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

College of Computer and Information Science
366 West Village H
Northeastern University
Boston, MA

February 3, 2012

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.
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.

**Find your way**
- Take terra bytes of data
- Descriptive & Predictive Models
- Find collaborators, friends
- Identify trends
Type of Analysis vs. Level of Analysis

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

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

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
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.
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

Mapping Transdisciplinary Tobacco Use Research Centers Publications

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.

Research questions:
1. Does space still matter in the Internet age?
2. 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?
3. Does the Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research institutions?

Contributions:
- Answer to Qs 1 + 2 is YES.
- Answer to Qs 3 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.
Mapping Science Exhibit at NEU

From left to right: Thomas Urell (Communications Officer, NEU Libraries), David Lazer (Associate Professor of Political Science and Computer Science), Katy Börner, William Wakeling (Dean, University Libraries).

Exhibit has been on display at more than 200 venues in 19 countries on 6 continents.

References


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 data 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.

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.
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.

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?

**Geospatial Analysis (Where)** Where is what science performed by whom? Science is global and needs to be studied globally.
VIVO On-The-Go

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

Download Data

General Statistics
- 36 publication(s) from 2001 to 2010 (.CSV File)
- 80 co-author(s) from 2001 to 2010 (.CSV File)

Co-Author Network (.GraphML File)

Save as Image (.PNG file)

Tables
- Publications per year (.CSV File)
- Co-authors (.CSV File)

http://vivo.iu.edu/vis/author-network/person25557
Run Sci2 Tool and Load Co-Author Network (GraphML File)

Visualize the file using Radial Graph layout.

Click on node to focus on it.
Hover over a node to highlight its co-authors.

Code and tutorials are linked from http://sci2.wiki.cns.edu

Develop VIVO Visualizations

See also Visualization in VIVO Workshop on Aug 24, 2011
http://wiki.cns.in.edu/display/PRES/VIVO+Presentation
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 data 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.
Needs-Driven Workflow Design using a modular data acquisition/analysis/modeling/visualization pipeline as well as modular visualization layers.
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)
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 110,000 times since December 2006.

### Computational Proteomics

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


![Drug-target network diagram](image)

**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 size of the drug (pentagon) 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?


![Product space diagram](image)

**Fig. 1.** The product space. (A) Hierarchically clustered proximity (a matrix representing the 175 IFC-4 product classes exported in the 1999–2005 period. (B) Network representation of the product space. Links are colored 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 Isenberg.
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.


Sci² Tool – “Open Code for S&T Assessment”

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


All 151 NEU NSF Awards active on Feb 2, 2012
Downloaded from http://www.nsf.gov/awardsearch
Co-PI Network
Edges go from lead PI to Co-PIs. All nodes with degree higher than 2 are labeled.
Many small and one larger collaboration network:

Bimodal Network of NSF Organization and Projects
Edges go from NSF organization (green) to project (red).
All NSF organization nodes are labeled.

Civil, Mechanical, and Manufacturing Innovation (CMMI)
Electrical, Communications and Cyber Systems (ECCS)
Computing and Communication Foundations (CCF)
Computer and Network Systems (CNS)
Chemical, Bioengineering, Environmental, and Transport Systems (CBET)
Mathematical Sciences (DMS)
Wordle.net of “Interest to Learn” response by users from more than 40 countries
A number of other projects recently adopted OSGi and/or CIShell:

- **Cytoscape** ([http://cytoscape.org](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](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.

- **Taverna Workbench** ([http://taverna.org.uk](http://taverna.org.uk)) Developed by the myGrid team ([http://mygrid.org.uk](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](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](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.lec.uno.es](http://sisob.lec.uno.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.