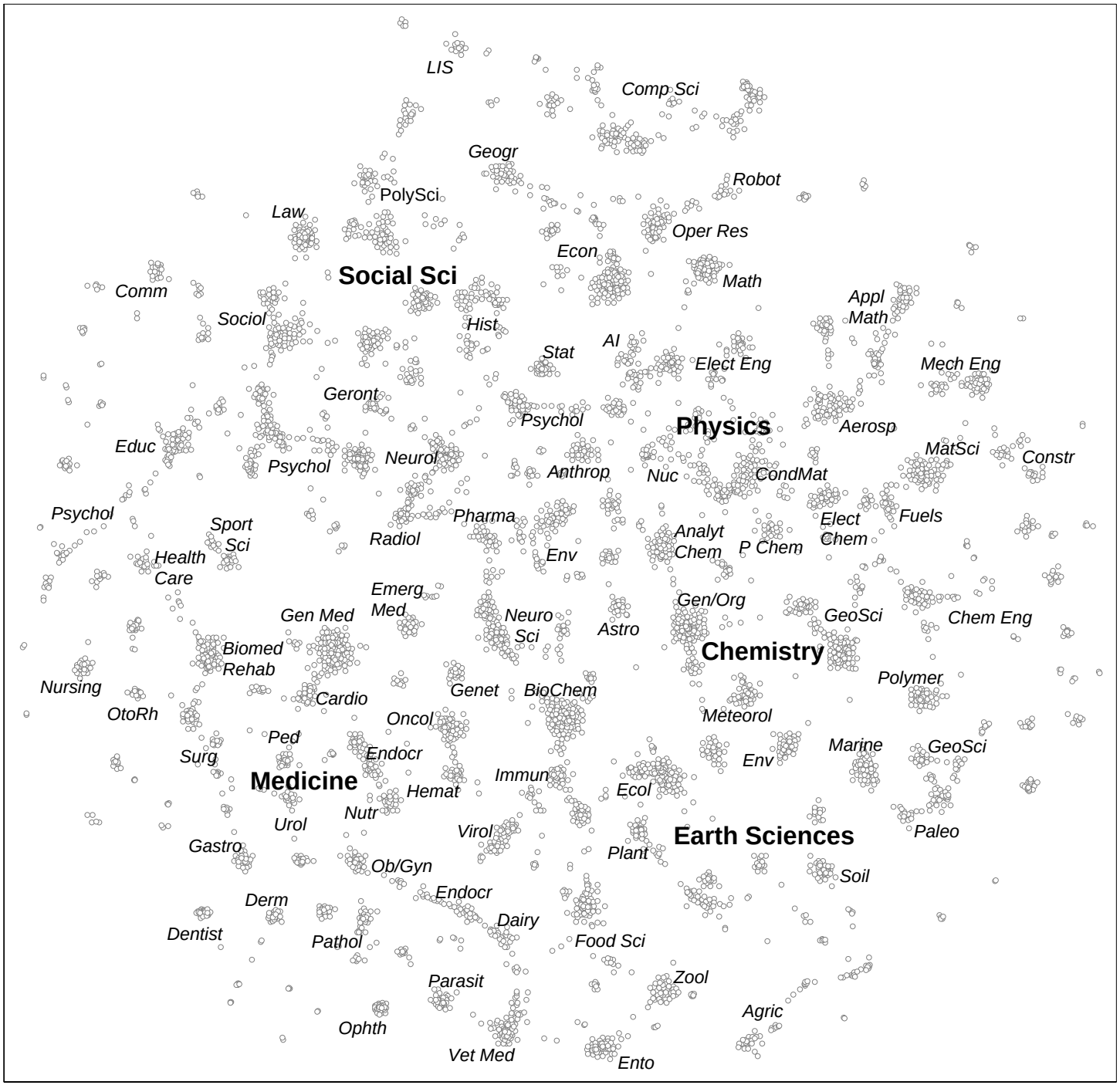



*Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. Scientometrics. 64(3), 351-374.*

# A 'Backbone' Map of Science & Social Science

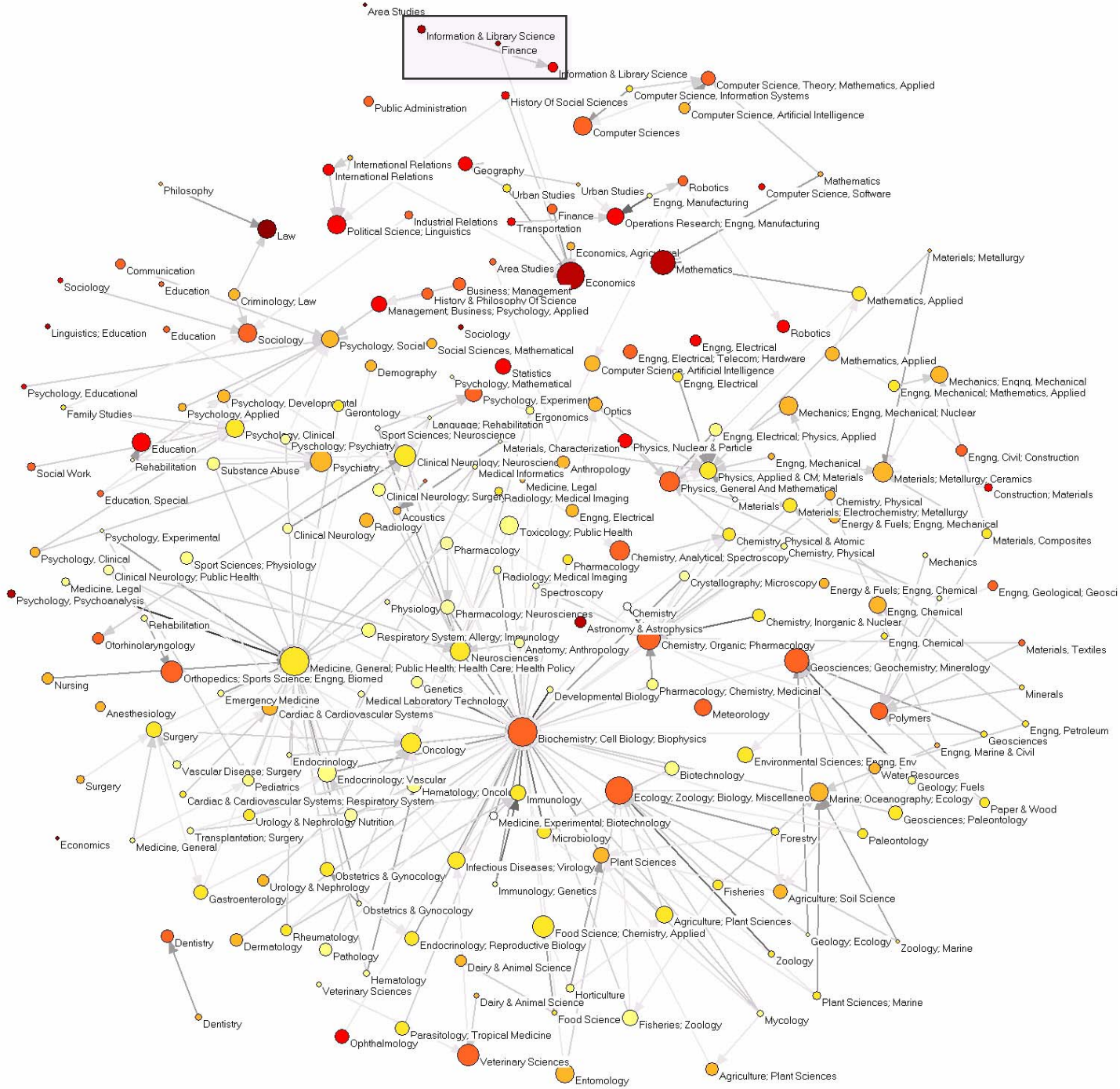
- The map is comprised of 7,121 journals from year 2000.
- Each dot is one journal
- An *IC-Jaccard* similarity measure was used.
- Journals group by discipline.
- Groups are labeled by hand.
- Large font size labels identify major areas of science.
- Small labels denote the disciplinary topics of nearby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



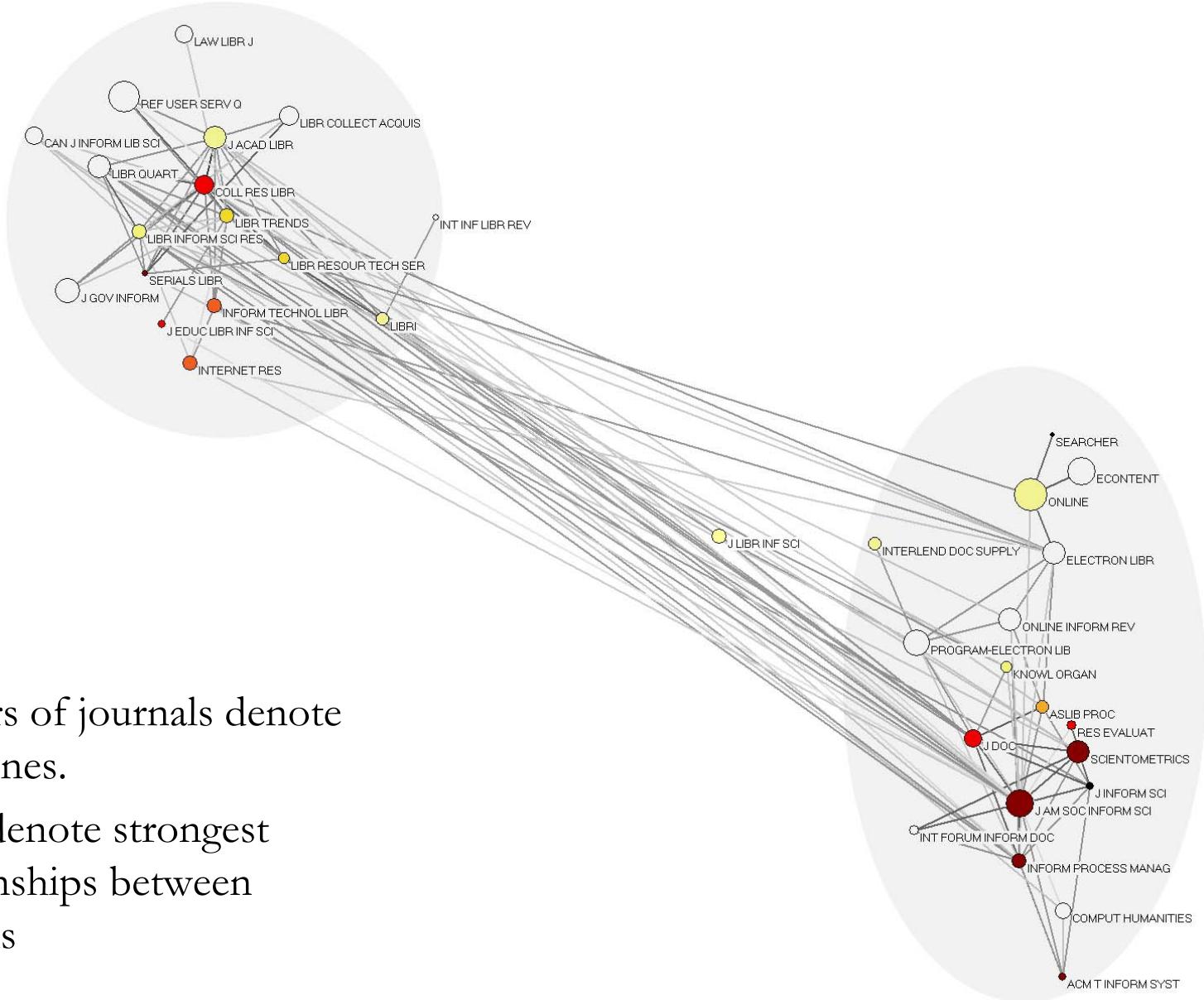

LIS
Comp Sci
Geogr
Robot
PolySci
Law
Oper Res
Econ
Social Sci
Math
Comm
Appl Math
Sociol
Hist
AI
Stat
Elect Eng
Mech Eng
Geront
Psychol
Physics
Aerosp
Educ
MatSci
Constr
Psychol
Neurol
Anthrop
Nuc
CondMat
Psychol
Sport Sci
Pharma
Elect Chem
Fuels
Radiol
Analyt Chem
P Chem
Health Care
Env
Emerg Med
Gen/Org
Gen Med
Neuro Sci
Astro
GeoSci
Chem Eng
Biomed Rehab
Chemistry
Polymer
Nursing
Genet
BioChem
Cardio
OtoRh
Oncol
Meteorol
Ped
Marine
GeoSci
Surg
Endocr
Env
Medicine
Immun
Ecol
Hemat
Nutr
Urol
Virol
Earth Sciences
Paleo
Gastro
Plant
Ob/Gyn
Soil
Endocr
Derm
Dairy
Dentist
Pathol
Food Sci
Parasit
Zool
Agric
Ophth
Vet Med
Ento

## Structural Map: Studying Disciplinary Diffusion

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


Area Studies
Information & Library Science
Finance
Information & Library Science
Computer Science, Theory; Mathematics, Applied
Computer Science, Information Systems
Computer Science, Artificial Intelligence
Public Administration
History Of Social Sciences
Computer Sciences
International Relations
International Relations
Geography
Philosophy
Urban Studies
Urban Studies
Robotics
Engng, Manufacturing
Mathematics
Computer Science, Software
Finance
Operations Research; Engng, Manufacturing
Industrial Relations
Transportation
Political Science; Linguistics
Law
Economics, Agricultural
Area Studies
Economics
Mathematics
Materials; Metallurgy
Communication
Sociology
Education
Criminology; Law
Business; Management
History & Philosophy Of Science
Management; Business; Psychology, Applied
Mathematics, Applied
Linguistics; Education
Education
Sociology
Sociology
Psychology, Social
Social Sciences, Mathematical
Robotics
Engng, Electrical
Engng, Electrical; Telecom; Hardware
Computer Science, Artificial Intelligence
Engng, Electrical
Mathematics, Applied
Demography
Statistics
Psychology, Mathematical
Mechanics; Engng, Mechanical
Engng, Mechanical; Mathematics, Applied
Psychology, Educational
Family Studies
Psychology, Developmental
Psychology, Applied
Gerontology
Psychology, Experimental
Optics
Ergonomics
Mechanics; Engng, Mechanical; Nuclear
Language; Rehabilitation
Psychology, Clinical
Psychology; Psychiatry
Sport Sciences; Neuroscience
Engng, Electrical; Physics, Applied
Education
Rehabilitation
Substance Abuse
Psychiatry
Clinical Neurology; Neuroscience
Materials, Characterization
Physics, Nuclear & Particle
Engng, Civil; Construction
Social Work
Medical Informatics
Anthropology
Engng, Mechanical
Physics, Applied & CM; Materials
Materials; Metallurgy; Ceramics
Medicine, Legal
Physics, General And Mathematical
Construction; Materials
Education, Special
Clinical Neurology; Surgery
Radiology; Medical Imaging
Materials
Chemistry, Physical
Materials; Electrochemistry; Metallurgy
Acoustics
Engng, Electrical
Energy & Fuels; Engng, Mechanical
Radiology
Toxicology; Public Health
Psychology, Experimental
Clinical Neurology
Chemistry, Physical & Atomic
Materials, Composites
Pharmacology
Psychology, Clinical
Sport Sciences; Physiology
Chemistry, Analytical; Spectroscopy
Chemistry, Physical
Pharmacology
Mechanics
Clinical Neurology; Public Health
Radiology; Medical Imaging
Crystallography; Microscopy
Energy & Fuels; Engng, Chemical
Engng, Geological; Geosci
Medicine, Legal
Spectroscopy
Psychology, Psychoanalysis
Physiology
Pharmacology; Neurosciences
Chemistry
Engng, Chemical
Rehabilitation
Astronomy & Astrophysics
Chemistry, Inorganic & Nuclear
Respiratory System; Allergy; Immunology
Chemistry, Organic; Pharmacology
Engng, Chemical
Materials, Textiles
Otorhinolaryngology
Anatomy; Anthropology
Neurosciences
Geosciences; Geochemistry; Mineralogy
Medicine, General; Public Health; Health Care; Health Policy
Orthopedics; Sports Science; Engng, Biomed
Nursing
Genetics
Pharmacology; Chemistry, Medicinal
Minerals
Emergency Medicine
Medical Laboratory Technology
Developmental Biology
Meteorology
Polymers
Anesthesiology
Cardiac & Cardiovascular Systems
Surgery
Biochemistry; Cell Biology; Biophysics
Engng, Petroleum
Geosciences
Oncology
Endocrinology
Environmental Sciences; Engng, Env
Engng, Marine & Civil
Vascular Disease; Surgery
Biotechnology
Water Resources
Endocrinology; Vascular
Pediatrics
Geology; Fuels
Surgery
Hematology; Oncology
Immunology
Ecology; Zoology; Biology, Miscellaneous
Marine; Oceanography; Ecology
Paper & Wood
Cardiac & Cardiovascular Systems; Respiratory System
Geosciences; Paleontology
Urology & Nephrology; Nutrition
Medicine, Experimental; Biotechnology
Transplantation; Surgery
Microbiology
Forestry
Paleontology
Economics
Medicine, General
Plant Sciences
Infectious Diseases; Virology
Obstetrics & Gynecology
Urology & Nephrology
Immunology; Genetics
Fisheries
Agriculture; Soil Science
Gastroenterology
Obstetrics & Gynecology
Agriculture; Plant Sciences
Food Science; Chemistry, Applied
Rheumatology
Geology; Ecology
Zoology; Marine
Dentistry
Dermatology
Endocrinology; Reproductive Biology
Pathology
Zoology
Dairy & Animal Science
Hematology
Veterinary Sciences
Horticulture
Plant Sciences; Marine
Dairy & Animal Science
Mycology
Food Science
Fisheries; Zoology
Dentistry
Parasitology; Tropical Medicine
Ophthalmology
Veterinary Sciences
Entomology
Agriculture; Plant Sciences

# Zoom Into Structural Map

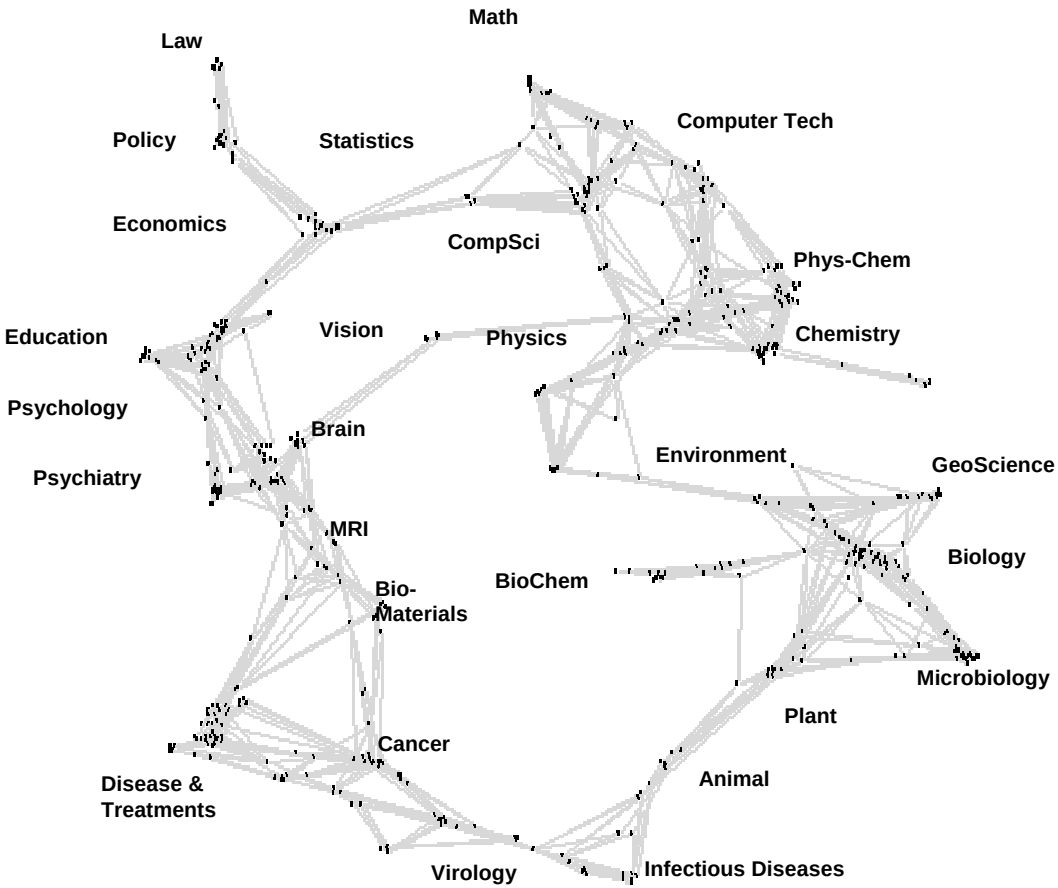



- Clusters of journals denote disciplines.
- Lines denote strongest relationships between journals

## Latest 'Base Map' of Science

*Kevin W. Boyack & Richard Klavans, unpublished work.*

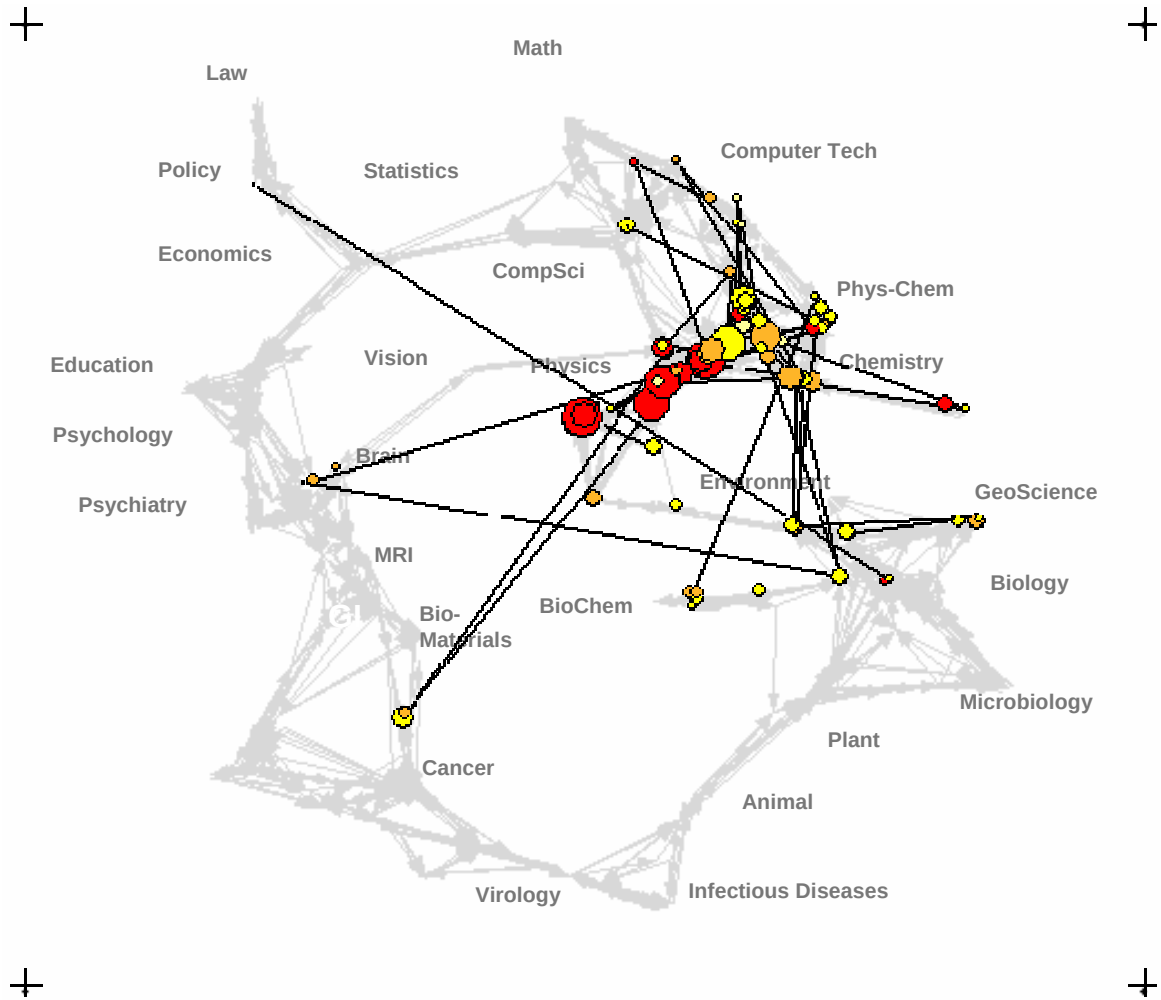
- Uses combined SCI/SSCI from 2002
  - 1.07M papers, 24.5M references, 7,300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts were reaggregated at the journal cluster level to calculate the
  - (x,y)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



# Science map applications: Identifying core competency

*Kevin W. Boyack & Richard Klavans, unpublished work.*

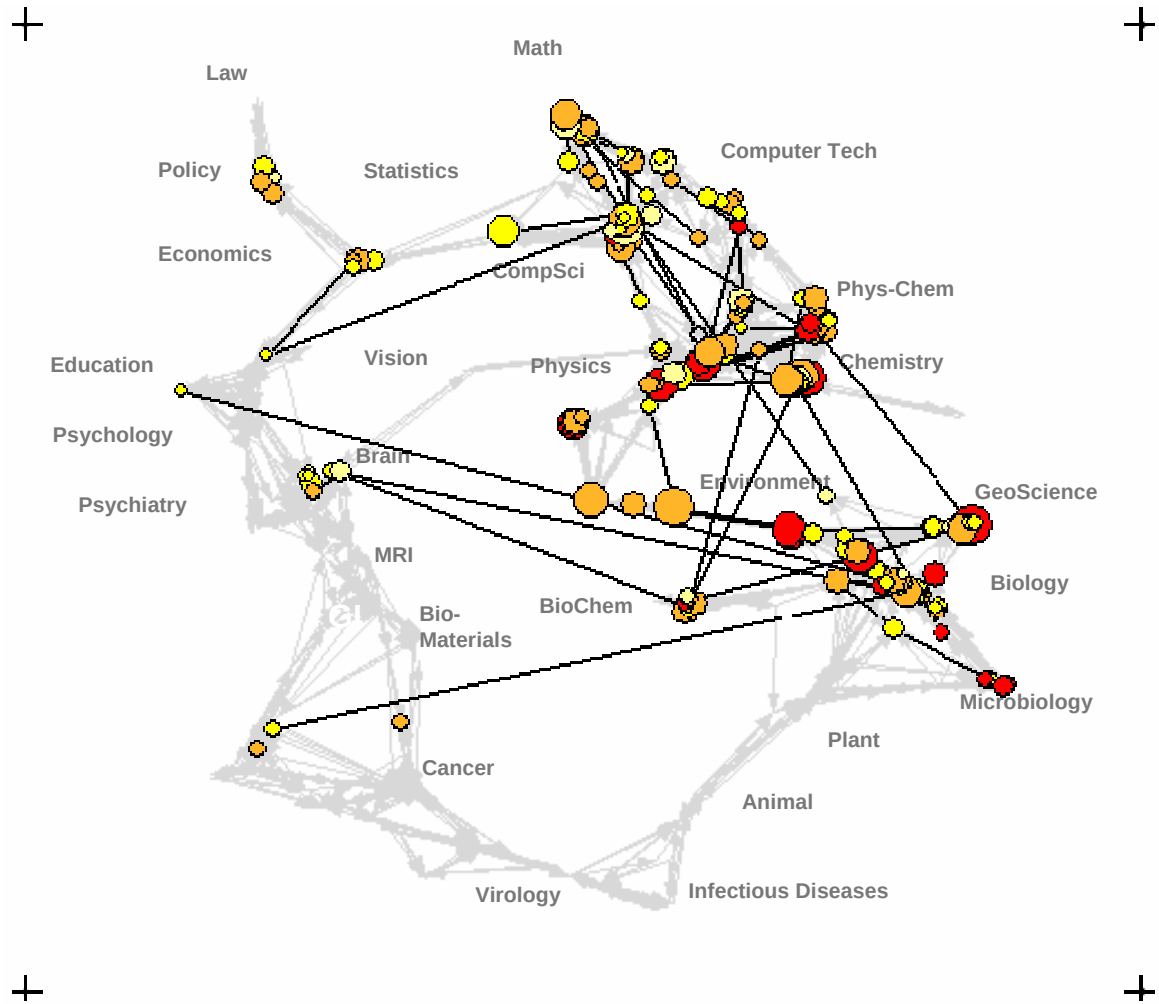
Funding patterns of the US Department of Energy (DOE)



# Science map applications: Identifying core competency

*Kevin W. Boyack & Richard Klavans, unpublished work.*

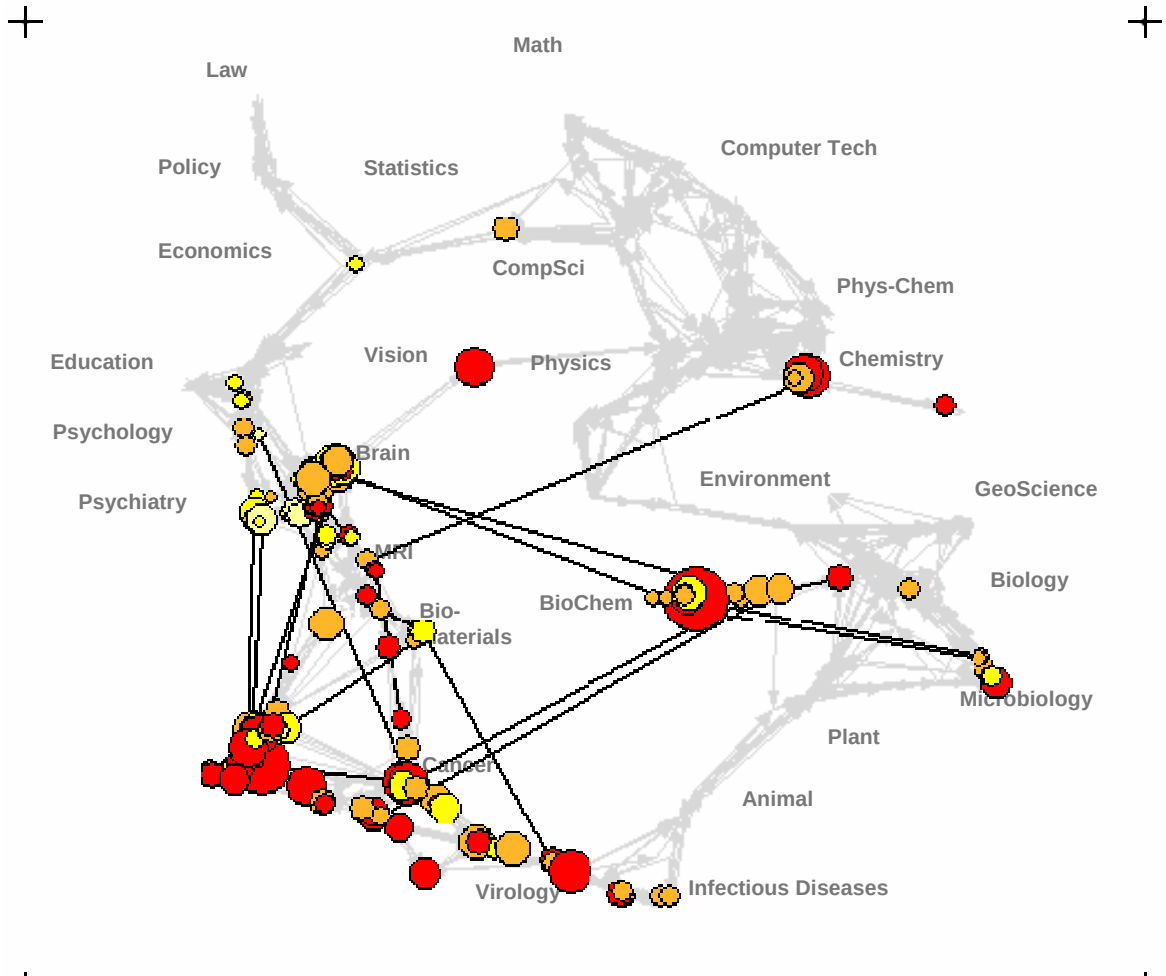
## Funding Patterns of the National Science Foundation (NSF)



# Science map applications: Identifying core competency

*Kevin W. Boyack & Richard Klavans, unpublished work.*

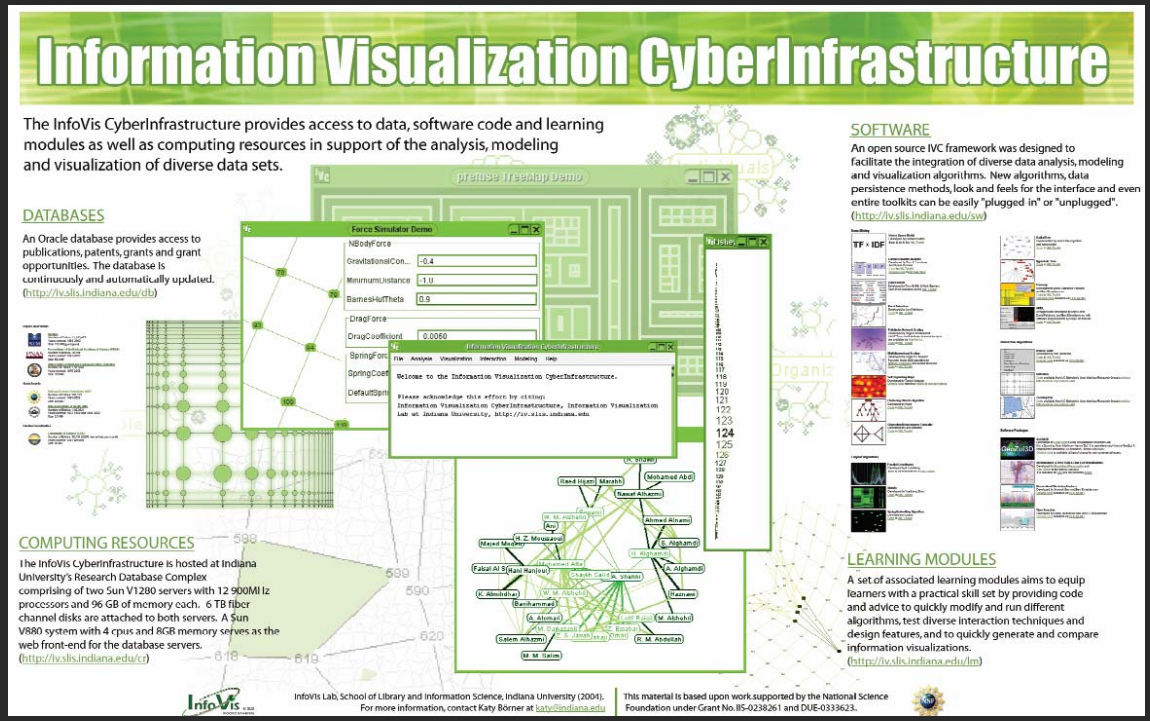
## Funding Patterns of the National Institutes of Health (NIH)




## Places & Spaces: Cartography of the Physical and the Abstract

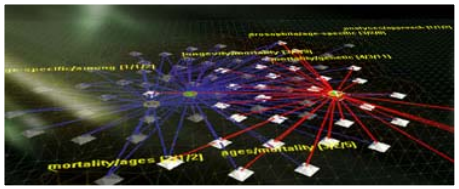
The first iteration of this science exhibit compares and contrasts first maps of our entire planet with the first maps of all of sciences.

http://vw.indiana.edu/places&spaces/


*CAREER: Visualizing Knowledge Domains. NSF IIS-0238261 award (Katy Börner, $440,000) Sept. 03-Aug. 08.*


*SEI: Network Workbench: A Large-Scale Network Analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research. NSF IIS-0513650 award (Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Craig Stewart (Senior Personnel), $1,120,926) Sept. 05 - Aug. 08.*



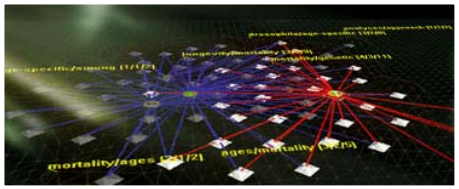
# Acknowledgements

I would like to thank the students in the Cyberinfrastructure for Network Science Center and the Information Visualization Laboratory at IU and my collaborators for their contributions to this work.

Support comes from the School of Library and Information Science, Indiana University's High Performance Network Applications Program, a Pervasive Technology Lab Fellowship, a 21st Century Grant, an Outstanding Junior Faculty Award, a SBC (formerly Ameritech) Fellow Grant, an NIH Evaluation Express award, and National Science Foundation grants DUE-0333623, CHE 0524661, IIS-0238261, and IIS-0513650. We also acknowledge Equipment Grants by SUN Microsystems, the 21st Century Fund, and Indiana University.

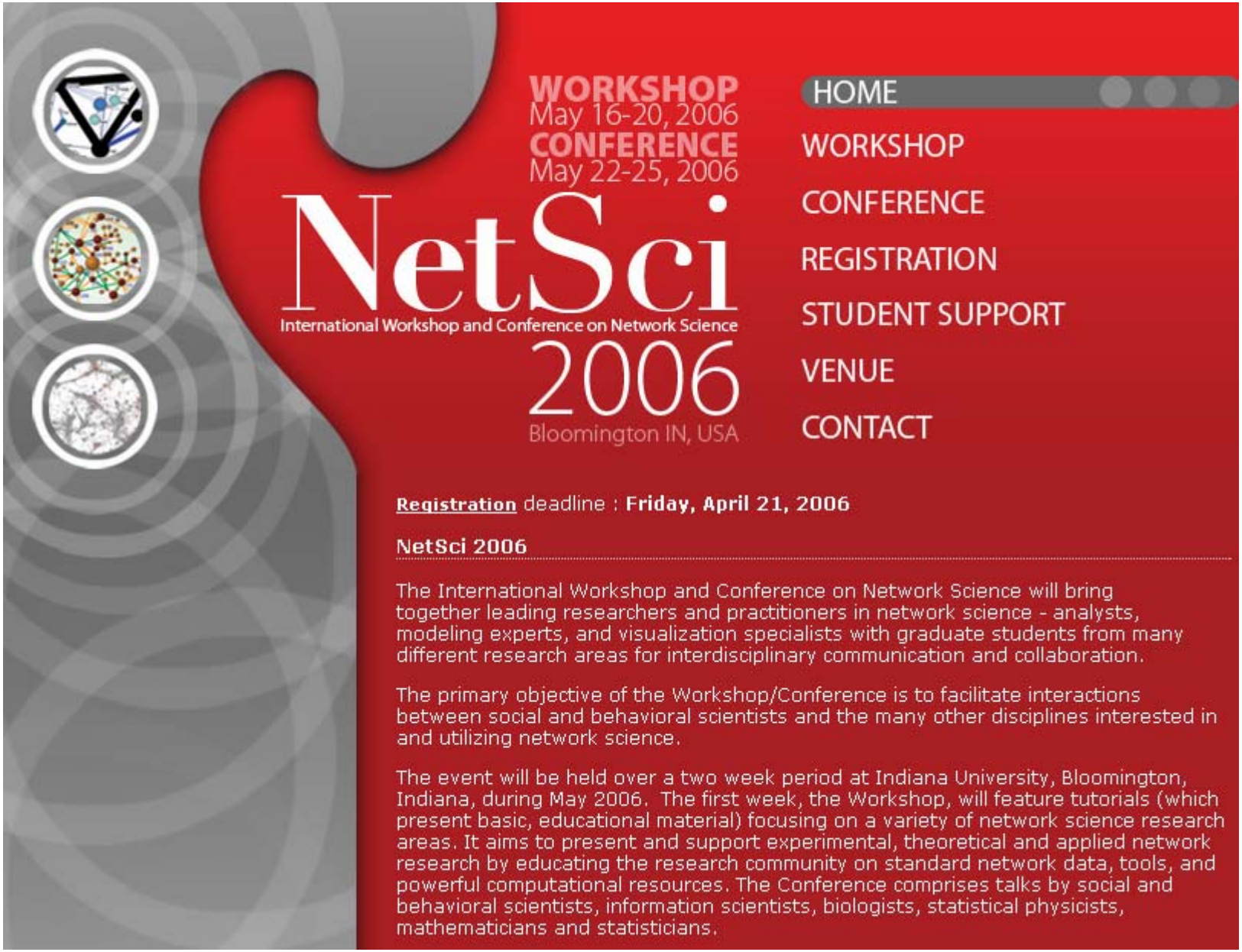
# Pointers to Papers & Expertise

(not covered in the talk)



# References

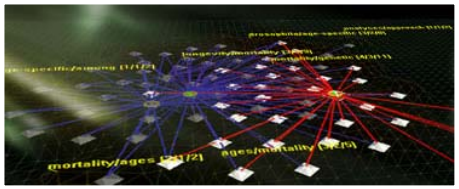

- Börner, Katy. Mapping All of Science: How to Collect, Organize and Make Sense of Mankind's Scholarly Knowledge and Expertise. Submitted to *Environment and Planning B*, Special Issue on *Mapping Humanity's Knowledge and Expertise in the Digital Domain*.
- Börner, Katy, Penumarthy, Shashikant, Meiss, Mark and Ke, Weimao. Mapping the Diffusion of Scholalry Knowledge Among Major U.S. Research Institutions. Accepted for *Scientometrics*. Dedicated issue on the *10th International Conference of the International Society for Scientometrics and Informetrics* held in Stockholm.
- Holloway, Todd, Božicevic, Miran and Börner, Katy. Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. Submitted to *Complexity*, Special issue on *Understanding Complex Systems*. Also available as cs.IR/0512085.
- Katy Börner. (2006) Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In Vladimir Geroimenko & Chaomei Chen (eds.) *Visualizing the Semantic Web*, Springer Verlag, 2nd Edition, chapter 11, pp. 183-198.
- Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. *Scientometrics,* 64(3), 351-374.
- Hook, Peter A. and Börner, Katy. (2005) Educational Knowledge Domain Visualizations: Tools to Navigate, Understand, and Internalize the Structure of Scholarly Knowledge and Expertise. In Amanda Spink and Charles Cole (eds.) *New Directions in Cognitive Information Retrieval*. Springer-Verlag, Netherlands, chapter 5, pp. 187-208.
- Börner, Katy, Dall'Asta, Luca, Ke, Weimao and Vespignani, Alessandro. (April 2005) Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *Complexity*, special issue on *Understanding Complex Systems*, 10(4): pp. 58 - 67. Also available as cond-mat/0502147.
- Ord, Terry J., Martins, Emília P., Thakur, Sidharth, Mane, Ketan K., and Börner, Katy. (2005) Trends in animal behaviour research (1968-2002): Ethoinformatics and mining library databases. *Animal Behaviour*, 69, 1399-1413. Supplementary Material.
- Mane, Ketan K. and Börner, Katy. (2004). Mapping Topics and Topic Bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5287-5290. Also available as cond-mat/0402380.
- Börner, Katy, Maru, Jeegar and Goldstone, Robert. (2004). The Simultaneous Evolution of Author and Paper Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl_1):5266-5273. Also available as cond-mat/0311459.




http://vw.indiana.edu/netsci06

# Extra Slides

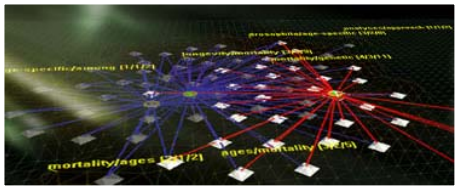
(not covered in the talk)



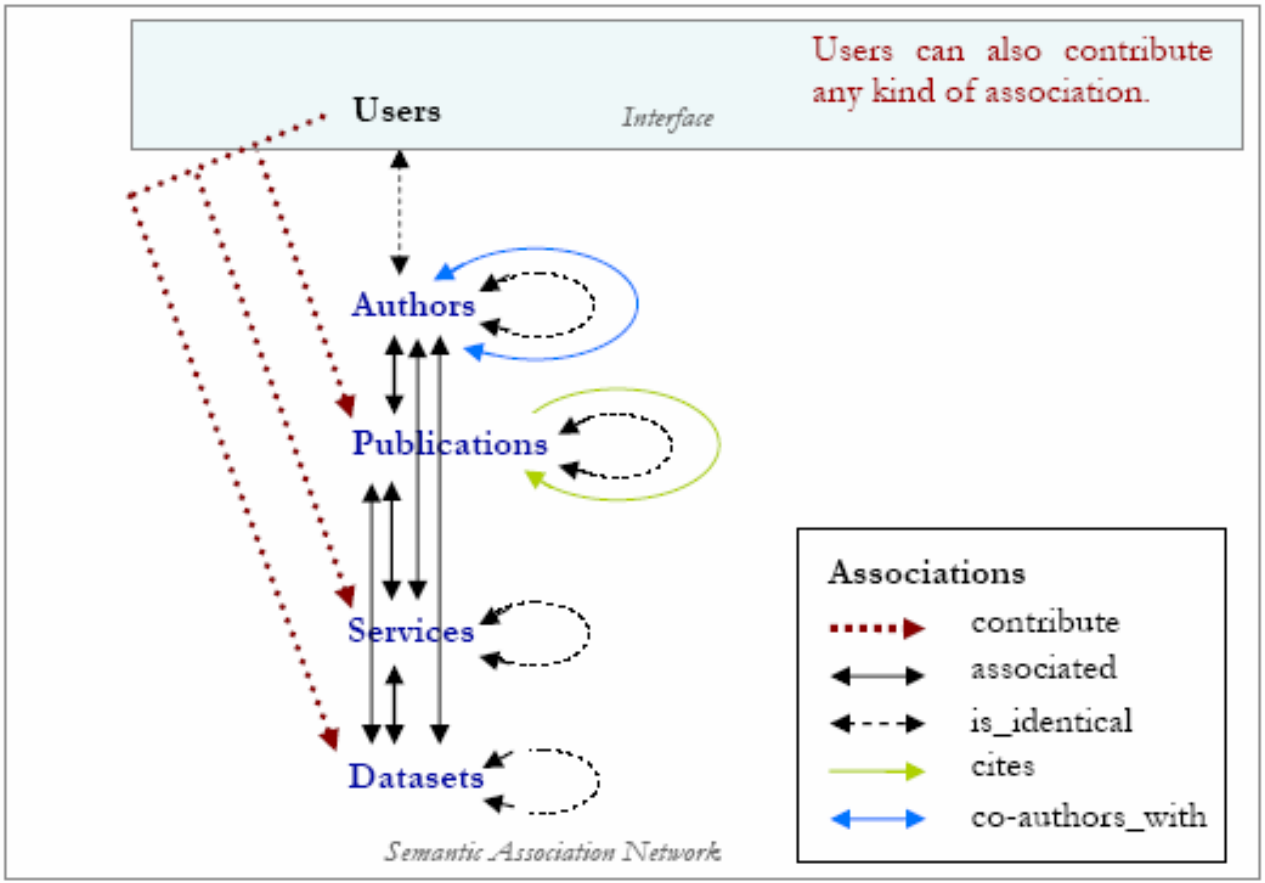
# Process of Mapping Knowledge Domains

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

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



# Semantic Association Networks




*Katy Börner. (2006) Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In Vladimir Geroimenko & Chaomei Chen (eds.) Visualizing the Semantic Web, Springer Verlag, 2nd Edition, chapter 11, pp. 183-198.*