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


Katy Börner & Shashikant Penumarthy
School of Library and Information Science
INDIANA UNIVERSITY
BLOOMINGTON
katy@indiana.edu

International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, July 24-28, 2005

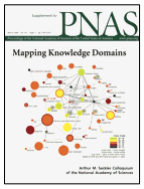
Outline

- Motivation
- Our Questions & Goals in This Study
- Related Work
- Dataset
- Data Analysis
- Results / Visualizations
- Discussion
- Future Work

Motivation

Knowledge domain visualizations help answer questions such as:

- What are the major research areas, experts, institutions, regions, nations, grants, publications, journals in xx research?
- Which areas are most insular?
- What are the main connections for each area?
- What is the relative speed of areas?
- Which areas are the most dynamic/static?
- What new research areas are evolving?
- Impact of xx research on other fields?
- How does funding influence the number and quality of publications?



Answers are needed by funding agencies, companies, and researchers.

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.

Diverse User Groups

- **Students** can gain an overview of a particular knowledge domain, identify major research areas, experts, institutions, grants, publications, patents, citations, and journals as well as their interconnections, or see the influence of certain theories.
- **Researchers** can monitor and access research results, relevant funding opportunities, potential collaborators inside and outside the fields of inquiry, the dynamics (speed of growth, diversification) of scientific fields, and complementary capabilities.
- **Grant agencies/R&D managers** could use the maps to select reviewers or expert panels, to augment peer-review, to monitor (long-term) money flow and research developments, evaluate funding strategies for different programs, decisions on project durations, and funding patterns, but also to identify the impact of strategic and applied research funding programs.
- **Industry** can use the maps to access scientific results and knowledge carriers, to detect research frontiers, etc. Information on needed technologies could be incorporated into the maps, facilitating industry pulls for specific directions of research.
- **Data providers** benefit as the maps provide unique visual interfaces to digital libraries.

- Last but not least, the availability of dynamically evolving maps of science (as ubiquitous as daily weather forecast maps) would dramatically improve the communication of scientific results to the **general public**.

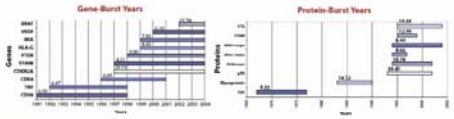
Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research

GOAL

To provide researchers, practitioners and students with a global map of a research domain, to help them answer questions such as: What are the major research areas, experts, institutions, regions, nations, grants, publications, journals in a certain area of research? Which areas are most insular? What are the main connections for each area? What is the relative speed of areas? What new research areas are evolving? How are the objects of study (e.g., genes, proteins) interconnected via papers?

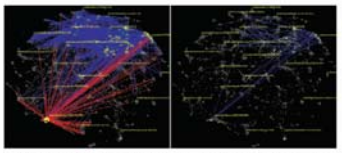
TOP-RESEARCHED GENES & PROTEINS

Identification of sudden interests in research/published papers on certain genes and proteins using Kleinberg's burst detection algorithm. The diagrams show the amount and the time spans of major burst for genes and proteins.



ASSOCIATION MAPS

A gene-gene, gene-paper, gene-protein, protein-paper and protein-protein map was generated.
The figure shows the gene-paper (left) and gene-gene (right) network. Highlighted in red is a single gene (CMM) and all its connections within the given network.



DATASETS

- 53804 Medline publication (1960 – Feb 2004)
- 299 Genes downloaded from *Entrez-Gene*
- 367 Proteins downloaded from Uniprot

PAPER-GENE-PROTEIN MAP

Shown here is the melanoma research area over the last 40 years. Gray dots represent papers, red dots denote proteins, blue dots indicate genes. Experts classified the shown research areas into two main categories:

- Applied Medical Sciences (left side) where research work occurs at organism level.
- Basic Molecular Sciences (right side) with studies related to genes and proteins.

TIME SERIES ANALYSIS

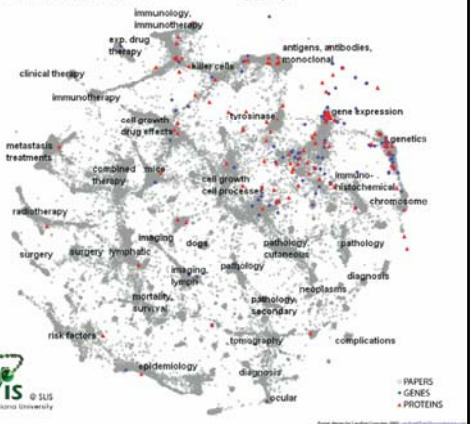
The structure & dynamics of melanoma research was examined:

1964-1973: Diagnostic and immunity based approaches dominate. Chemotherapy emerges as a new area for cancer treatment.

1974-1983: Chemotherapy gains popularity as viable treatment. Monoclonal studies involving tagging cancerous cells using antigens start.

1984-1993: Research on metastasis behavior of cancer dominates.

1994-2003: Gene-expression and mutation related studies gain popularity.



Kevin W. Boyack, Ketan K. Mane, Katy Börner (in press) *Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research*. Accepted for the Information Visualization Conference 2004.

For more information, contact Katy Börner at katy@indiana.edu.

This material is based upon work supported by the National Science Foundation under Grants No. IIS-0238261 and DUE-0333623.

InfoVis © SLIS Indiana University

Examining the Evolution and Distribution of Patent Classifications

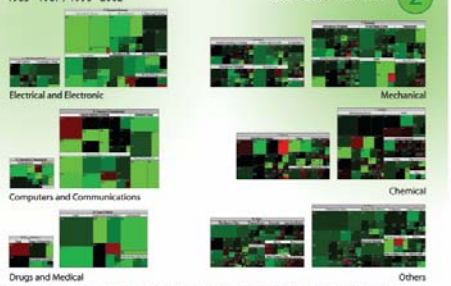
1 Patents Granted Over the Last 20 Years

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.

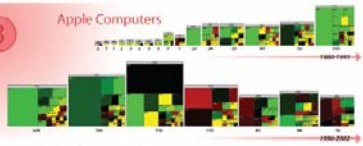
2 Fast Growth Domains / Slow Growth Domains

1983 1987 / 1998 2002 — 1983 - 1987 / 1998 - 2002



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.

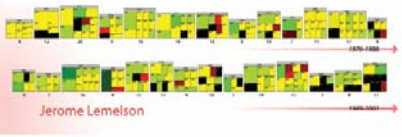
3 Apple Computers



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.

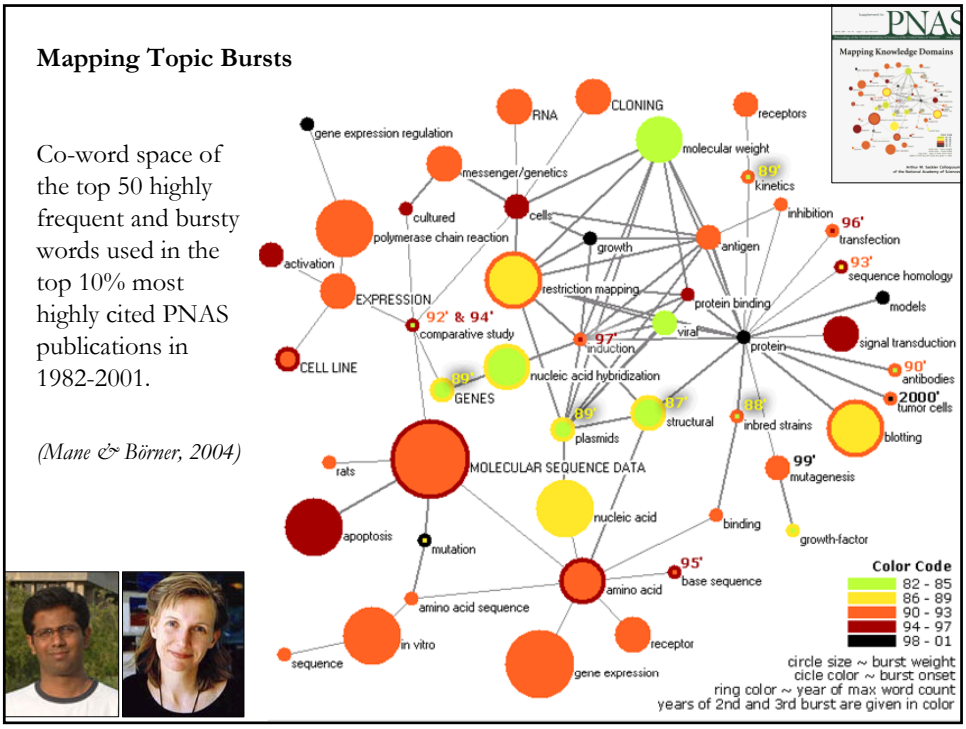
The material is based upon work supported bythe National Science Foundation under Grant No. IIS-0238261.

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

Mapping 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)



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

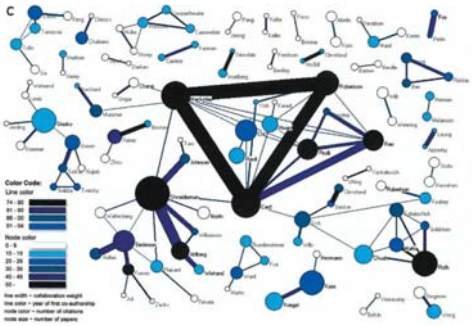
Börner, Dall'Asta, Ke & Vespignani (2005) Complexity, 10(4):58-67.

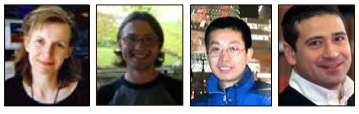
Research question:

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

Contributions:

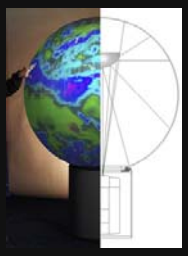
- New approach to allocate citational credit.
- Novel weighted graph representation.
- Visualization of the growth of weighted co-author network.
- Centrality measures to identify author impact.
- Global statistical analysis of paper production and citations in correlation with co-authorship team size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



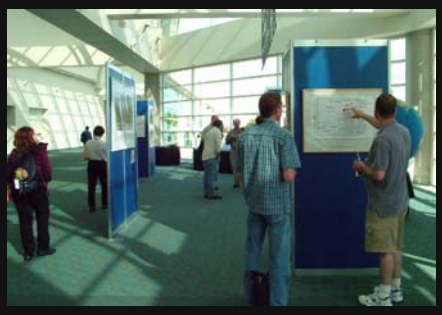


This physical & virtual science exhibit compares and contrasts first maps of our entire planet with the first maps of all of sciences.

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



Places & Spaces at ESRI in San Diego, CA



This Study

Our Questions & Goals in this Study

Questions

- Does space still matter in the Internet age?
- 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?
- Does the Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research instructions? **GUESS!**

Goals

- Identify geographically and statistically significant instances of institutions that act as major information sources,
- Correlate their behavior as *information sources* (number of citations their papers receive), *information sinks* (number of references to papers produced at other institutions), and *self-consumers* (number of self citations),
- Use direct citation linkage to identify their interrelation based on the amount of directly exchanged information, and
- Analyze and visualize the importance of proximity in geographic space for information exchange.

Related Work

The diffusion of tangible objects (people, goods, etc.) but also of intangible objects (ideas, activity levels, etc) has been studied in different fields

- Physics, e.g., heat diffusion,
- Robotics, e.g., communication among mobile robots (Arai, Yoshida et al. 1993),
- Social network analysis (Granovetter 1973; 2002),
- Bibliometrics/scientometrics/webometrics (Katz 1994; Thelwall 2002),
- Geography, e.g., migration studies (Ravenstein 1885; Thornwaite 1934; Tobler 1995), and
- Biology, e.g., neuronal migration in the nervous system (Thurner, Wick et al. 2002).

- Diverse activity, impact, and linkage measures exist to judge the research vitality or quality of research or to quantify the research contribution of institutions (Narin, Olivastro et al. 1994).
- Few studies have attempted to analyze the geographical concentration of highly cited authors, institutions, countries. Batty's (2003) work nicely shows that the distribution of citation counts is highly skewed, with most citations being associated with a few individuals working at a small number of institutions in an even smaller number of places and countries.

Dataset & Data Analysis

- Complete Proceedings of the National Academy of Sciences (PNAS) publications covering 1982-2001.
- Dataset contains 47,073 papers published by 18,994 unique authors, who work at 2,822 institutions (Institutions comprise academic institutions, research labs and corporate entities. An initial data cleaning step was performed to remove suffixes such as INC, MED.)
- Identify the 500 most cited research institutions.
- Compute spatial inter-citation patterns for four time slices.
- Determine information sources and sinks.
- Communicate results.

Identification of Unique Institutions & Their Properties

- Determine the number of citations received by each paper. (The paper most highly cited by papers within the set received 612 citations.)
- Attribution of citational credit to first author.
- Identification of the unique list of institutional affiliations of all first authors.
- Computation of spatial location for all institutions.
- Merge of institutions with identical names that are spatially close. (For example, Indiana University has several campuses.)
- Recalculation of total citation counts for merged institutions. (Indiana University as single entity might qualify to be in the top 500 most highly cited institution list, but when the campuses are split, none of the individual campuses might have the requisite number of citations.)

Problems

- Attribution of citational credit is incorrect.
- When should institutions be merged?
- Computation of locations for non-US institutions.

Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.

Unique Institution IDs

Problem

We could not find automatic means to compute locations for non-US institutions. Decision was made to exclude them. University of Tokyo received 1,797 and would have made the top 500 list.

The US 5-digit zip code assigns postal codes based on the position of a certain geographic location in a hierarchy of geographic significance based on area.

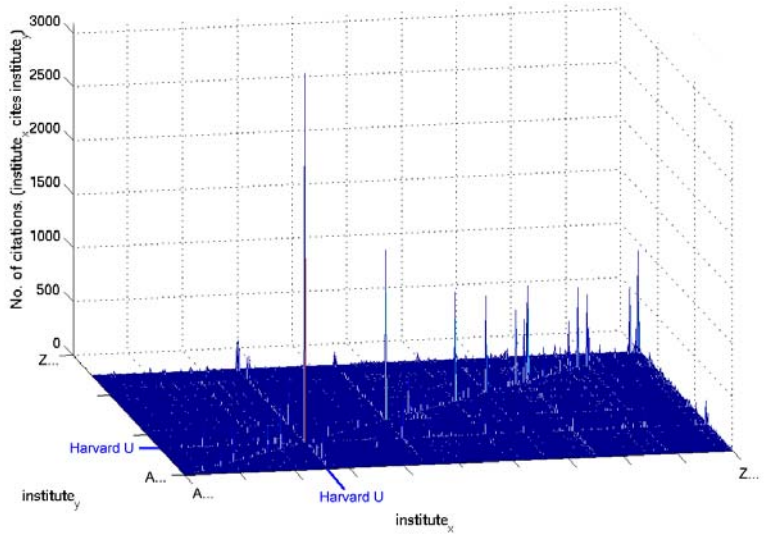
(Show Ben Fry's visualization.)

- 1st digit = major region
- 2nd & 3rd digit = state and county information.
- 4th & 5th digit = towns and cities within a county.

A unique ID was created for each institution by concatenating the (abbreviated) name of the institution with its zip code, e.g., INDIANA UNIV47401 and INDIANA UNIV47405 were collapsed into INDIANA UNIV47401. Aims is to compromise between maintaining geographic identity and statistical significance.

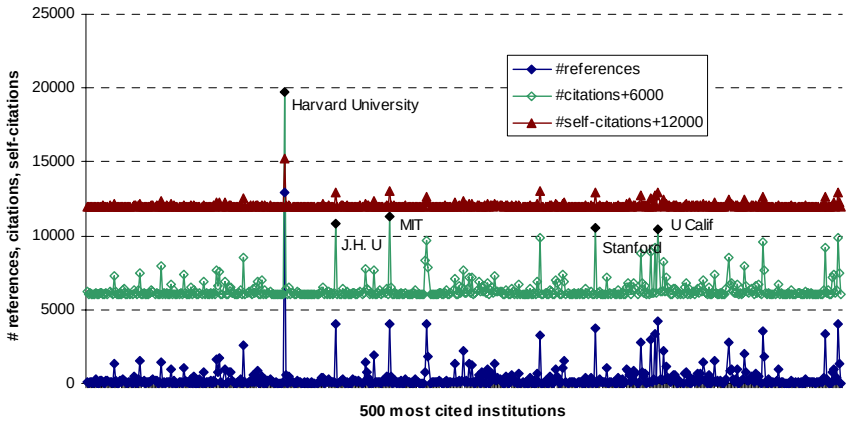
Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.

Inter-citation Matrix for Top-500 US Institutions



Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.

Total Citation Counts for Top-500 US Institutions



x-axis is sorted alphabetically.

Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.

Visualization Using ArcGIS

- Map of U.S. is used as base map/reference system.
- States are color coded based on the population size in the year 2000. Lighter shades of green represent lower population.
- Overlaid are the top 500 institutions.

- Citations per institution (excluding self citations) are represented by a 'citation stick'.
- The stick height is a function of the normalized number of citations received by a certain institution in relation to the maximum number of citations that any institution received:

$$height = \left(\sin\left(\frac{\#\,citations}{\max\#\,citations}\right)+1\right)*k.$$

The utilization of sin guaranties that small differences between institutions with low citation counts are visible and that the huge differences among the institutions with high citation counts is less distorting. k is a scaling factor.

Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.

Visualization Using ArcGIS

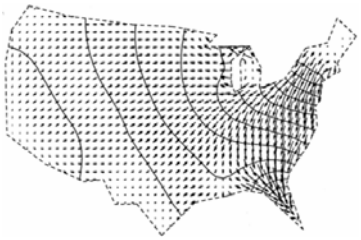
Geographic location and number of received citations for the top 500 U.S. institutions.

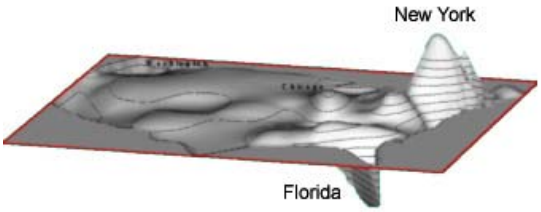
Using Tobler's (1995) analogy of flow of energy in a vector potential field, highly cited institutions exhibit a high pressure for the diffusion of information whereas other institutions are mostly importing information and hence act as information sinks.

Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.

Visualizations by Tobler

- Migration potentials are shown as contours (left).
- The pressure to move in US based on a continuous spatial gravity model (right).





Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.

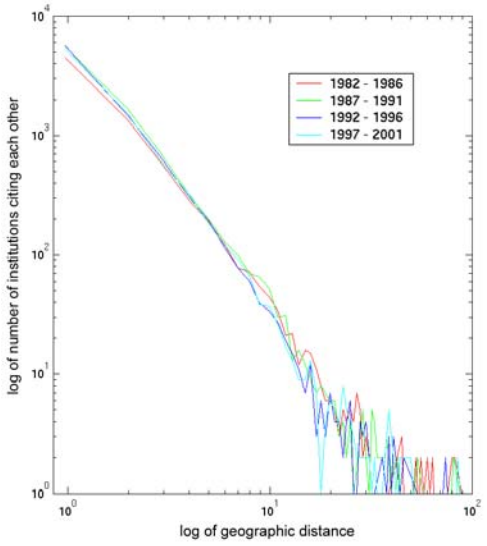
Geospatial Distribution of Citations

Log-log plot showing the variation of the number of institutions that cite each other over geographic distance among them for four time slices.

Distance was calculated by applying the Euclidean form formulae to xy coordinates obtained using the Albers projection.

1.5 units of geographic distance equal approximately 100 miles.

$\gamma_{82\text{-}86} = 1.94$ (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%)



Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.

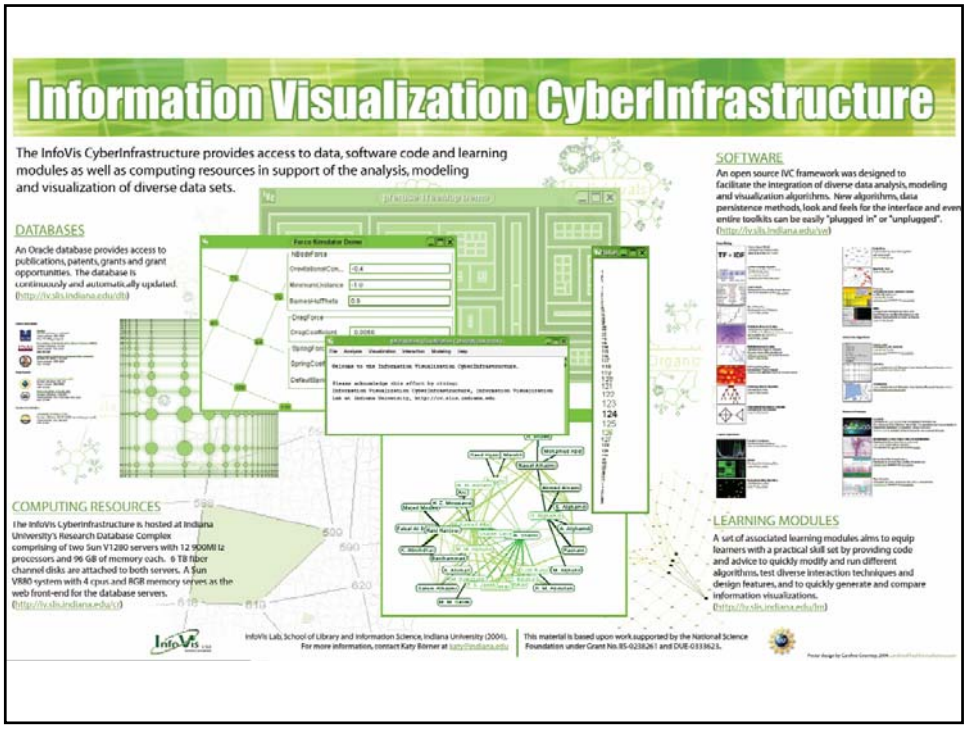
Discussion

- Novel approach to analyzing the dual role of institutions as information producers and consumers and to study the diffusion of information among them.
- Application of the approach to a large scale 20-year data set.

  While the PNAS data set nicely represents major research results from diverse areas of science over a 20 year time span, it **does not cover any specific discipline completely** nor does it represent **any authors' entire life work**. In the PNAS data set (and most other publication data sets) **there is no means to attribute a certain percentage of a paper to each co-author** (and his/her institution). **Non-U.S. institutions had to be excluded** from this analysis as no information about their longitude/latitude information was available to us. Obviously, the number of co-authorships or co-PI-ships, co-citations of papers, and co-occurrence of words in papers are **additional valid indicators** for information diffusion among institutions.
- Counterintuitive result: Place does matter even in the Internet age.

Future Work

- Try to explain why the Internet does not lead to a more global citation behavior. Reasons for local collaborations might comprise 'winner takes all' funding schemes, the demands of complex, large-scale instrumentation, and the need to gain experience, train researchers, and sponsor protégés, see also (Katz 1994), p. 32 or the importance of social linkages for making citation linkages (Wellman, White, Nazer, 2004).
- Apply the approach to studying the dual role of authors as information producers and consumers or the diffusion of information among companies via publication and patent citations, email exchanges, etc.
- Develop techniques that can visualize diffusion patterns among many different static or moving instances. A first attempt to visualize social diffusion patterns was made in (Börner and Penumarthy 2003). Future work will address the analysis and visualization of diffusion patterns of tangible and intangible objects over space and time.
- Visualize growth processes in an effective yet aesthetically pleasing way.



References

- Boyack, Kevin W., Klavans, R. and Börner, Katy. (in press). Mapping the Backbone of Science. *Scientometrics*.
- Hook, Peter A. and Börner, Katy. (in press) 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.
- Katy Börner. (in press) 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.
- 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.

Katy Börner & Shashikant Penumarthy, Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.