Visual Analytics: Mining, Mapping, and Accelerating Science and Technology

Katy Börner
Director, Cyberinfrastructure for Network Science Center
School of Informatics and Computing, Indiana University, USA
Science, Technology and Innovation Visiting Research Fellow at OECD, France

Presentation at EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR RESEARCH & INNOVATION
in Brussels, Belgium

June 23, 2014

Language Communities of Twitter - Eric Fischer - 2012

Descriptive & Predictive Models

Terra bytes of data

Find your way

Find collaborators, friends

Identify trends
Descriptive Models

Multiple levels: Micro ... Macro


Different Levels of Abstraction/Analysis

- Macro/Global Population Level
- Meso/Local Group Level
- Micro Individual Level
### Type of Analysis vs. Level of Analysis

<table>
<thead>
<tr>
<th></th>
<th>Micro/Individual (1-100 records)</th>
<th>Meso/Local (101–100,000 records)</th>
<th>Macro/Global (100,000 &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>
Individual Co-PI Network
Ke & Börner (2006)

Mapping the Evolution of Co-Authorship Networks
Mapping the Evolution of Co-Authorship Networks

Mapping Transdisciplinary Tobacco Use Research Centers Publications

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

Supported by NIH/NCI Contract HHSN261200800812
Mapping Indiana’s Intellectual Space
Ke, Light & Börner (2010)

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

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.

Acceleration in the Co-development of Patented Technologies, 1996-2001 and 2006-11, see OECD Scoreboard 2013, p.171
**Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions**


**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.

---

The Global 'Scientific Food Web'


**Contributions:**
- Comprehensive global analysis of scholarly knowledge production and diffusion on the level of continents, countries, and cities.
- Quantifying knowledge flows between 2000 and 2009, we identify global sources and sinks of knowledge production. Our knowledge flow index reveals, where ideas are born and consumed, thereby defining a global 'scientific food web'.
- While Asia is quickly catching up in terms of publications and citation rates, we find that its dependence on knowledge consumption has further increased.
Country Mobility Network, 1996-2011, see OECD Scoreboard 2013, p. 62
### Type of Analysis vs. Level of Analysis

<table>
<thead>
<tr>
<th></th>
<th>Micro/Individual (1-100 records)</th>
<th>Meso/Local (101–100,000 records)</th>
<th>Macro/Global (100,000 &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 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></td>
<td>PNAS publications</td>
</tr>
<tr>
<td><strong>Topical Analysis (What?)</strong></td>
<td>Base knowledge from one individual</td>
<td></td>
<td>VxOrd/Topic NIH funding</td>
</tr>
<tr>
<td><strong>Network Analysis (With Whom?)</strong></td>
<td>NSF Co-PI network of one researcher</td>
<td>Co-author network NIH's core competency</td>
<td></td>
</tr>
</tbody>
</table>

### Predictive Models (Why?)

Example: Collective allocation of science funding as an alternative to peer review
From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review

Existing (left) and proposed (right) funding systems. Reviewers in blue; investigators in red.

In the proposed system, all scientists are both investigators and reviewers: every scientist receives a fixed amount of funding from the government and discretionary distributions from other scientists, but each is required in turn to redistribute some fraction of the total they received to other investigators.

Current Model is Expensive:
If four professors work four weeks full-time on a proposal submission, labor costs are about $30k [1]. With typical funding rates below 20%, about five submission-review cycles might be needed resulting in a total expected labor cost of $150k. The average NSF grant is $128k per year. U.S. universities charge about 50% overhead (ca. $42k), leaving about $86k.

In other words, the four professors lose $150k-$86k = $64k of paid research time by obtaining a grant to perform the proposed research.

To add: Time spent by researchers to review proposals. In 2012 alone, NSF convened more than 17,000 scientists to review 53,556 proposals.

From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review


Assume
Total funding budget in year $y$ is $t_y$
Number of qualified scientists is $n$

Each year,
the funding agency deposits a fixed amount into each account, equal to the total funding budget divided by the total number of scientists: $t_y/n$.
Each scientist must distribute a fixed fraction, e.g., 50%, of received funding to other scientists (no self-funding, COIs respected).

Result
Scientists collectively assess each others’ merit based on different criteria; they “fund-rank” scientists; highly ranked scientists have to distribute more money.

Example:
Total funding budget per year is 2012 NSF budget
Given the number of NSF funded scientists, each receives a $100,000 basic grant.
Fraction is set to 50%

In 2013, scientist $S$ receives a basic grant of $100,000 plus $200,000 from her peers, i.e., a total of $300,000.
In 2013, $S$ can spend 50% of that total sum, $150,000, on her own research program, but must donate 50% to other scientists for their 2014 budget.

Rather than submitting and reviewing project proposals, $S$ donates directly to other scientists by logging into a centralized website and entering the names of the scientists to donate to and how much each should receive.
From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review


Model Run and Validation:

It uses citations as a proxy for how each scientist might distribute funds in the proposed system.

Dataset: 37M articles from TR 1992 to 2010 Web of Science (WoS) database with 770M citations and 4,195,734 unique author names. The 867,872 names who had authored at least one paper per year in any five years of the period 2000–2010 were used in validation.

For each pair of authors we determined the number of times one had cited the other in each year of our citation data (1992–2010).

NIH and NSF funding records from IU’s Scholarly Database provided 347,364 grant amounts for 109,919 unique scientists for that time period.

Simulation run begins in year 2000, in which every scientist was given a fixed budget of $B = 100k. In subsequent years, scientists distribute their funding in proportion to their citations over the prior 5 years.

The model yields funding patterns similar to existing NIH and NSF distributions.
Visualizing STI Model Results

Example: Places & Spaces: Mapping Science Exhibit

Mapping Science Exhibit on display at MEDIA X, Stanford University
Map of Scientific Collaborations from 2005-2009


Language Communities of Twitter

Language Communities of Twitter - Eric Fischer - 2012

Illuminated Diagram Display on display at the Smithsonian in DC. 
http://scimaps.org/exhibit_info/#ID
Science Maps in “Expedition Zukunft” science train visiting 62 cities in 7 months 12 coaches, 300 m long Opening was on April 23rd, 2009 by German Chancellor Merkel

http://www.expedition-zukunft.de
Places & Spaces Digital Display in North Carolina State’s brand new Immersion Theater

Places & Spaces: Mapping Science Exhibit
http://scimaps.org

Maps are available for sale and the exhibit can be hosted by anyone.
Visualizing STI Model Results

Example: The Information Visualization MOOC

The Information Visualization MOOC
ivmooc.cns.iu.edu

Students from more than 100 countries
350+ faculty members
#ivmooc

Course Schedule

- **Session 1** – Workflow design and visualization framework
- **Session 2** – “When:” Temporal Data
- **Session 3** – “Where:” Geospatial Data
- **Session 4** – “What:” Topical Data

**Mid-Term**

**Students work in teams with clients.**
- **Session 5** – “With Whom:” Trees
- **Session 6** – “With Whom:” Networks
- **Session 7** – Dynamic Visualizations and Deployment

**Final Exam**

Final grade is based on Midterm (30%), Final (40%), Client Project (30%).
Needs-Driven Workflow Design

**Stakeholders**

**Types and levels of analysis** determine data, algorithms & parameters, and deployment

**Data**

**READ** ➔ **ANALYZE** ➔ **VISUALIZE** ➔ **DEPLOY**

- **Validation**
- **Interpretation**

Visually encode data ➔ Overlay data ➔ Select visualiz. type

**Types and levels of analysis**

**Data**

**READ** ➔ **ANALYZE** ➔ **VISUALIZE** ➔ **DEPLOY**

- **Validation**
- **Interpretation**

Visually encode data ➔ Overlay data ➔ Select visualiz. type
Clients

http://ivmooc.cns.iu.edu/clients.html

Diogo Carmo
Tutorial at OECD: “Open Source Tools for S&T Data Analysis and Visualization”

**Speaker:** Katy Börner  
**Date:** 25 June, 2014  
**Time:** 9:30-13:30  
**Venue:** OECD Conference Centre, Room E, 2 rue André Pascal, Paris 75116

This tutorial is designed for researchers and practitioners interested to use advanced data mining algorithms and visualizations in their research and daily decision making. It gives an overview of open source tools for the analysis and visualization of science and technology (S&T) data. A specific focus is the Science of Science (Sci2) Tool that supports temporal, geospatial, topical, and network analysis and visualization of scholarly datasets at the micro (individual), meso (local), and macro (global) levels. Open data from OECD and other government agencies will be used to demonstrate different analysis and visualization workflows.

The tutorial has two parts: Part 1 provides an overview of diverse international efforts to (1) standardize and federate micro-level datasets of S&T activity, e.g., publication, patent, grant, social media data; (2) design open code tools and online services that are interoperable; (3) develop means to share and teach open datasets and tools. Part 2 is reserved for “hands-on” training. If you plan to attend this part, please bring your laptop and pre-install the Sci2 (v 1.1 beta) tool prior to the workshop—the tool is freely available at [http://sci2.cns.iu.edu](http://sci2.cns.iu.edu).

For a preview see: [http://cns.iu.edu/docs/presentations/2012-borner-sci2tutorial-oecd.pdf](http://cns.iu.edu/docs/presentations/2012-borner-sci2tutorial-oecd.pdf)
Load **One** File and Run **Many** Analyses and Visualizations

<table>
<thead>
<tr>
<th>Times Cited</th>
<th>Publication Year</th>
<th>City of Publisher</th>
<th>Country</th>
<th>Journal Title (Full)</th>
<th>Subject Category</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2011</td>
<td>NEW YORK</td>
<td>USA</td>
<td>COMMUNICATIONS OF THE ACM</td>
<td>Computer Science</td>
<td>Borner, K</td>
</tr>
</tbody>
</table>

Co-author and many other bi-modal networks.
References


All papers, maps, tools, talks, press are linked from http://cns.iu.edu
These slides will soon be at http://cns.iu.edu/docs/presentations

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