Analysis and Visualization of Science

What is science?

Why do we analyze and visualize science?

How do we analyze and visualize science?
Authors are mortal. Papers are immortal. Monsters = ‘the unknown’ or voids. Impact of funding on science (yellow). Good and bad years.
Science as accumulation of knowledge.
“Scholarly brick laying”.
Standing on the shoulders of giants.
Densely knit communities.
The importance of weak links.
Areas of science are tube shaped.
Crust of science can represent “funding” or “usage”.
This drawing attempts to show the “structure” of science. Many are interested to understand the “dynamics” of science.
Why Map Science?
Cartographic maps of physical places have guided mankind’s explorations for centuries.

They enabled the discovery of new worlds while also marking territories inhabited by the unknown.

Without maps, we would be lost.
Domain maps of abstract semantic spaces aim to serve today’s explorers navigating the world of science.

These maps are generated through a scientific analysis of large-scale scholarly datasets in an effort to connect and make sense of the bits and pieces of knowledge they contain.

They can be used to identify objectively major research areas, experts, institutions, collections, grants, papers, journals, and ideas in a domain of interest. Science maps can provide overviews of “all-of-science” or of a specific area.

They can show homogeneity vs. heterogeneity, cause and effect, and relative speed. They allow us to track the emergence, evolution, and disappearance of topics and help to identify the most promising areas of research.
Information Needs for Science Map User Groups

Advantages for Funding Agencies
- 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.

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.

Analysis and Visualization of Science
<table>
<thead>
<tr>
<th>Type of Analysis vs. Scale of Level of Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Micro/Individual (1-100 records)</strong></td>
</tr>
<tr>
<td><strong>Statistical Analysis/Profiling</strong></td>
</tr>
<tr>
<td><strong>Temporal Analysis (When)</strong></td>
</tr>
<tr>
<td><strong>Geospatial Analysis (Where)</strong></td>
</tr>
<tr>
<td><strong>Topical Analysis (What)</strong></td>
</tr>
<tr>
<td><strong>Network Analysis (With Whom?)</strong></td>
</tr>
</tbody>
</table>
### Process of Computational Scientometrics

<table>
<thead>
<tr>
<th>Data Extraction</th>
<th>Unit of Analysis</th>
<th>Measures</th>
<th>Layout (often one code does both similarity and ordination steps)</th>
<th>Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searches</td>
<td>Common Choices</td>
<td>Counts/Frequencies</td>
<td>Similarity: Scalar (unit by unit matrix)</td>
<td>Ordination: Dimensionality Reduction: Eigenvector/Eigenvalue solutions</td>
</tr>
<tr>
<td></td>
<td>• ISI</td>
<td>• Attributes (eg, terms)</td>
<td>• Direct citation</td>
<td>• Factor Analysis (FA) and Principal Components Analysis (PCA)</td>
</tr>
<tr>
<td></td>
<td>• INSPEC</td>
<td>• Author citations</td>
<td>• Co-citation</td>
<td>• Multi-dimensional scaling (MDS)</td>
</tr>
<tr>
<td></td>
<td>• Eng Index</td>
<td>• Co-citations</td>
<td>• Combined linkage</td>
<td>• LSA</td>
</tr>
<tr>
<td></td>
<td>• Medline</td>
<td>• By year</td>
<td>• Co-word/co-term</td>
<td>• Pathfinder networks (PFNet)</td>
</tr>
<tr>
<td></td>
<td>• ResearchIndex</td>
<td>• Thresholds</td>
<td>• Co-classification</td>
<td>• Self-organizing maps (SOM) incl. SOM, ET-maps, etc.</td>
</tr>
<tr>
<td></td>
<td>• Patents</td>
<td>• By counts</td>
<td>• Latent Semantic Analysis</td>
<td>• Cluster analysis</td>
</tr>
<tr>
<td></td>
<td>• etc.</td>
<td>Vector (unit by attribute matrix)</td>
<td>(words/terms) incl. Singular Value Decomposition (SVD)</td>
<td></td>
</tr>
<tr>
<td>Broadening</td>
<td>Journal</td>
<td>Correlation (if desired)</td>
<td></td>
<td>Analysis</td>
</tr>
<tr>
<td>By citation</td>
<td>Document</td>
<td>• Pearson’s R on any of above</td>
<td></td>
<td></td>
</tr>
<tr>
<td>By terms</td>
<td>Author</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


### Computational Scientometrics: Studying Science by Scientific Means


- Places & Spaces: Mapping Science exhibit, see also [http://scimaps.org](http://scimaps.org).


Science of Science Cyberinfrastructure

Overview

What cyberinfrastructure will be required to measure, model, analyze, and communicate scholarly data and, ultimately, scientific progress?

This talk presents our efforts to create a science of science cyberinfrastructure that supports:

• Data access and federation via the Scholarly Database, http://sdb.slis.indiana.edu,

• Data preprocessing, modeling, analysis, and visualization using plug-and-play cyberinfrastructures such as the Sci2 Tool, http://sci2.cns.iu.edu, and

• Communication of science to a general audience via the Mapping Science Exhibit at http://scimaps.org.

The following demos should be particularly interesting for those interested to

• Map their very own domain of research,

• Test and compare data federation, mining, visualization algorithms on large scale datasets,

• Use advanced network science algorithms in their own research.
The Scholarly Database at Indiana University provides free access to 25,000,000 papers, patents, and grants. Since March 2009, users can also download networks, e.g., co-author, co-investigator, co-inventor, patent citation, and tables for burst analysis.
Explicitly designed for SoS research and practice, well documented, easy to use.
Empowers many to run common studies while making it easy for exports to perform novel research.
Advanced algorithms, effective visualizations, and many (standard) workflows.
Supports micro-level documentation and replication of studies.
Is open source—anybody can review and extend the code, or use it for commercial purposes.

SUMMARY
- Existing metrics have known flaws
- A reliable, open, joined-up data infrastructure is needed
- Data should be collected on the full range of scientists' work
- Social scientists and economists should be involved

Acknowledgments
This work is supported in part by the Cyberinfrastructure for Network Science center and the School of Library and Information Science at Indiana University, the National Science Foundation under Grant No. 0855391 and HSS-0719690, and the James S. McDonnell Foundation.
Supported Input file formats:
- GraphML (*.xml or *.graphml)
- XGMML (*.xml)
- Pajek .NET (*.net) & Pajek .Matrix (*.mat)
- NWB (*.nwb)
- TreeML (*.xml)
- Edge list (*.edge)
- CSV (*.csv)
- ISI (*.isi)
- Scopus (*.scopus)
- NSF (*.nsf)
- Bibtext (*.bib)
- Endnote (*.enw)

Output file formats:
- GraphML (*.xml or *.graphml)
- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)
- CSV (*.csv)

Network Extraction

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

---

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

**Individual Level**
A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks (p. 54-65)
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U (p. 49-53)

**Institution Level**

**Scientific Field Level**
D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data (p. 77-85)
All papers, maps, cyberinfrastructures, talks, press are linked from http://cns.iu.edu