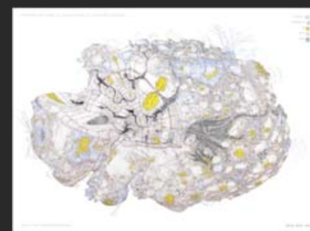
Models of Science

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

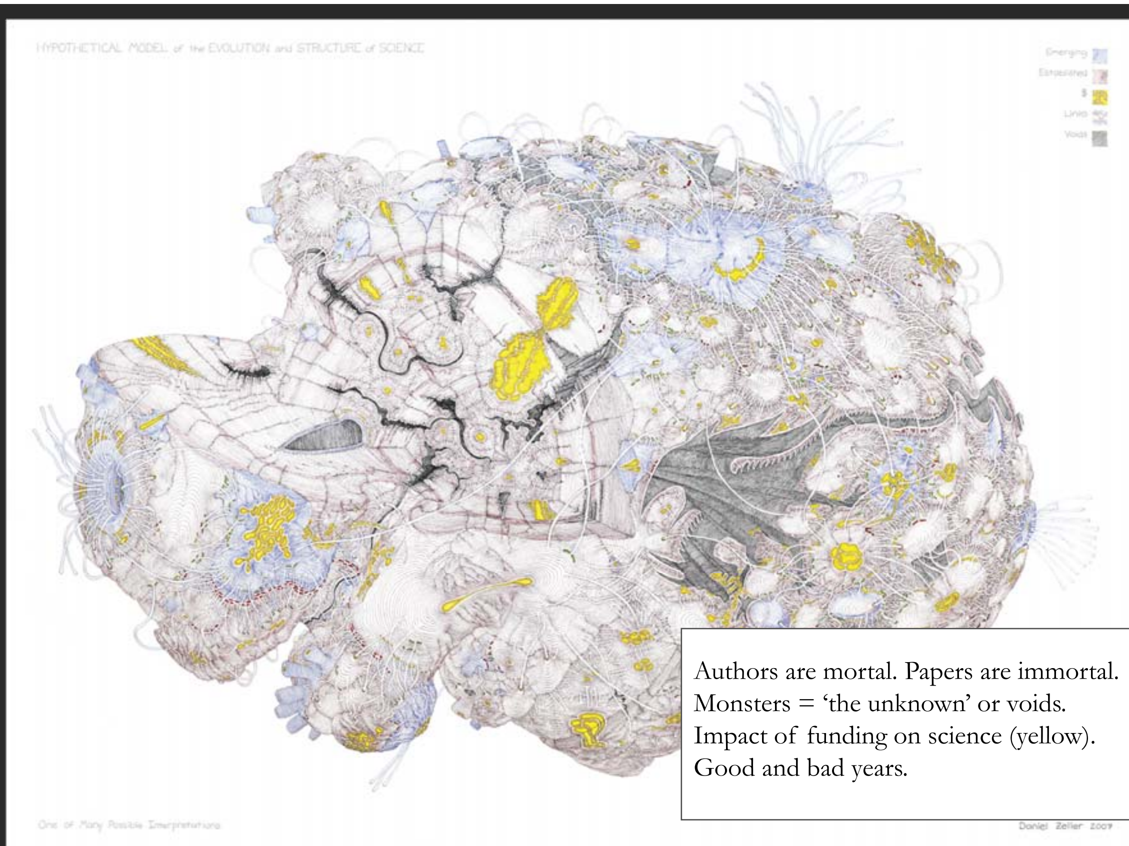
With special thanks to the members at the Cyberinfrastructure for Network Science Center.
Other collaborators are mentioned in slides.

NEH Summer Institute - Distinguished Lecture Series
UNC Charlotte, Charlotte, NC
June 2, 2011



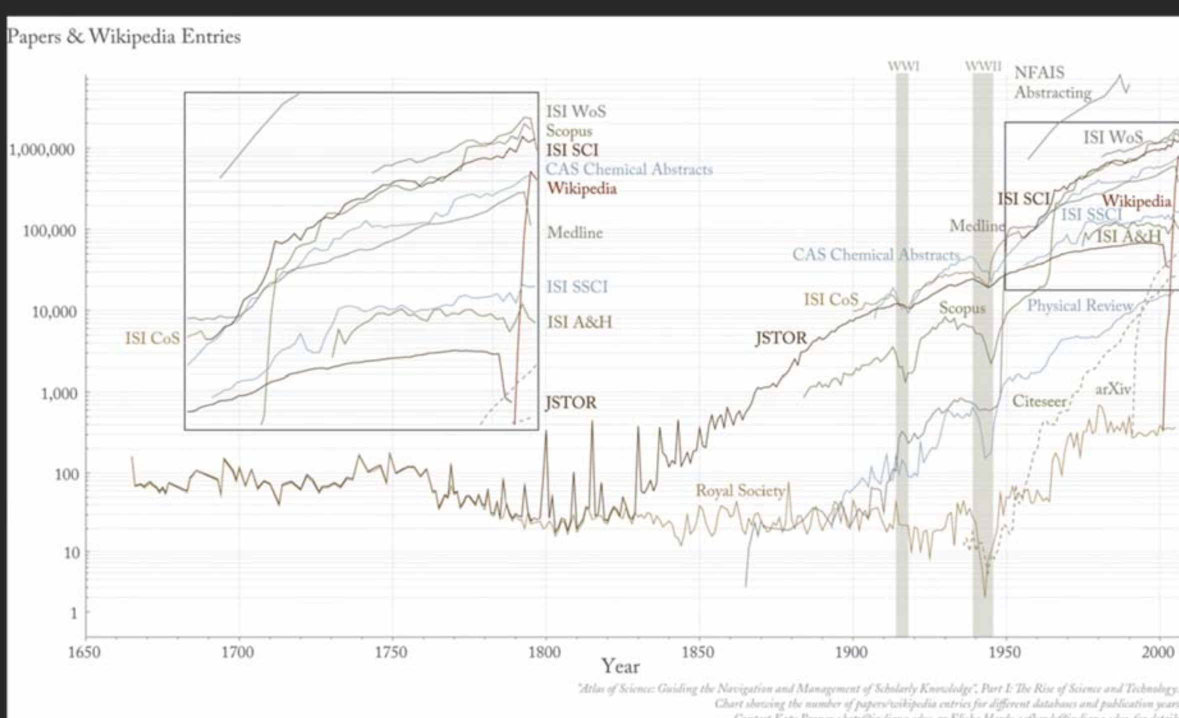
Four Parts:

1. **Conceptualizing Science**
2. Model Inspirations from Other Sciences
3. Models of Science
4. Tools to Model and Map Science

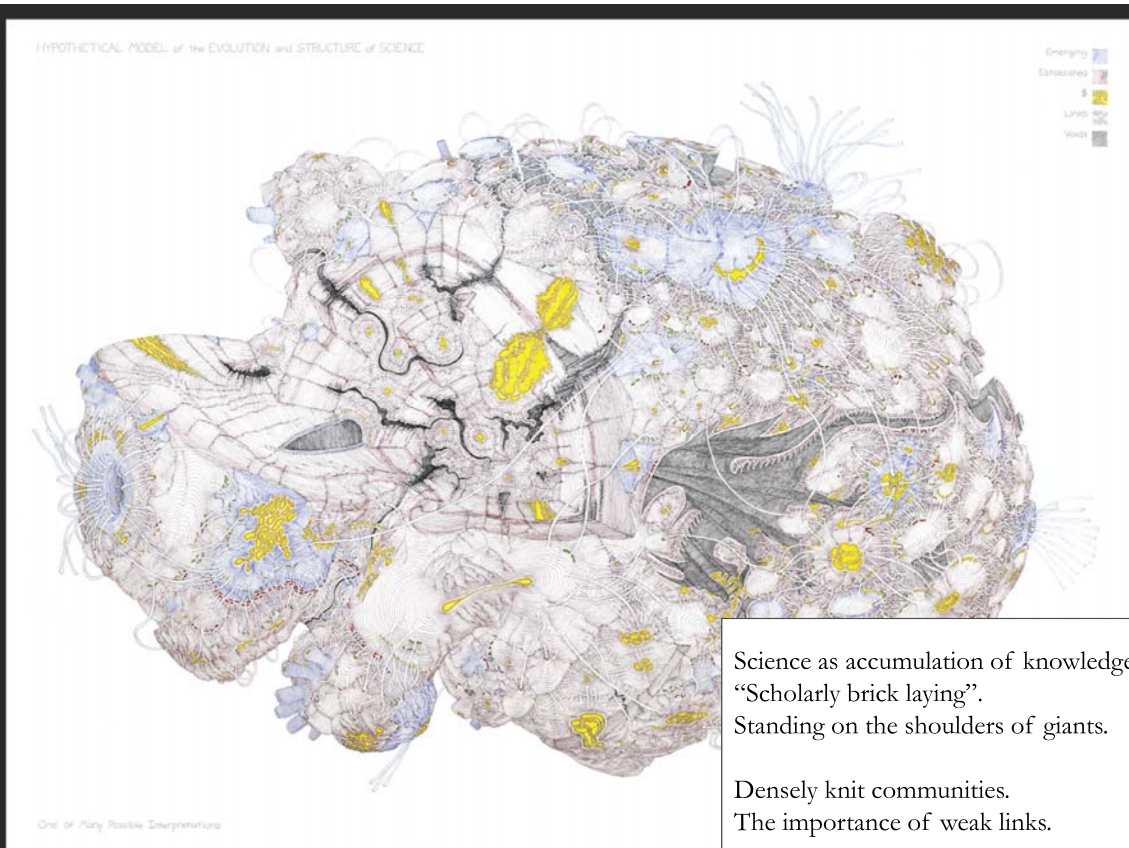


Authors are mortal. Papers are immortal.
Monsters = 'the unknown' or voids.
Impact of funding on science (yellow).
Good and bad years.

Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007



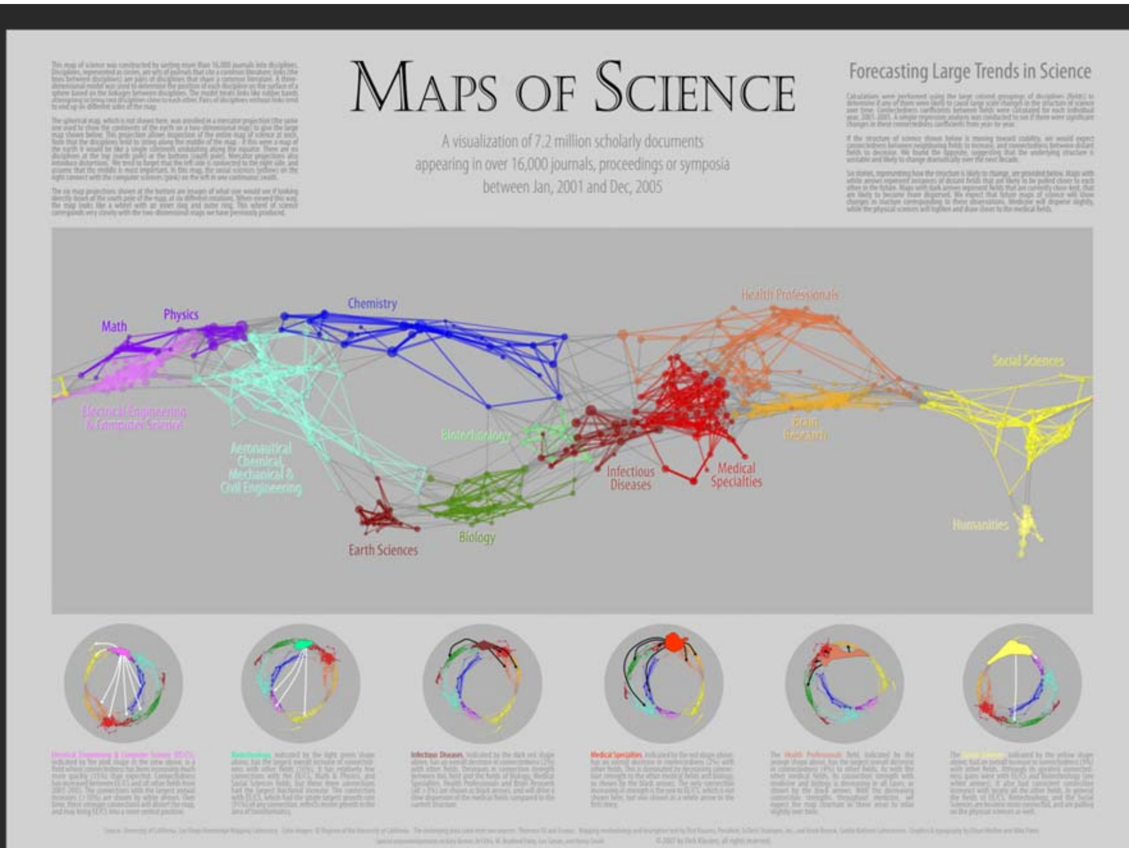
Atlas of Science - Katy Borner - 2010



Science as accumulation of knowledge.
“Scholarly brick laying”.
Standing on the shoulders of giants.

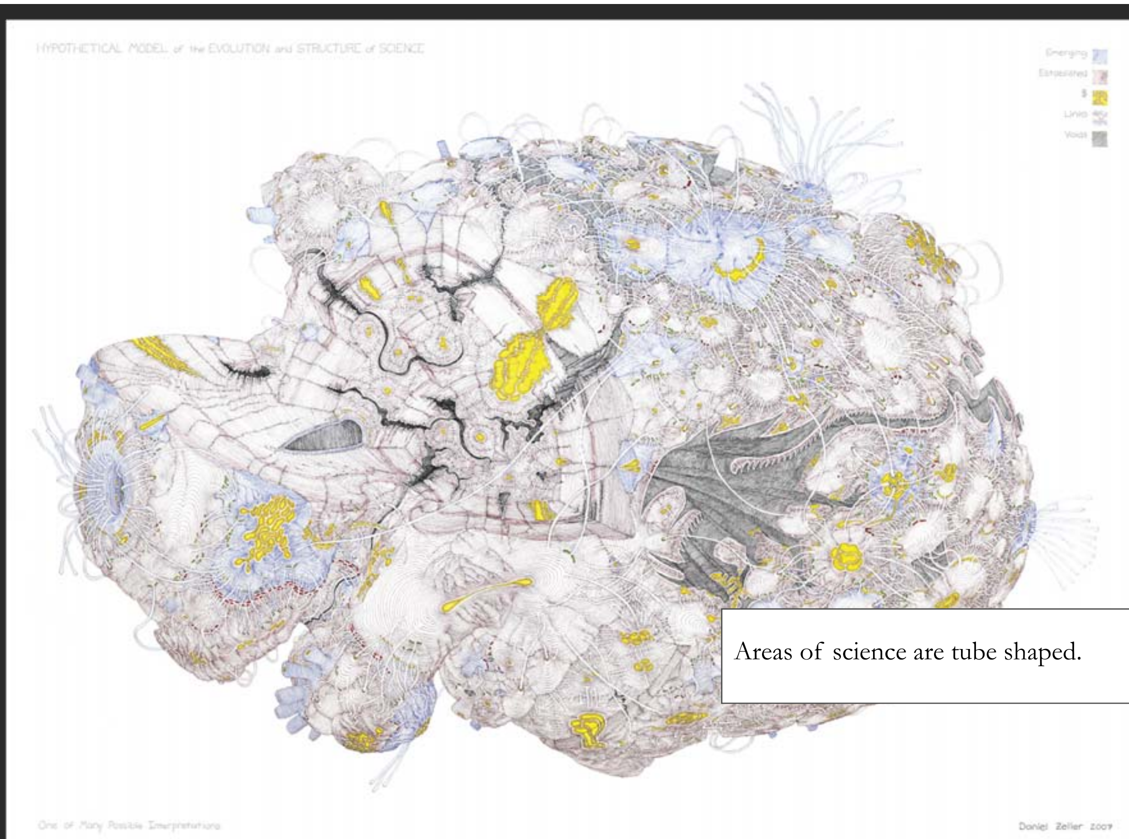
Densely knit communities.
The importance of weak links.

Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007



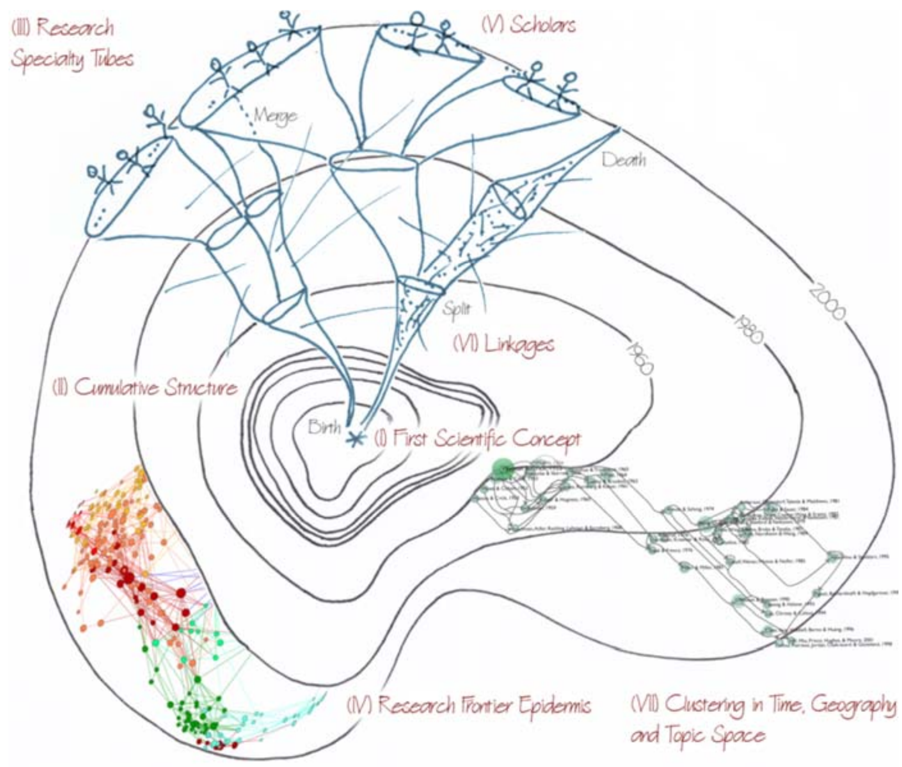
Maps of Science: Forecasting Large Trends in Science - Richard Klavans, Kevin Boyack - 2007

In Terms of Geography - Andre Skupin - 2005

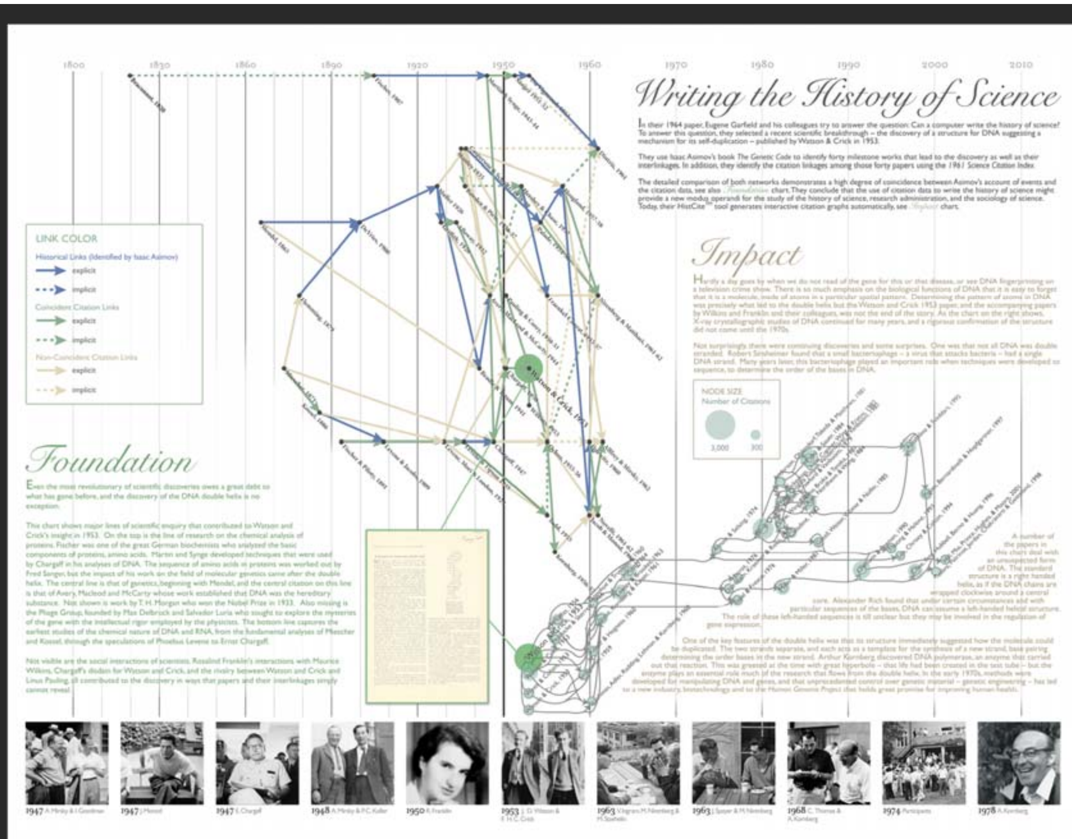


Areas of science are tube shaped.

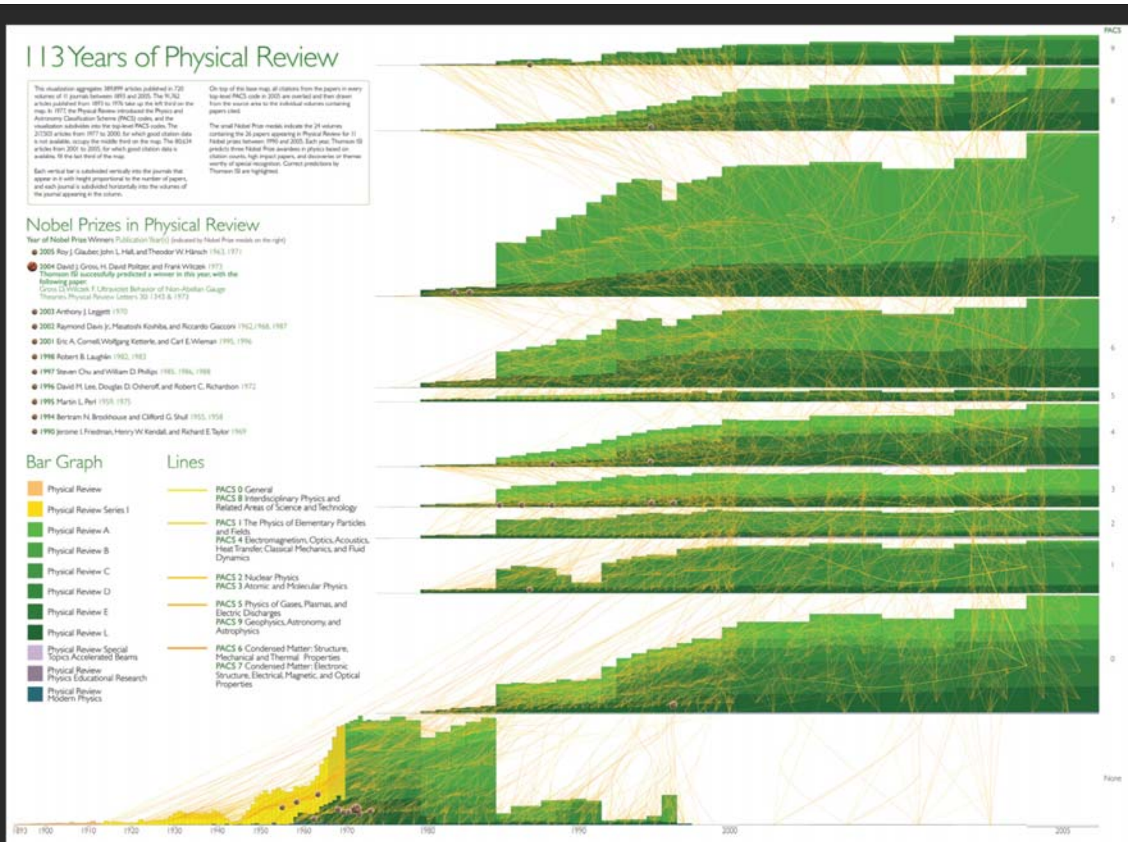
Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007



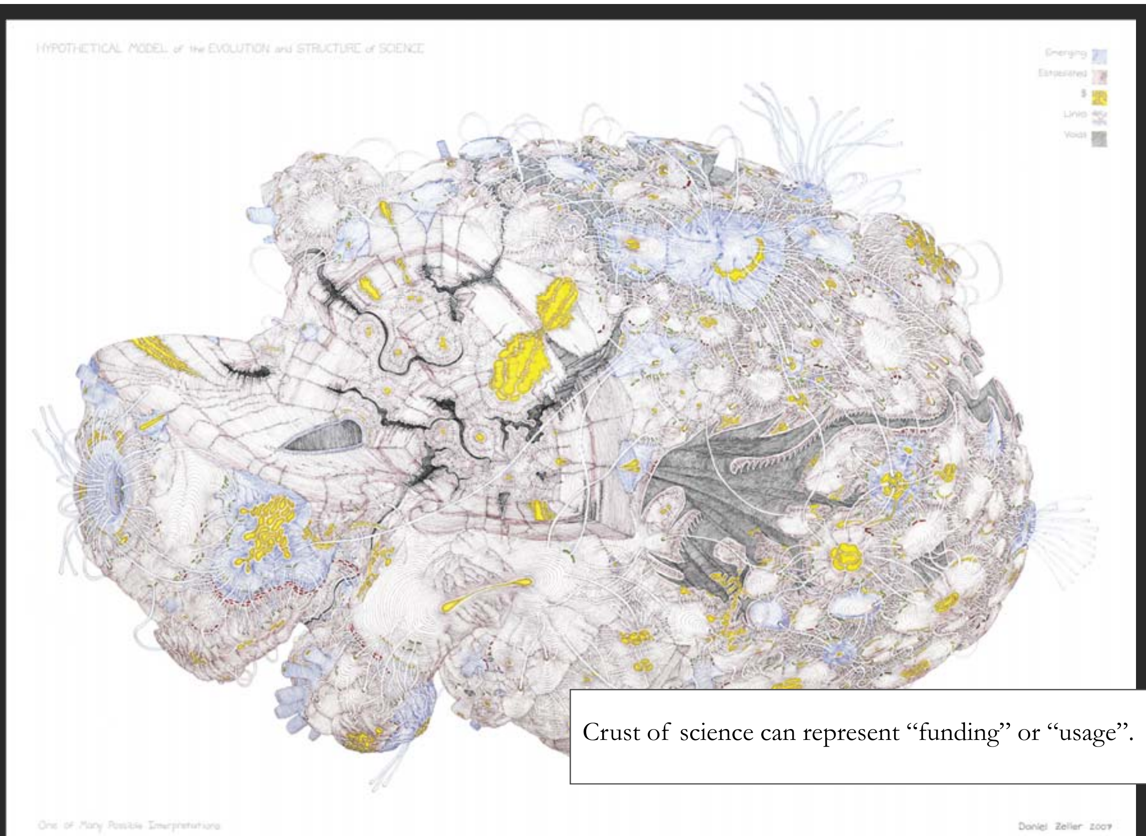
Atlas of Science - Katy Borner - 2010



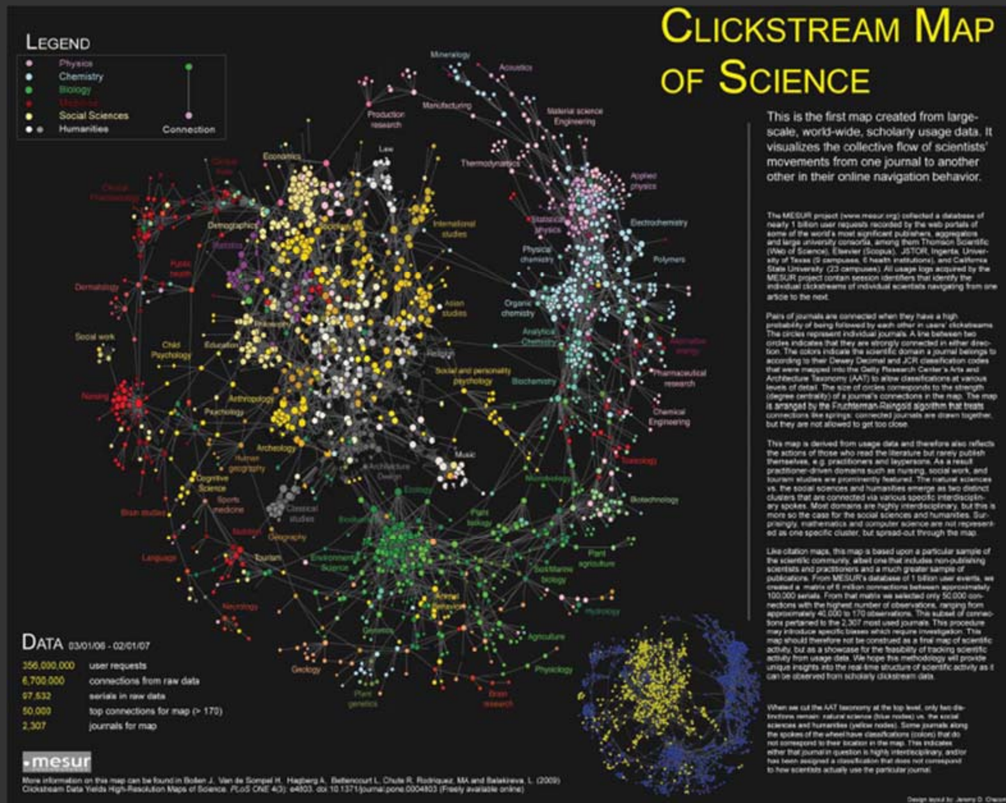
HistCite™ Visualization of DNA Development - Eugene Garfield, Elisha Hardy, Katy Borner, Ludmila Pollock, Jan Witkowski- 2006



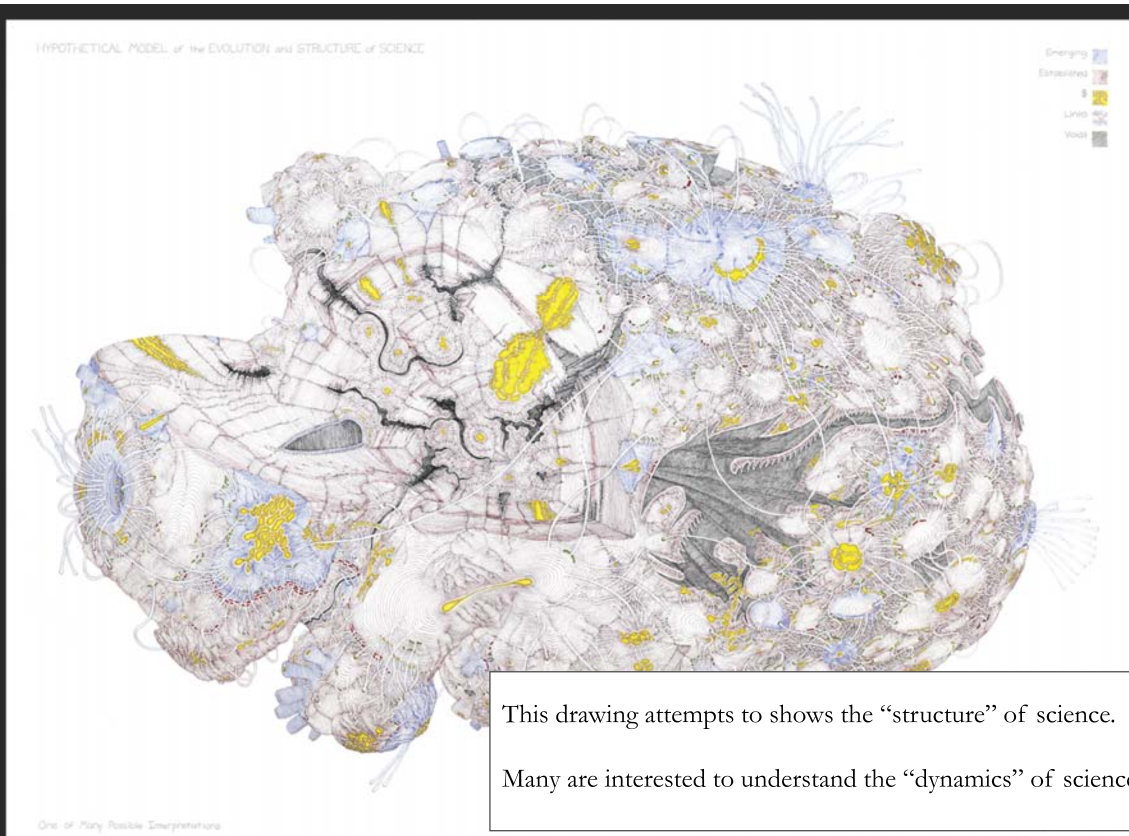
114 Years of Physical Review - Bruce W. Herr II, Russell Duhon, Katy Borner, Elisha Hardy, Shashikant Penumarthy - 2007



Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007



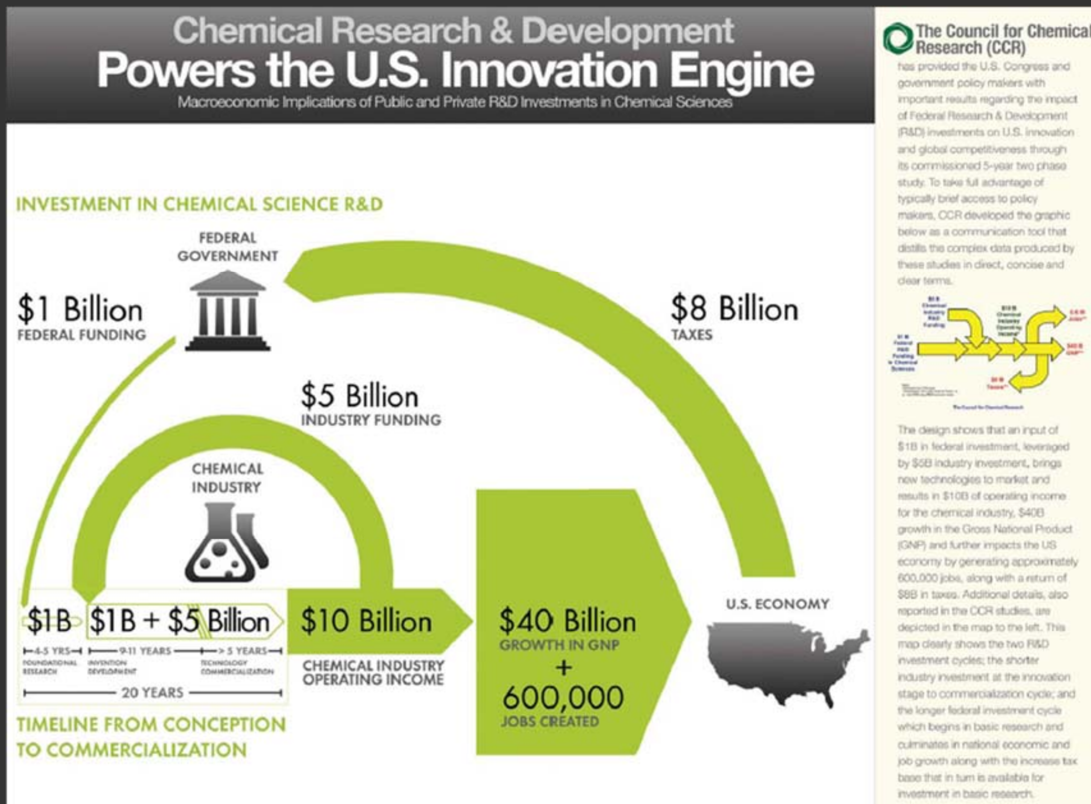
A Clickstream Map of Science – Bollen, Johan, Herbert Van de Sompel, Aric Hagberg, Luis M.A. Bettencourt, Ryan Chute, Marko A. Rodriquez, Lyudmila Balakireva - 2008



This drawing attempts to shows the "structure" of science.

Many are interested to understand the "dynamics" of science.

Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007

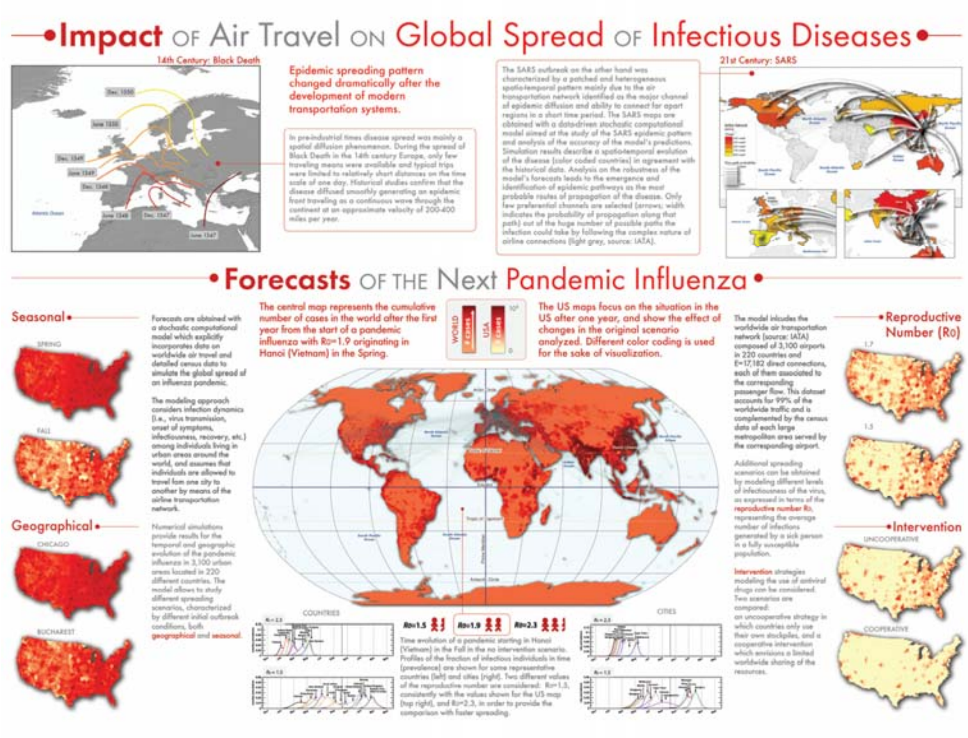


Council for Chemical Research - Chemical R&D Powers the U.S. Innovation Engine. Washington, DC. Courtesy of the Council for Chemical Research - 2009

Four Parts:

1. Conceptualizing Science
2. **Model Inspirations from Other Sciences**
3. Models of Science
4. Tools to Model and Map Science

Modeling of Science
Learning from Epidemiology



Impact of Air Travel on Global Spread of Infectious Diseases - Vittoria Colizza, Alessandro Vespignani - 2007

Modeling Science
Learning from Economics

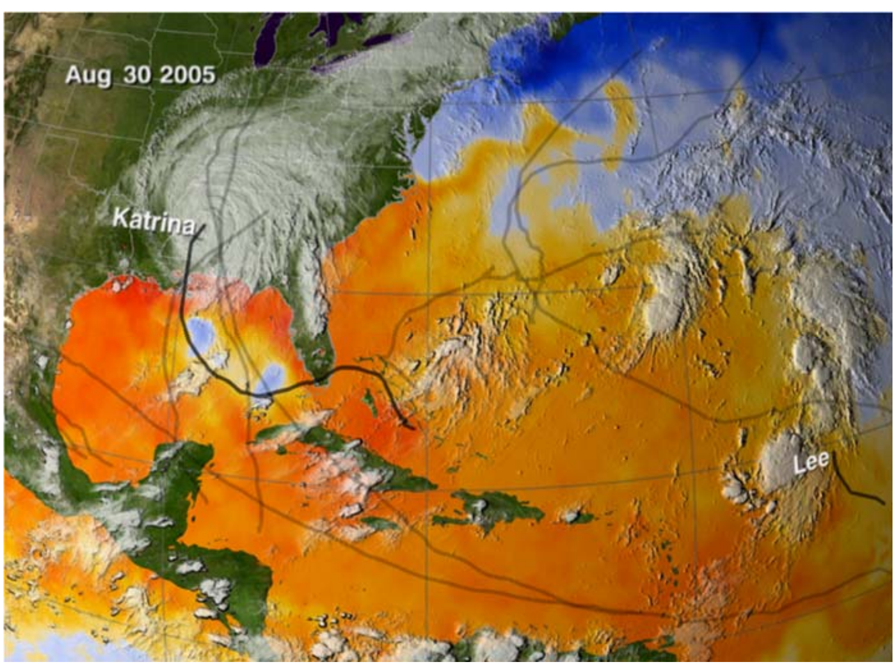
Self amplifying downward spiral | 'systemic' meltdown with intertwined breakdowns | 'war room' analyses | market wind tunnel |power market test bed |
Regulators feel duty–bound to adhere to generally accepted and well-vetted techniques

"… while any new technical device or medical drug has extensive testing for efficiency, reliability and safety before it ever hits the market, we still implement new economic measures without any prior testing." Dirk Helbing

Modeling Science
Learning from Economics

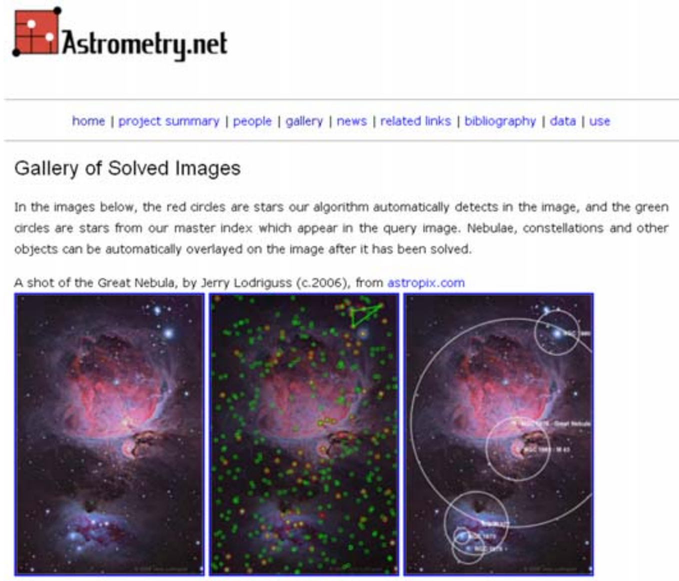
Logicland Participative Global Simulation - Michael Ashauer, Maia Gusberti, Nik Thoenen - 2002

Mapping Science
Learning from Meteorology



Named Storms, available online at http://svs.gsfc.nasa.gov/vis/a000000/a003200/a003279

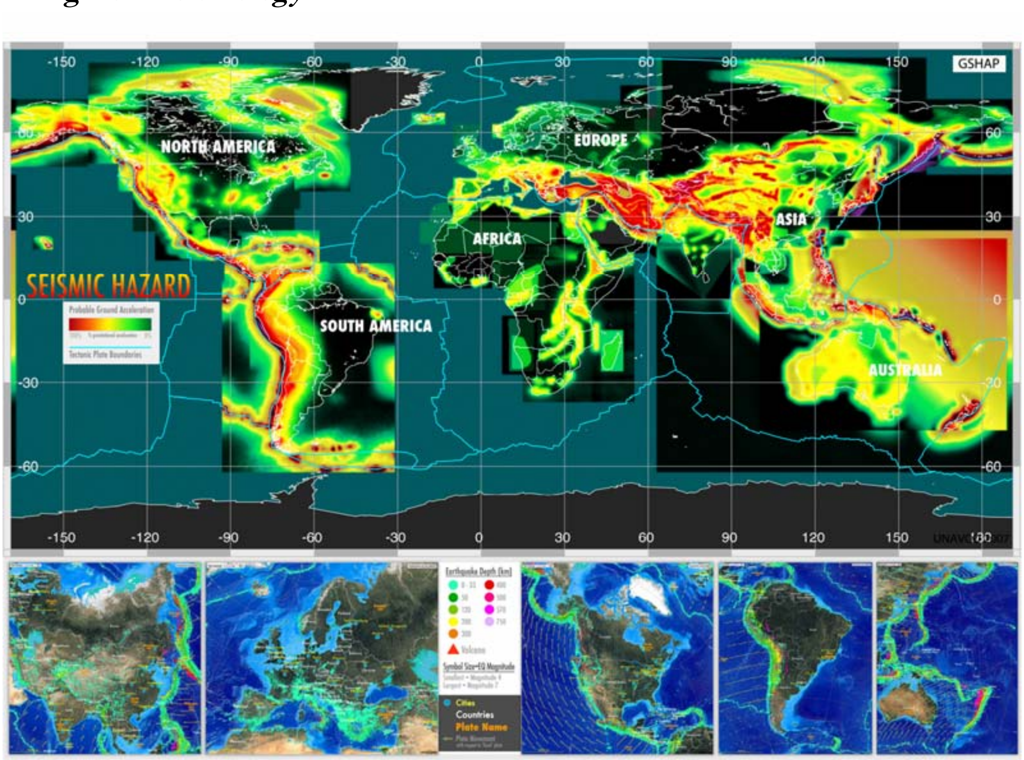
Patch-working Models/Studies/Maps of Science
Learning from Astronomy



http://www.astrometry.net/gallery.html

http://cosmo.nyu.edu/hogg/research/2006/09/28/astrometry_google.pdf

Patch-working Models/Studies/Maps of Science
Learning from Seismology



Tectonic Movements and Earthquake Hazard Predictions - Martin W. Hamburger, Lou Estey, Chuck Meertens, Elisha Hardy - 2005

Four Parts:

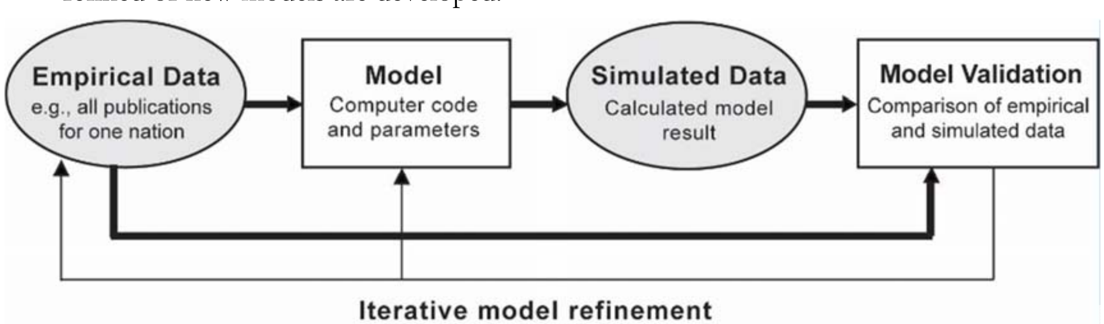
1. Conceptualizing Science
2. Model Inspirations from Other Sciences
3. **Models of Science**
4. Tools to Model and Map Science

An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of process models

Katy Börner, Kevin W. Boyack, Staša Milojević, Steven Morris. (2011) In *Scharnhorst, Andrea, Börner, van den Besselaar (Eds) Models of Science Dynamics. Springer Verlag.*

Modeling Process

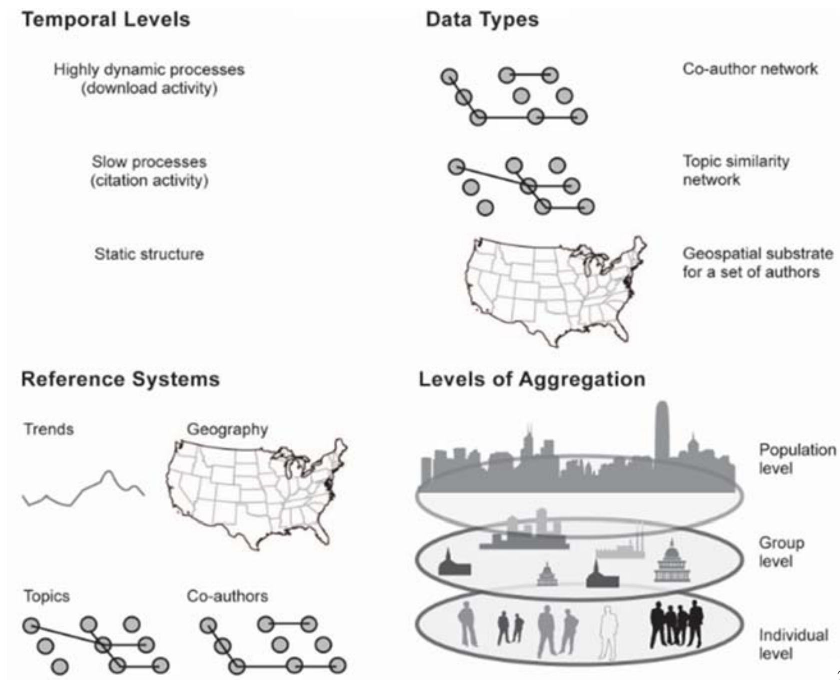
1. Formulation of a scientific hypothesis about the identification of a specific structure or dynamics. Often, this hypothesis is based on analysis of patterns found in empirical data.
2. Algorithm design and implementation using either tools (e.g., NetLogo, RePast) or custom codes that attempt to mathematically describe the structure or dynamics of interest.
3. Simulated data are calculated by running the algorithm and validated by comparison with empirical data.
4. Resulting insights frequently inspire new scientific hypotheses, and the model is iteratively refined or new models are developed.

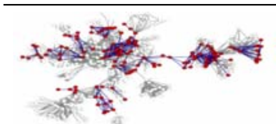


An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of process models
Katy Börner, Kevin W. Boyack, Staša Milojević, Steven Morris. (2011) In *Scharnhorst, Andrea, Börner, van den Besselaar (Eds) Models of Science Dynamics. Springer Verlag.*

Multi-level and multi-perspective models

It is often desirable to model a system at multiple levels using different vantage points.





Descriptive Models vs. Process Models

Descriptive Models

Aim to describe the major features of a (typically static) data set, e.g., statistical patterns of article citation counts, networks of citations, individual differences in citation practice, the composition of knowledge domains, and the identification of research fronts as indicated by new but highly cited papers.

Bibliometrics, Scientometrics, or KDVis

Process Models

Aim to simulate, statistically describe, or formally reproduce the statistical and dynamic characteristics of interest. Of particular interest are models that *"conform to the measured data not only on the level where the discovery was originally made but also at the level where the more elementary mechanisms are observable and verifiable"* (Willinger, Govindan, Jamin, Paxson, & Shenker, 2002), p.2575.

Statistical Physics and Sociology

Descriptive Models

Examples:

- Detect advances of scientific knowledge via "longitudinal mapping" (Garfield, 1994).
- Synthesis of specialty narratives from co-citation clusters (Small, 1986).
- Identify cross-disciplinary fertilization via "passages through science" (Small, 1999, 2000).
- Understand scholarly information foraging (Sandstrom, 2001).
- Knowledge discovery in un-connected terms (Swanson & Smalheiser, 1997).
- Determine areas of expertise for specific researcher, research group via "invisible colleges" (note that researchers self definition might differ from how field defines him/her) (Crane, 1972).
- Identify profiles of authors, also called CAMEOS, to be used to for document retrieval or to map an author's subject matter and studying his/her publishing career, or to map the social and intellectual networks evident in citations to and from authors and in co-authorships (White, 2001).

- Identification of scientific frontiers http://www.science-frontiers.com/.
- *ISI's Essential Science Indicators* http://essentialscience.com/
- Import-export studies (Stigler, 1994).
- Evaluation of 'big science' facilities using 'converging partial indicators' (Martin, 1996; Martin & Irvine, 1983).
- Input (levels of funding, expertise of scientists, facilities used) - output (publications, patents, Nobel prices, improved health, reduced environment insults, etc. - influenced by political, economic, financial, and legal factors studies (Kostroff & DelRio, 2001).
- Determine influence of funding on research output (Boyack & Borner, 2002).

- How to write highly influential paper (van Dalen & Henkens, 2001).

Process Models

Can be used to predict the effects of

- Large collaborations vs. single author research on information diffusion.
- Different publishing mechanisms, e.g., E-journals vs. books on co-authorship, speed of publication, etc.
- Supporting interdisciplinary collaborations (shallow science? or decrease in duplication?).
- Many small vs. one large grant on # publications, Ph.D. students, etc.
- Resource distribution on research output.
- …

In general, process model provide a means to analyze the structure and dynamics of science -- to study science using the scientific methods of science as suggested by Derek J. deSolla Price about 40 years ago.

We now do have the data, code and compute power to do this!

Process Models

In *Sociology*, several mathematical models of network evolution have been developed (Banks & Carley, 95). Most assume a fixed number of edges.

Snijders' Simulation Investigation for Empirical Network Analysis (SIENA) (http://stat.gamma.rug.nl/snijders/siena.html) is a probabilistic model for the evolution of social networks. It assumes a directed graph with a fixed set of actors.

Recent work in *Statistical Physics* aims to design models and analytical tools to analyze the statistical mechanics of topology and dynamics of real world networks. Of particular interest is the identification of elementary mechanisms that lead to the emergence of *small-world* (Albert & Barabási, 2002; Watts, 1999) and *scale free network structures* (Barabási, Albert, & Jeong, 2000). The models assume nodes of one type (e.g., web page, paper, author).

Examples:

- **Watts-Strogatz Model** for Small World Networks
- **Albert-Barabasi Model** for Scale Free Networks

The Watts-Strogatz Model for Small World Networks

First model that generates graphs with small average path length and high clustering coefficients.

- Starting configuration is a regular lattice.
- Each edge is examined and is redirected with a probability p to another target node (chosen randomly).

Regular network (left) drastically changes from a set of tiny isolated clusters of nodes to a giant cluster joined by almost everybody.

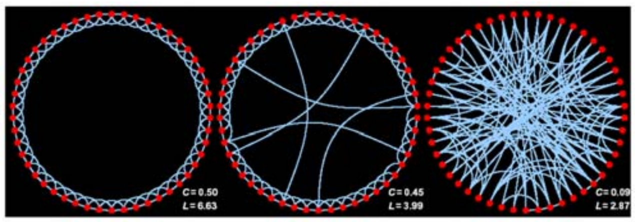


Figure 1. Watts-Strogatz model interpolates between a regular lattice *(left)* and a random graph *(right)*. Randomly rewiring just a few edges *(center)* reduces the average distance between nodes, *L*, but has little effect on the clustering coefficient, *C*. The result is a "small-world" graph.

(Source: D.J. Watts and S. Strogatz. Collective Dynamics of `Small-World' Networks. Nature, Vol. 393(6):pp. 440-442, June 1998.)

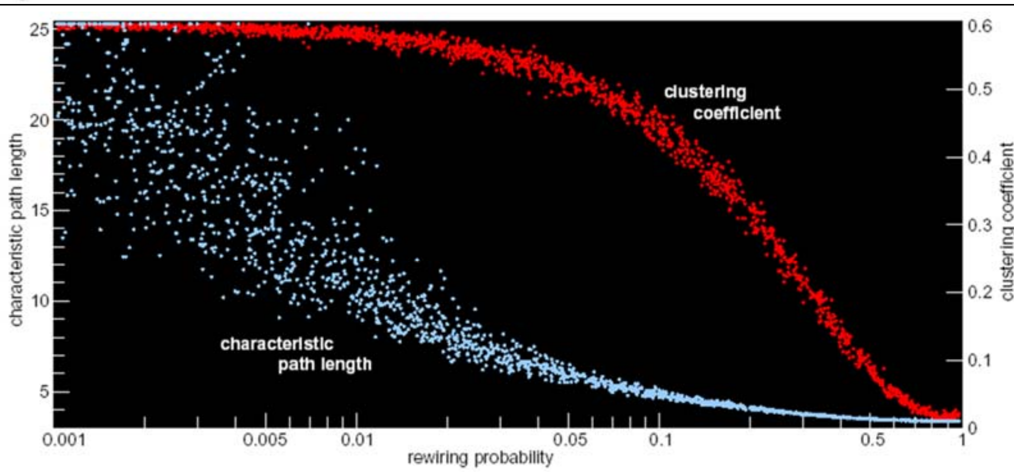
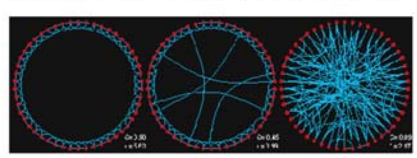


Figure 2. As the probability of rewiring increases in the Watts-Strogatz model, the characteristic path length falls off long before the clustering coefficient drops. Results are from 2,000 random graphs, each with 300 vertices and 900 edges.

C(p) and l(p) as a function of rewiring probability p.



Dorogovtsev, S. N. & Mendes, J. F. F. (2002). Evolution of networks. cond-mat/0106144

Variations of the WS Model

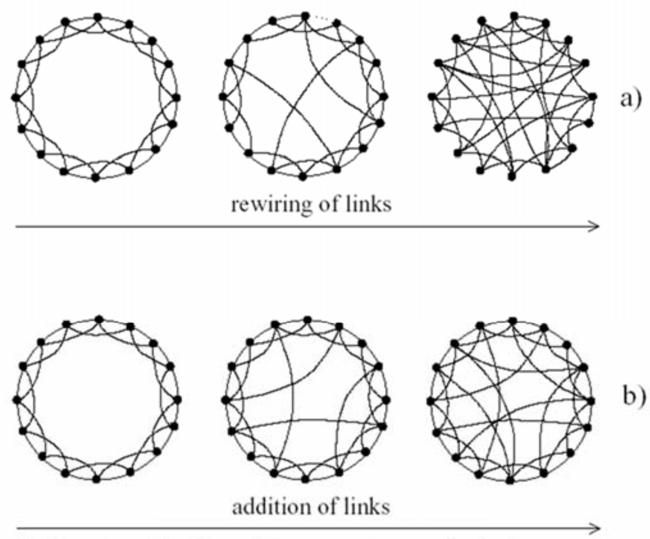


FIG. 10. Small-world networks in which the crossover from a regular lattice to a random network is realized. (a) The original Watts-Strogatz model with the rewiring of links [11]. (b) The network with the addition of shortcuts [133,134].

The Barabasi–Albert (BA) Model for Scale Free Networks

Many large networks are scale free, their degree distribution follows a power law for large k.

Random graph theory and the small world model cannot reproduce this feature.

(1) Growth: Starting with a small number (m_0) of nodes, at every time step, we add a new node with m(<= m_0) edges that link the new node to m nodes already present in the system.

(2) Preferential attachment: The probability p that a new node will be connected to node i depends on the degree k_i of node i, such that

$$p(k_i) = \frac{k_i}{\sum_j k_j}.$$

After t time-steps the network has $N = t+m_0$ nodes and mt edges.

This network evolves into a stationary scale-free state with the probability that a node has k edges following a power law with an exponent $\gamma_{BA} = 3$.

(Source: A.-L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509512.)

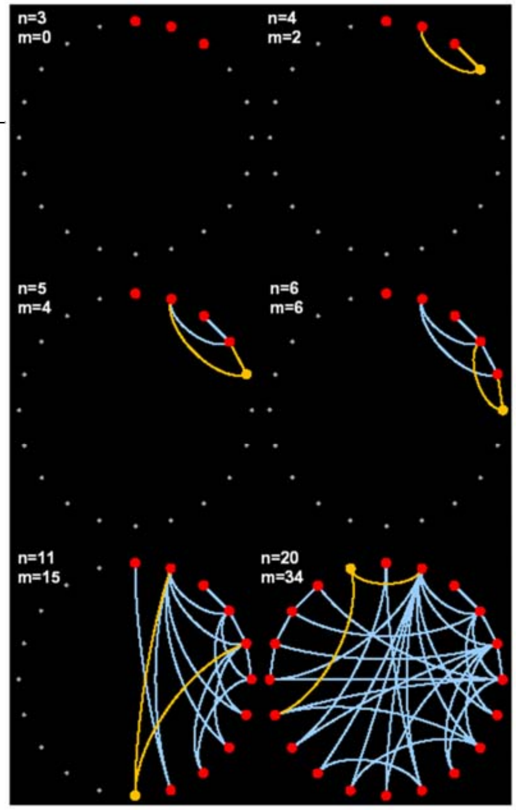


Figure 3. Model of Barabási, Albert and Jeong grows a graph by adding both vertices and edges. In the example shown here, each stage adds one new vertex and two new edges *(yellow)*.

Reka Albert, Albert-Laszlo Barabasi (2002). [Statistical mechanics of complex networks.](#) cond-mat/0106096: Evolving Networks

Alternative models that simulate preferential attachment

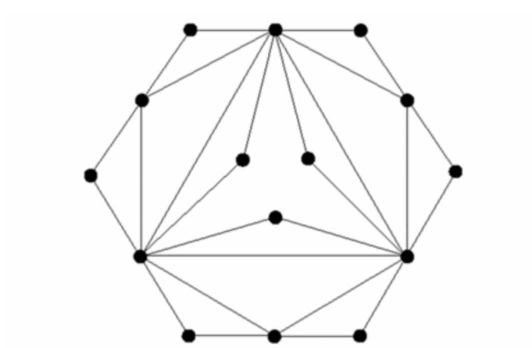


FIG. 23. A simple deterministic growing graph. At time $t = 0$, the graph is a triangle. At each time step every edge of the graph generates a new vertex which connects to both ends of the edge.

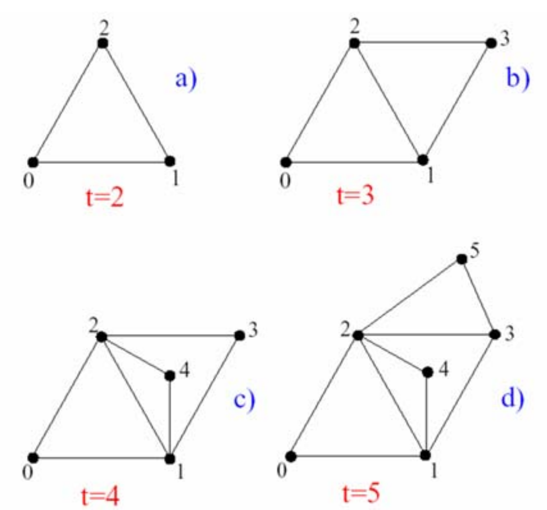


FIG. 21. Illustration of a simple model of a scale-free growing network [166]. In the initial configuration, $t = 2$, three vertices are present, $s = 0, 1, 2$ (a). At each increment of time, a new vertex with two edges is added. These edges are attached to the ends of a randomly chosen edge of the network.

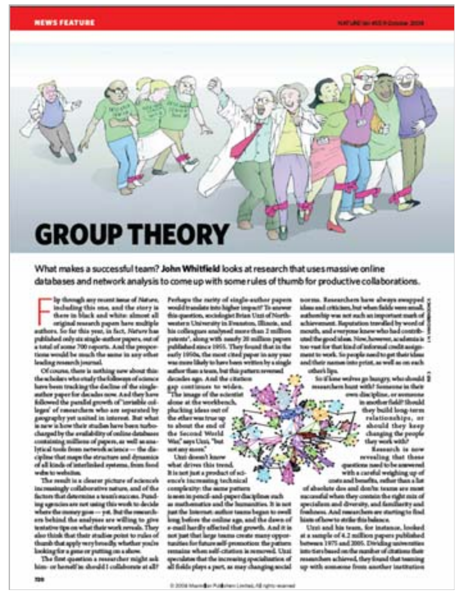
Information Diffusion Among Major U.S. Research Institutions

Börner, Katy, Penumarthy, Shashikant, Meiss, Mark & Ke, Weimao. (2006). Mapping the Diffusion of Information among Major U.S. Research Institutions. Scientometrics. Vol. 68(3), 415 - 426.



Questions:

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

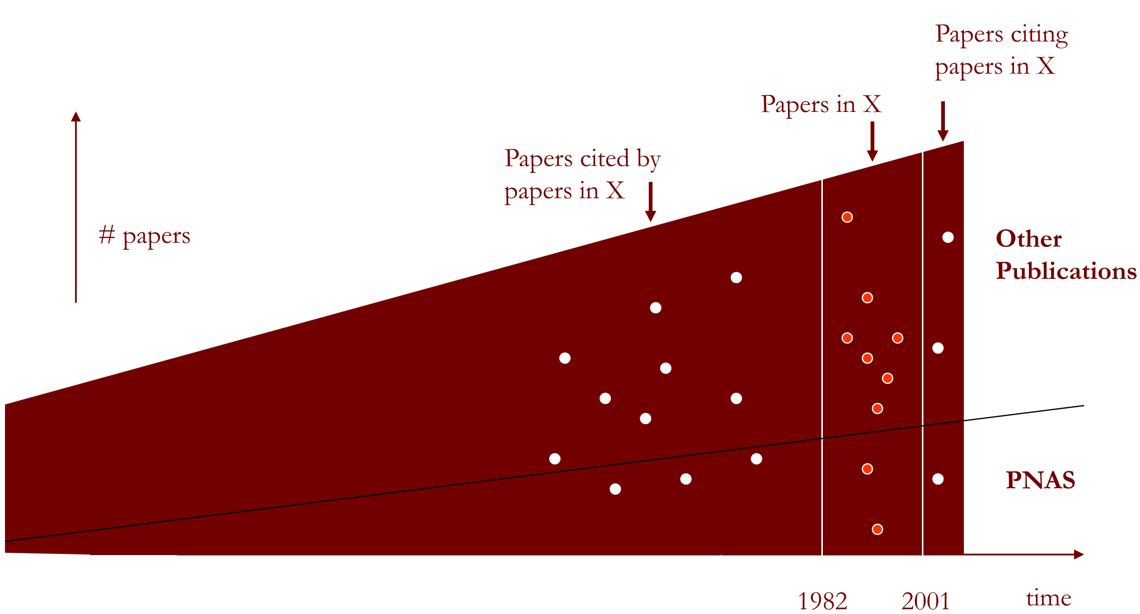


Contributions:

- Answer to Q1 is YES.
- Answer to Q2 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.

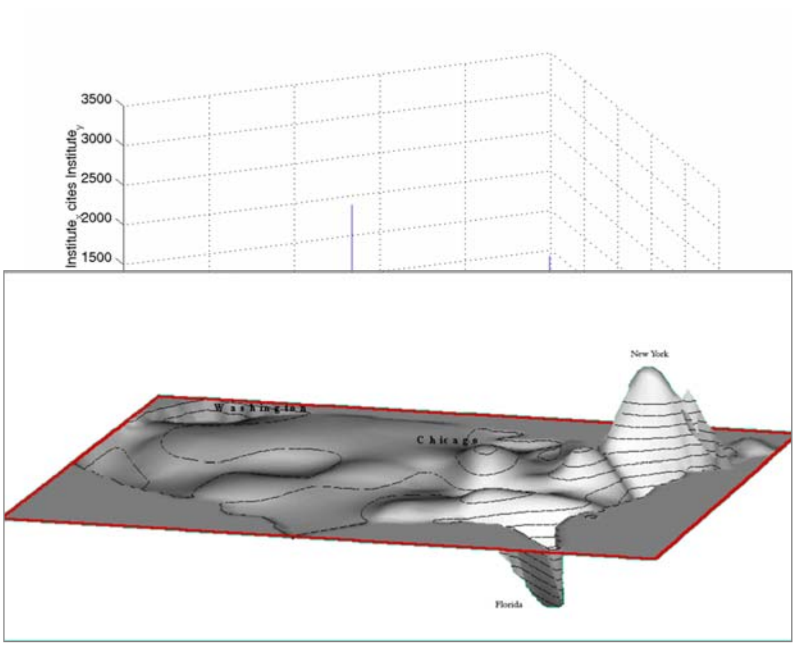
20-Year PNAS Dataset (1982-2001)

Coverage in terms of time span, total number of papers, and complete author's work



Citation Matrix

Unsymmetrical direct citation linkage patterns among the top 500 institutions. High peak values in the diagonal reflect the high amount of self-citations for all institutions. Medium peak horizontal and vertical lines denote references from and citations to papers written at Harvard University.



Information Sources (Export) and Sinks (Import)

Calculate ratio of the number of citations received by an institution divided by the sum of received citations and references made, multiplied by 100.

131 have a value between 0-40% acting mostly as information producers = information sources.

71 have a value between 60-100% and act mostly as information consumers – they reference a large number of papers but the number of citations they receive is comparably low = information sinks.

(Tobler, 1995)

Geographic Location of Received Citations

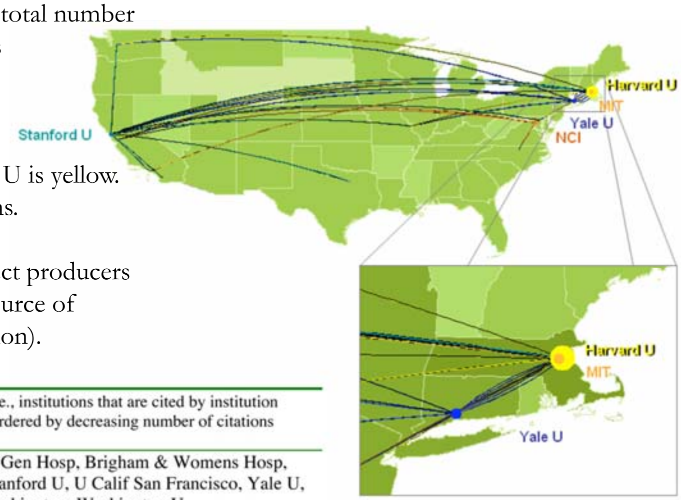
Unsymmetrical direct citation linkage patterns among the top 500 institutions. High peak values in the diagonal reflect the high amount of self-citations for all institutions. Medium peak horizontal and vertical lines denote references from and citations to papers written at Harvard University.

Information Flow Among the Top-5 Consumers and Their Top-10 Producers

U.S. states are color coded based on the total number of citations received by their institutions (excluding self citations).

Dots indicate the five producers.
Each has a different color, e.g., Harvard U is yellow.
Dot area size depicts number of citations.

Lines represent citations that interconnect producers and consumers shaded from colored (source of information) to white (sink of information).

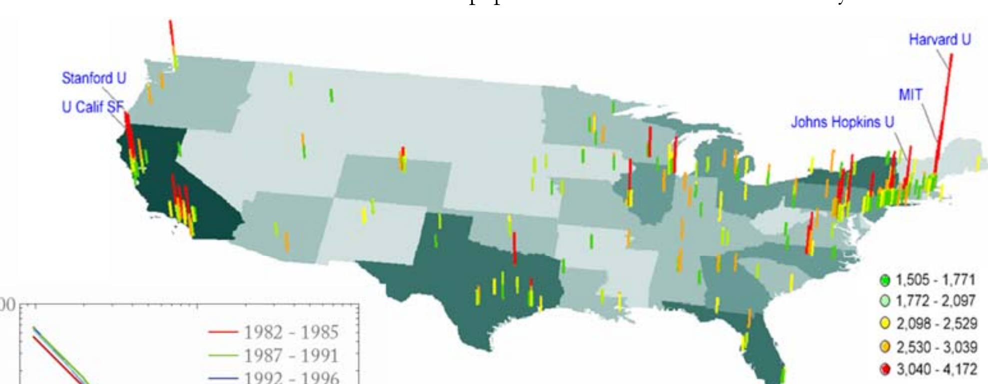


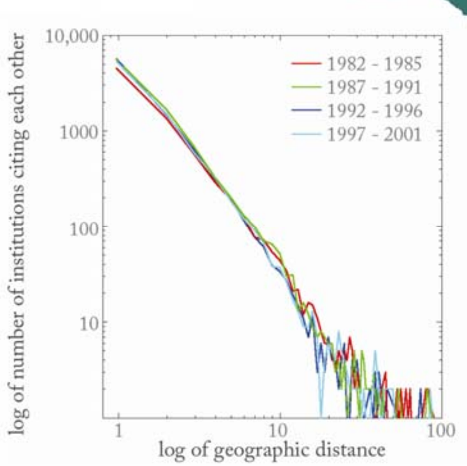
Consumers, i.e., citing institutions	# citations made	Top ten producers, i.e., institutions that are cited by institution listed in first column ordered by decreasing number of citations received.
Harvard U	13,552	MIT, Massachusetts Gen Hosp, Brigham & Womens Hosp, Johns Hopkins U, Stanford U, U Calif San Francisco, Yale U, Rockefeller U, U Washington, Washington U
U Calif SF	4,682	Harvard U, MIT, Stanford U, Johns Hopkins U, U Washington, Washington U, U Calif Berkeley, U Texas, U Calif SD, U Calif LA
MIT	4,655	Harvard U, Whitehead Inst Biomed Res, Johns Hopkins U, Stanford U, U Calif SF, Yale U, Rockefeller U, U Calif LA, Massachusetts Gen Hosp, U Calif Berkeley
NCI (zip: 20814)	4,519	Harvard U, NCI (zip: 20205), NCI (zip: 21701), MIT, Duke U, Johns Hopkins U, NIAID NICHHD, Stanford U, U Calif SF
Yale U	4,464	Harvard U, MIT, Stanford U, Rockefeller U, Johns Hopkins U, Washington U, U Calif SF, U Washington, NCI, Massachusetts Gen Hosp

Paper also shows top-5 producers and their top-10 consumers.

Changes in Citation Behavior Over Time

Unsymmetrical direct citation linkage patterns among the top 500 institutions. High peak values in the diagonal reflect the high amount of self-citations for all institutions. Medium peak horizontal and vertical lines denote references from and citations to papers written at Harvard University.





1982-1986: **1.94** (R=91.5%)
1987-1991: **2.11** (R=93.5%)
1992-1996: **2.01** (R=90.8%)
1997-1901: **2.01** (R=90.7%)

As time progresses and the amount of produced papers increases, space seems to matter more. Authors are more likely to cite papers generated by authors at close-by institutions.

Modeling the Co-Evolving Author-Paper Networks

Börner, Katy, Maru, Jeegar & Goldstone, Robert. (2004). The Simultaneous Evolution of Author and Paper Networks. PNAS. Vol. 101(Suppl. 1), 5266-5273.

The TARL Model (Topics, Aging, and Recursive Linking) incorporates

- A partitioning of authors and papers into topics,
- Aging, i.e., a bias for authors to cite recent papers, and
- A tendency for authors to cite papers cited by papers that they have read resulting in a rich get richer effect.

The model attempts to capture the roles of authors and papers in the production, storage, and dissemination of knowledge.

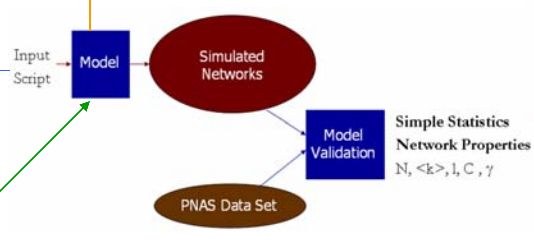
Model Assumptions

- Co-author and paper-citation networks co-evolve.
- Authors come and go.
- Papers are forever.
- Only authors that are 'alive' are able to co-author.
- All existing (but no future) papers can be cited.
- Information diffusion occurs directly via co-authorships and indirectly via the consumption of other authors' papers.

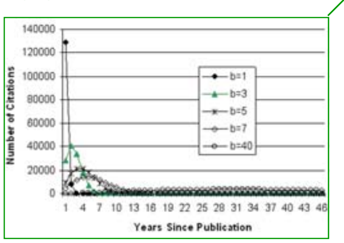
- Preferential attachment is modeled as an *emergent property* of the elementary, local networking activity of authors reading and citing papers, but also the references listed in papers.


```
------------------------------------------
Model Parameters (0=without, 1=with)
------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
------------------------------------------
Model Initialization Values
------------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1  # Levels References are Considered
```

```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select #_read_ papers from existing papers;
                get all references of read_ papers up to #_reference_path_length;
                for each new_paper_reference do {
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```



Aging function



Model Validation

The properties of the networks generated by this model are validated against a 20-year data set (1982-2001) of documents of type article published in the Proceedings of the National Academy of Science (PNAS) – about 106,000 unique authors, 472,000 co-author links, 45,120 papers cited within the set, and 114,000 citation references within the set.

Table 3 Statistics for SIM data

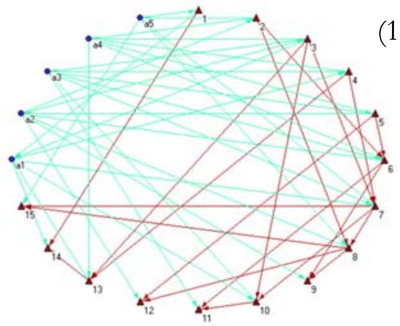
Year	#p	#a	#r	#c	a/#ca
1981	1624	3953	0	756	8.21
1982	1040	5200	31200	112161	4
1983	1118	5590	33540	21397	4
1984	1197	5985	35910	10224	4
1985	1275	6375	38250	6184	4
1986	1353	6765	40590	4687	4
1987	1432	7160	42960	3573	4
1988	1510	7550	45300	2816	4
1989	1589	7945	47670	2219	4
1990	1667	8335	50010	1853	4
1991	1745	8725	52350	1634	4
1992	1824	9120	54720	1431	4
1993	1902	9510	57060	1167	4
1994	1981	9905	59430	1040	4
1995	2059	10295	61770	767	4
1996	2137	10685	64110	632	4
1997	2216	11080	66480	522	4
1998	2294	11470	68820	400	4
1999	2373	11865	71190	265	4
2000	2451	12255	73530	125	4
2001	2529	12645	75870	0	4
Total	**37316**		**1070760**	**173853**	

Table 2. PNAS Statistics

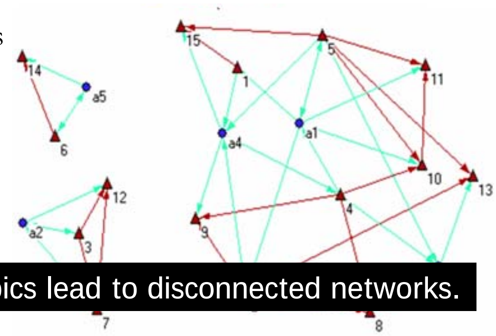
Year	#p	#a	#r	#c	a/#ca
1982	1669	5201	46665	156690	3.92
1983	1611	5142	46685	161437	3.98
1984	1695	5583	49834	174161	4.22
1985	1846	6325	55662	191750	4.38
1986	2042	7209	64379	218229	4.76
1987	1924	7061	59110	207729	4.88
1988	2035	7471	63116	215227	4.8
1989	2088	7959	65883	215437	5.01
1990	2066	8031	66019	207138	5.15
1991	2382	9559	77740	223102	5.25
1992	2500	9812	80949	211238	5.29
1993	2413	9770	79848	193867	5.55
1994	2600	10656	86176	187353	5.56
1995	2476	10429	82021	151249	5.66
1996	2765	11803	99061	148622	5.96
1997	2618	11255	96788	122908	6.12
1998	2711	12328	100973	107764	6.48
1999	2603	12182	97018	76080	6.69
2000	2501	12201	94181	44131	7.6
2001	2575	13038	97450	16357	8.4
Total	**45120**		**1509558**	**3230469**	

The TARL Model: The Effect of Parameters

(0000)

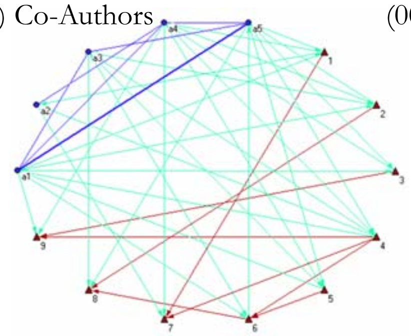


(1000) Topics

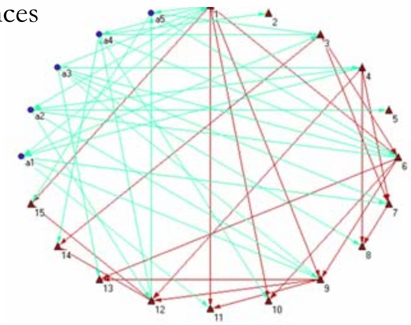


Topics lead to disconnected networks.

(0100) Co-Authors



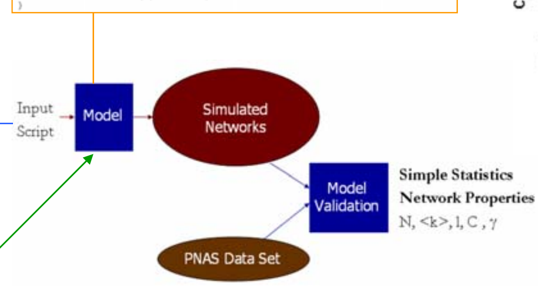
(0010) References



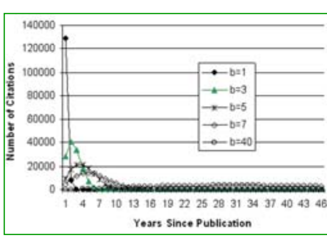
Co-authoring leads to fewer papers.


```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select #_read_ papers from existing papers;
                get all references of read_ papers up to #_reference_path_length;
                for each new_paper_reference do {
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```

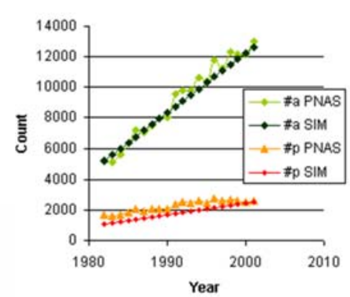
```
---------------------------------------------
Model Parameters (0=without, 1=with)
---------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
---------------------------------------------
Model Initialization Values
---------------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1   # Levels References are Considered
```



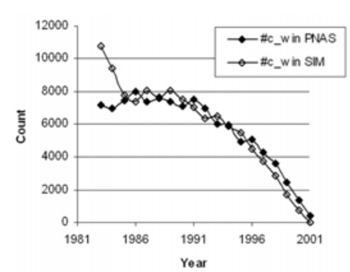
Aging function



Counts for Papers and Authors



Counts for Citations



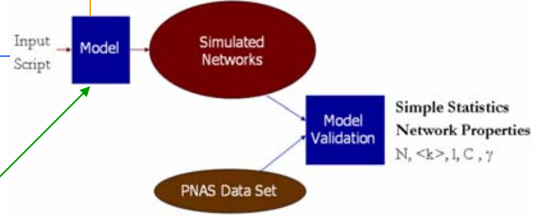
```
------------------------------------------
Model Parameters (0=without, 1=with)
------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
------------------------------------------
Model Initialization Values
------------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1  # Levels References are Considered
```

```
// Initialization
generate #_papers papers and assign a random topi
generate #_authors authors and assign a random to
randomly assign #_co-authors+1 authors to papers o
// Simulation
for each year do (
    add #_new_authors new authors, deactivate aut
    for each topic do (
        randomly partition set of authors into autho
        for each author_group do (
            for each new_paper to be produced,
                generate new_paper;
                randomly select #_read_ paper
                get all references of read_ pap
                for each new_paper_reference
                    select a time_slice from (
                    randomly select a paper p
                    add the new_paper_refere
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```

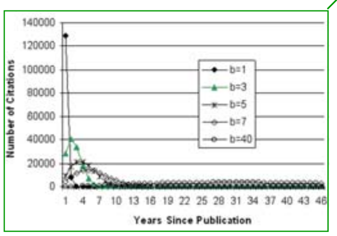
Table 2. Properties of co-author & paper citation networks comprising number of nodes n, average node degree <k>, path length l, cluster coefficient C, and power law exponent γ. Source references are given in the left column.

Network	n	<k>	l	C	γ	Reference
Co-authorship networks						
LANL	52,909	9.7	5.9	0.43	--	Newman, (2001a; 2001b; 2001c)
MEDLINE	1,520,251	18.1	4.6	0.066	--	
SPIRES	56,627	1.73	4.0	0.726	1.2	
NCSTRL	11,994	3.59	9.7	0.496	--	
Math.	70,975	3.9	9.5	0.59	2.5	Barabasi et al., (2002)
Neurosci.	209,293	11.5	6	0.76	2.1	
PNAS	105,915	8.97	5.89	0.399	2.54	
Paper-citation networks						
ISI	783,339	8.57	--	--	3	Redner, (1998)
PhysRev	24,296	14.5	--	--	3	
PNAS	45,120	3.53	--	0.081	2.29	
SIM	37,114	2.13	--	0.074	2.05	

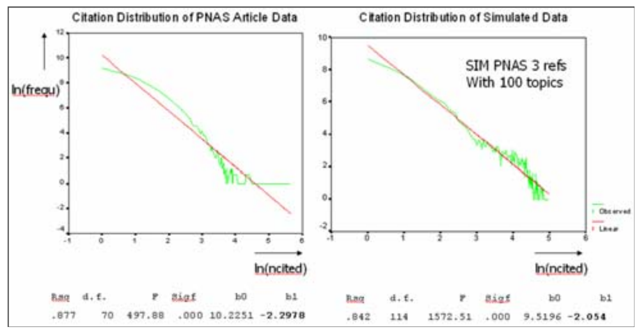
Co-Author and Paper-Citation Network Properties



Aging function

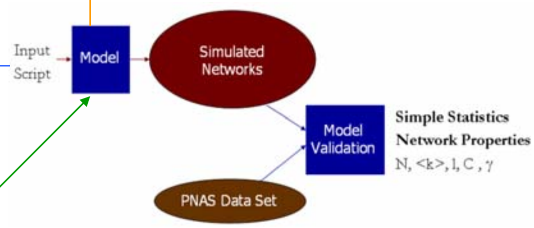


Power Law Distributions

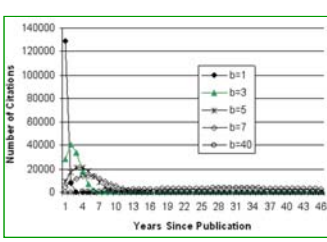



```
------------------------------------------
Model Parameters (0=without, 1=with)
------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
------------------------------------------
Model Initialization Values
------------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1  # Levels References are Considered
```


```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do (
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do (
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do (
            for each new_paper to be produced, do (
                generate new_paper;
                randomly select #_read_ papers from existing papers;
                get all references of read_ papers up to #_reference_path_length;
                for each new_paper_reference do (
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```



Aging function



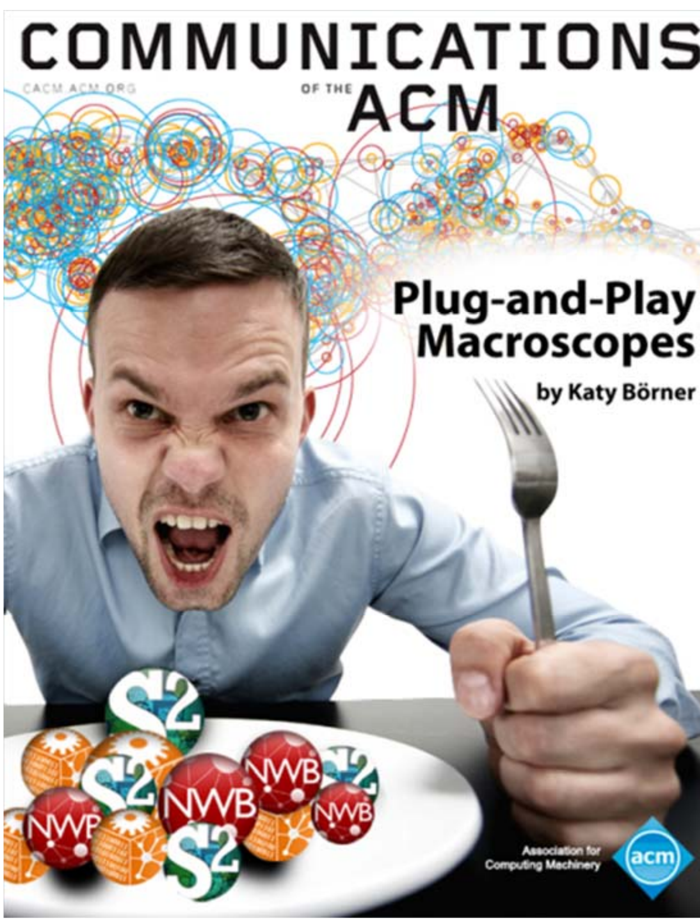
Topics: The number of topics is linearly correlated with the clustering coefficient of the resulting network: C= 0.000073 * #topics. Increasing the number of topics increases the power law exponent as authors are now restricted to cite papers in their own topics area.

Aging: With increasing b, and hence increasing the number of older papers cited as references, the clustering coefficient decreases. Papers are not only clustered by topic, but also in time, and as a community becomes increasingly nearsighted in terms of their citation practices, the degree of temporal clustering increases.

References/Recursive Linking: The length of the chain of paper citation links that is followed to select references for a new paper also influences the clustering coefficient. Temporal clustering is ameliorated by the practice of citing (and hopefully reading!) the papers that were the earlier inspirations for read papers.

Four Parts:

1. Conceptualizing Science
2. Model Inspirations from Other Sciences
3. Models of Science
4. Tools to Model and Map Science



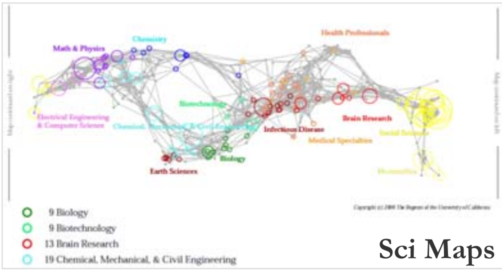
Börner, Katy. (March 2011). Plug-and-Play Macroscopes. *Communications of the ACM, 54(3), 60-69.*

Video and paper are at http://www.scivee.tv/node/27704

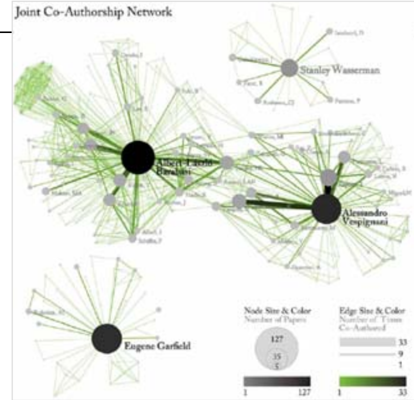
Sci² Tool – "Open Code for S&T Assessment"

http://sci2.cns.iu.edu

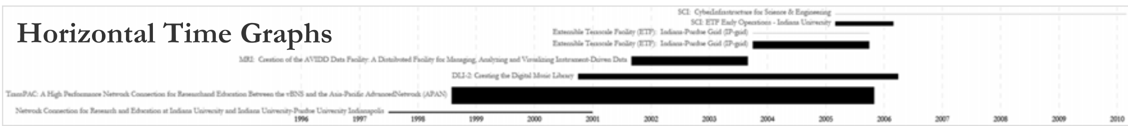
OSGi/CIShell powered tool with NWB plugins and many new scientometrics and visualizations plugins.



Sci Maps



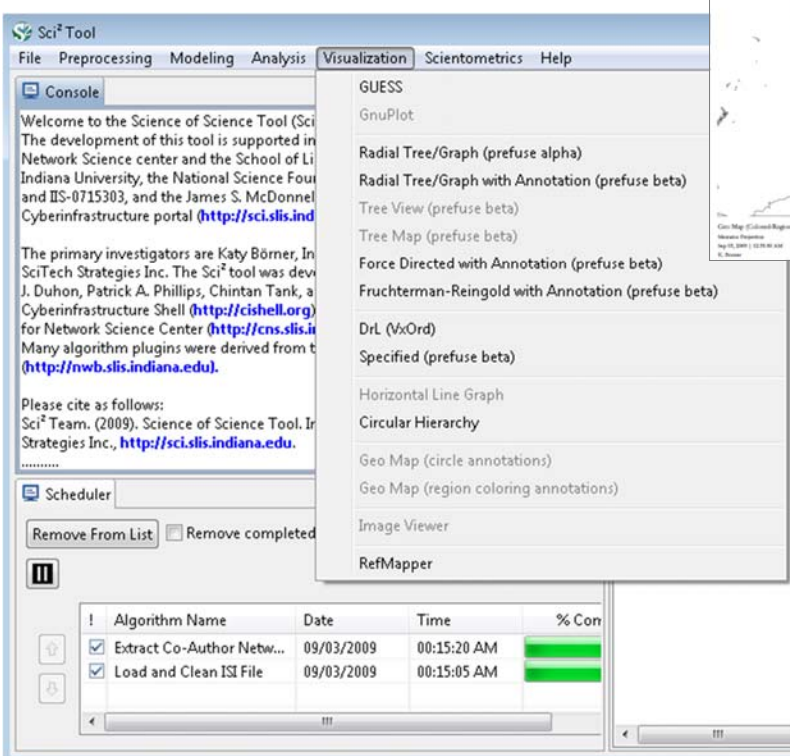
GUESS Network Vis

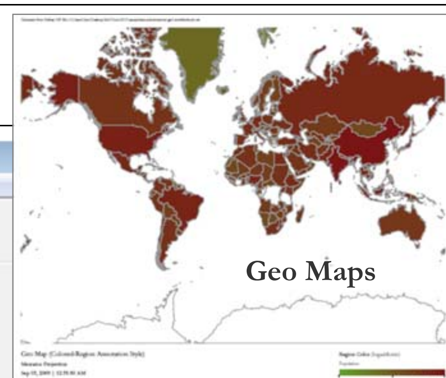


Horizontal Time Graphs

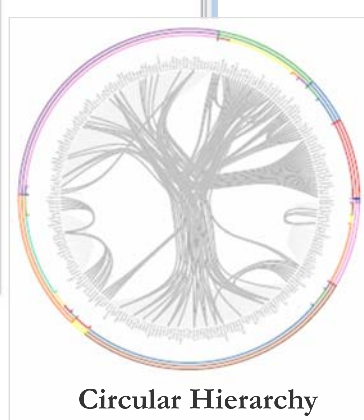
Börner, Katy, Huang, Weixia (Bonnie), Linnemeier, Micah, Duhon, Russell Jackson, Phillips, Patrick, Ma, Nianli, Zoss, Angela, Guo, Hanning & Price, Mark. (2009). Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Scholarly Database and the Network Workbench Tool. Proceedings of ISSI 2009: 12th International Conference on Scientometrics and Informetrics, Rio de Janeiro, Brazil, July 14-17 . Vol. 2, pp. 619-630.

Sci² Tool





Geo Maps



Circular Hierarchy

Type of Analysis vs. Level of Analysis

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

Sci² Tool: Algorithms

Preprocessing
Extract Top N% Records
Extract Top N Records
Normalize Text
Slice Table by Line

Extract Top Nodes
Extract Nodes Above or Below Value
Delete Isolates

Extract top Edges
Extract Edges Above or Below Value
Remove Self Loops
Trim by Degree
MST-Pathfinder Network Scaling
Fast Pathfinder Network Scaling

Snowball Sampling (in nodes)
Node Sampling
Edge Sampling

Symmetrize
Dichotomize
Multipartite Joining

Geocoder

Extract ZIP Code

Modeling
Random Graph
Watts-Strogatz
Small World
Barabási-Albert Scale-Free
TARL

Analysis
Network Analysis Toolkit (NAT)
Unweighted & Undirected
- Node Degree
- Degree Distribution
- ---
- K-Nearest Neighbor (Java)
- Watts-Strogatz Clustering Coefficient
- Watts Strogatz Clustering Coefficient over K
- ---
- Diameter
- Average Shortest Path
- Shortest Path Distribution
- Node Betweenness Centrality
- ---
- Weak Component Clustering
- Global Connected Components
- ---
- Extract K-Core
- Annotate K-Coreness
- ---
- HITS

Weighted & Undirected
- Clustering Coefficient
- Nearest Neighbor Degree
- Strength vs Degree
- Degree & Strength
- Average Weight vs End-point Degree
- Strength Distribution
- Weight Distribution
- Randomize Weights
- ---
- Blondel Community Detection
- ---
- HITS

Unweighted & Directed
- Node Indegree
- Node Outdegree
- Indegree Distribution
- Outdegree Distribution
- ---
- K-Nearest Neighbor
- Single Node in-Out Degree Correlations
- ---
- Dyad Reciprocity
- Arc Reciprocity
- Adjacency Transitivity
- ---
- Weak Component Clustering
- Strong Component Clustering
- ---

Sci² Tool: Algorithms cont.

- --------------------------------
- Extract K-Core
- Annotate K-Coreness
- --------------------------------
- HITS
- PageRank

Weighted & Directed
- HITS
- Weighted PageRank

Textual
Burst Detection

Visualization
GnuPlot
GUESS
Image Viewer

Radial Tree/Graph (prefuse alpha)
Radial Tree/Graph with Annotation (prefuse beta)
Tree View (prefuse beta)
Tree Map (prefuse beta)
Force Directed with Annotation (prefuse beta)
Fruchterman-Reingold with Annotation (prefuse beta)

DrL (VxOrd)
Specified (prefuse beta)

Horizontal Bar Graph
Circular Hierarchy
Geo Map (Circle Annotation Style)
Geo Map (Colored-Region Annotation Style)
Science Map (Circle Annotation)

Scientometrics
Remove ISI Duplicate Records
Remove Rows with Multitudinous Fields
Detect Duplicate Nodes
Update Network by Merging Nodes

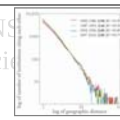
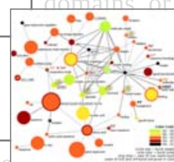
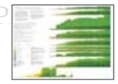
Extract Directed Network
Extract Paper Citation Network
Extract Author Paper Network

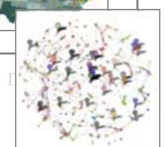
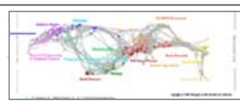
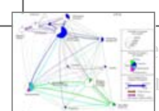
Extract Co-Occurrence Network
Extract Word Co-Occurrence Network
Extract Co-Author Network
Extract Reference Co-Occurrence (Bibliographic Coupling) Network
----
Extract Document Co-Citation Network

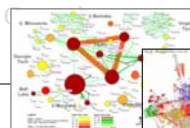
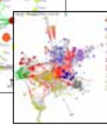
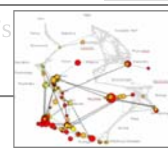
Soon:
Database support for ISI and NSF data.

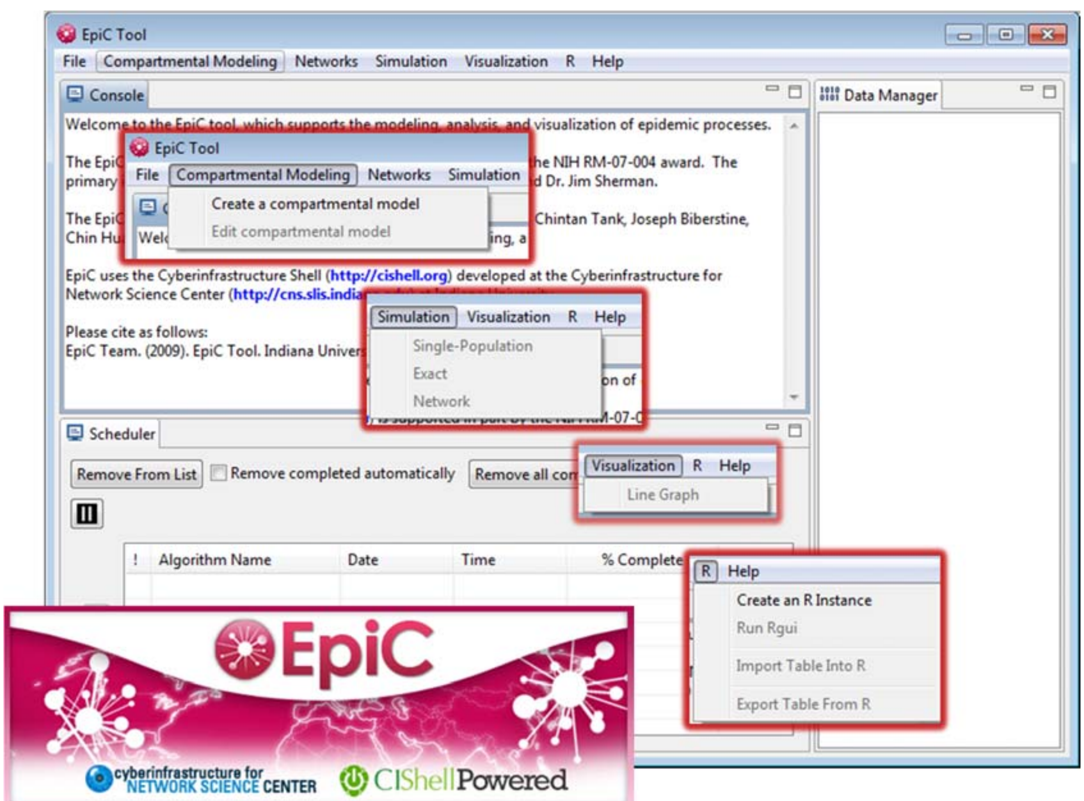
Type of Analysis vs. Level of Analysis

	Micro/Individual (1-100 records)	*Meso/Local (101–10,000 records)*	*Macro/Global (10,000 < records)*
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NS... SA, all of sci...
Temporal Analysis (When)	Funding portfolio of one individual	...ic bursts ...f PNAS	113 Years of P... Research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a s... intellectual l...	PNAS ...
Topical Analysis (What)	...s.	...flows in ... research	VxOrd/Topic ... NIH funding
Network Analysis (With Whom?)	NSF ...work of one	C... k	NIH's ... cy



References

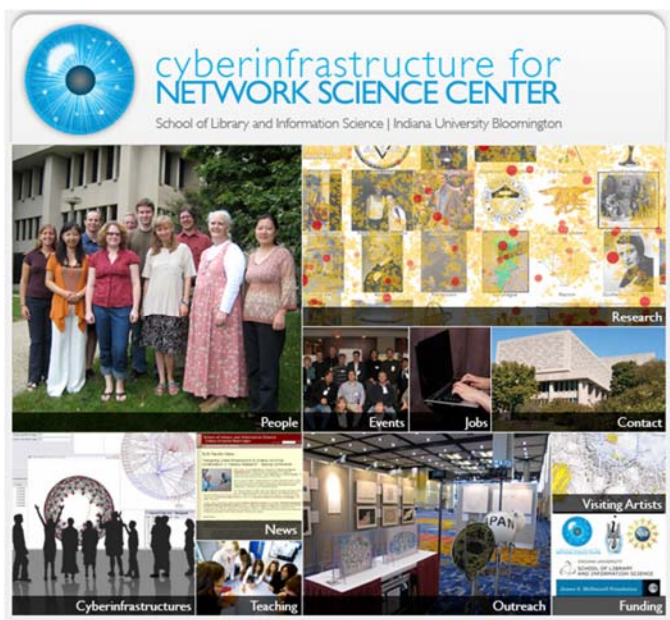
Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). **Visualizing Knowledge Domains.** In Blaise Cronin (Ed.), *ARIST*, Medford, NJ: Information Today, Volume 37, Chapter 5, pp. 179-255. http://ivl.slis.indiana.edu/km/pub/2003-borner-arist.pdf

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). http://www.pnas.org/content/vol101/suppl_1/

Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (2007). **Network Science.** In Blaise Cronin (Ed.), *ARIST*, Information Today, Inc., Volume 41, Chapter 12, pp. 537-607.
http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf

Börner, Katy (2010) **Atlas of Science.** MIT Press.
http://scimaps.org/atlas

Scharnhorst, Andrea, Börner, Katy, van den Besselaar, Peter (2011) **Models of Science Dynamics.** Springer Verlag.



All papers, maps, tools, talks, press are linked from http://cns.iu.edu

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