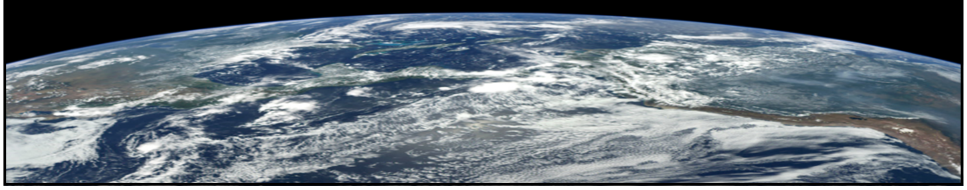
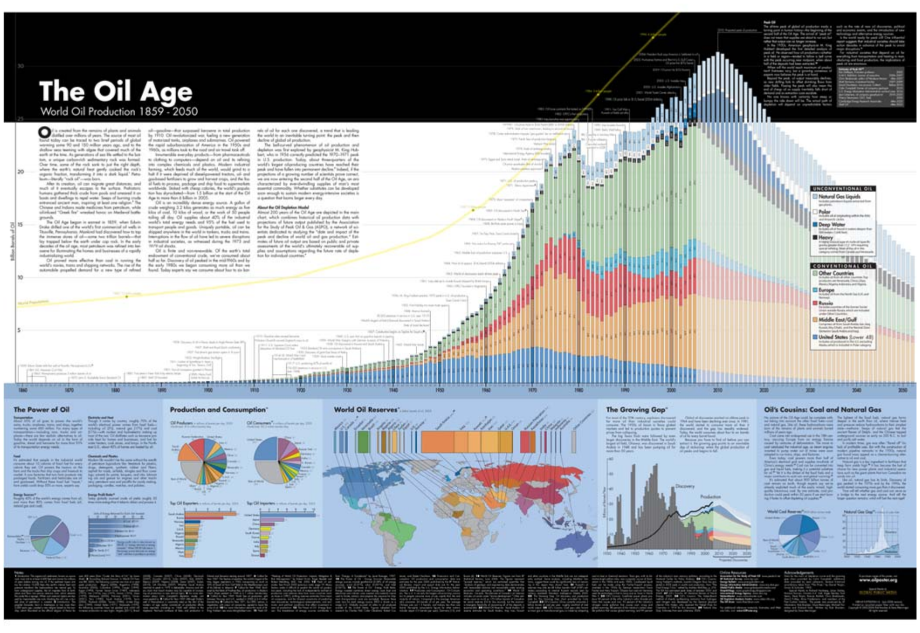
Modelling Co-Evolving Scholarly Networks and the Collective Allocation of Research Funding, and Broadcasting STI Forecasts

Katy Börner, Cyberinfrastructure for Network Science Center
School of Informatics and Computing, Indiana University, USA
Royal Netherlands Academy of Arts and Sciences (KNAW), The Netherlands

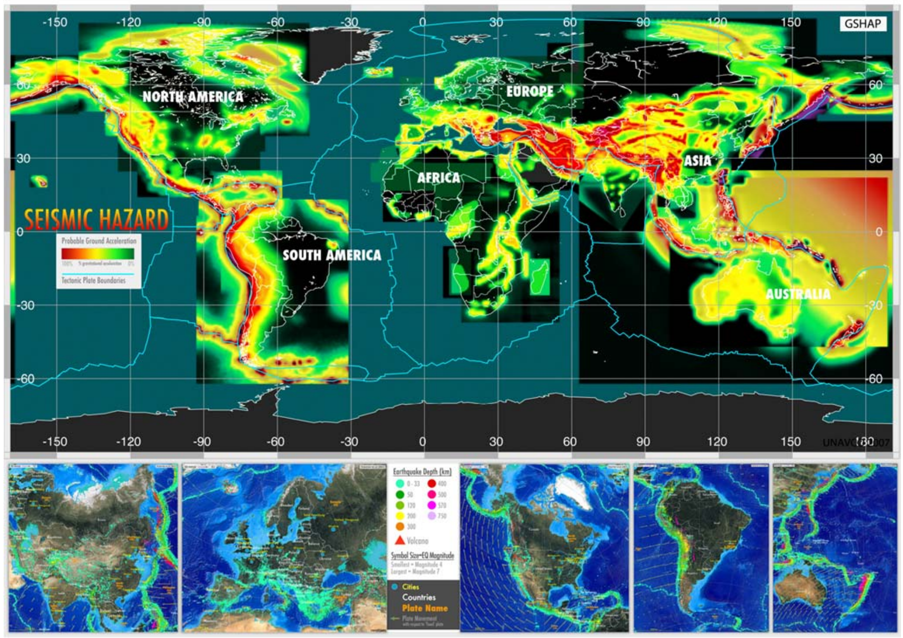
Planning, Prediction, Scenarios--Using Simulations and Maps Conference
European Academy of Technology and Innovation Assessment, Bonn, Germany

May 11, 2015

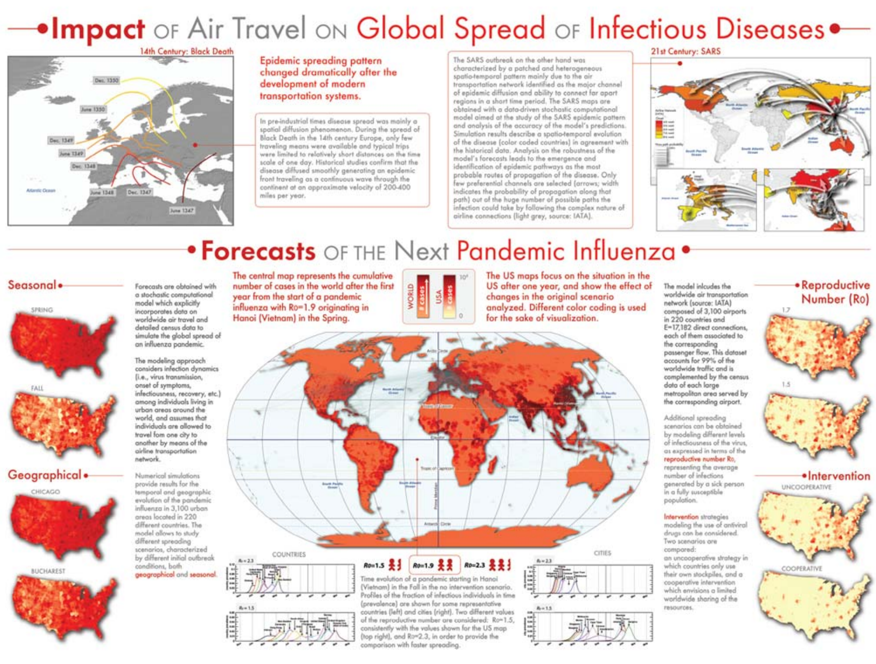




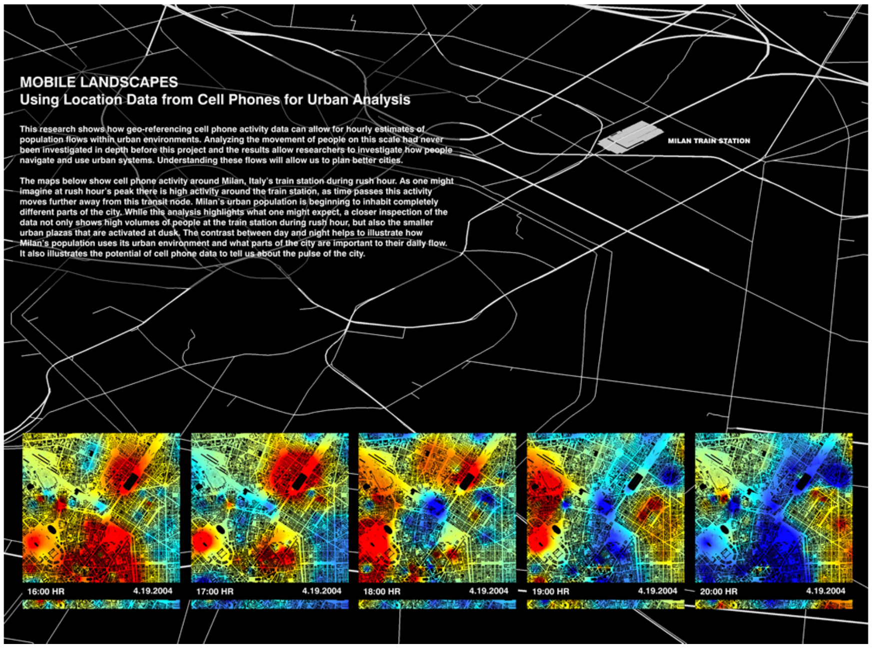
The Oil Age: World Oil Production 1859-2050 - Rob Bracken, Dave Menninger, Michael Poremba, Richard Katz - 2005



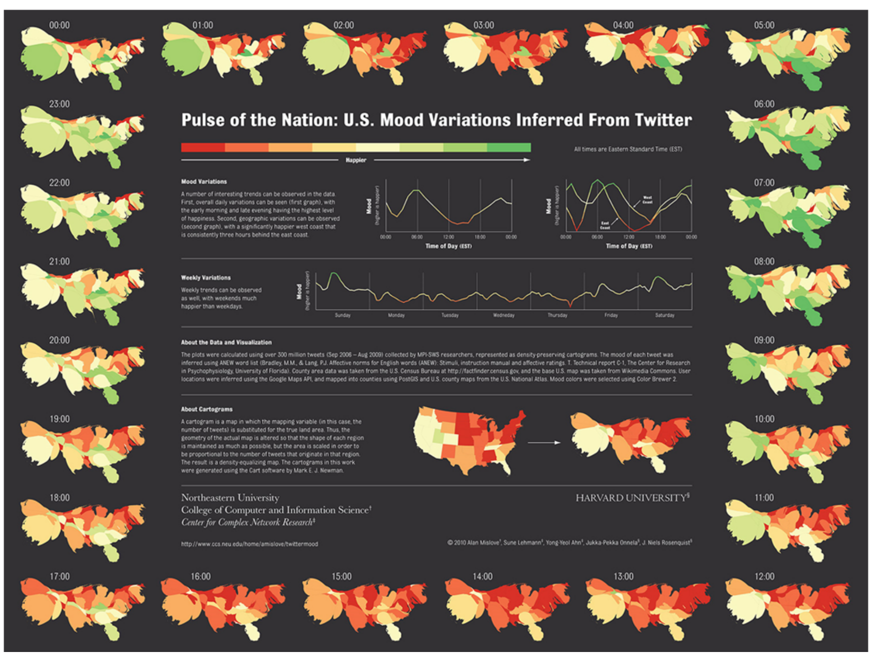
Tectonic Movements and Earthquake Hazard Predictions - Martin W. Hamburger, Lou Estey, Chuck Meertens, Elisha Hardy - 2005



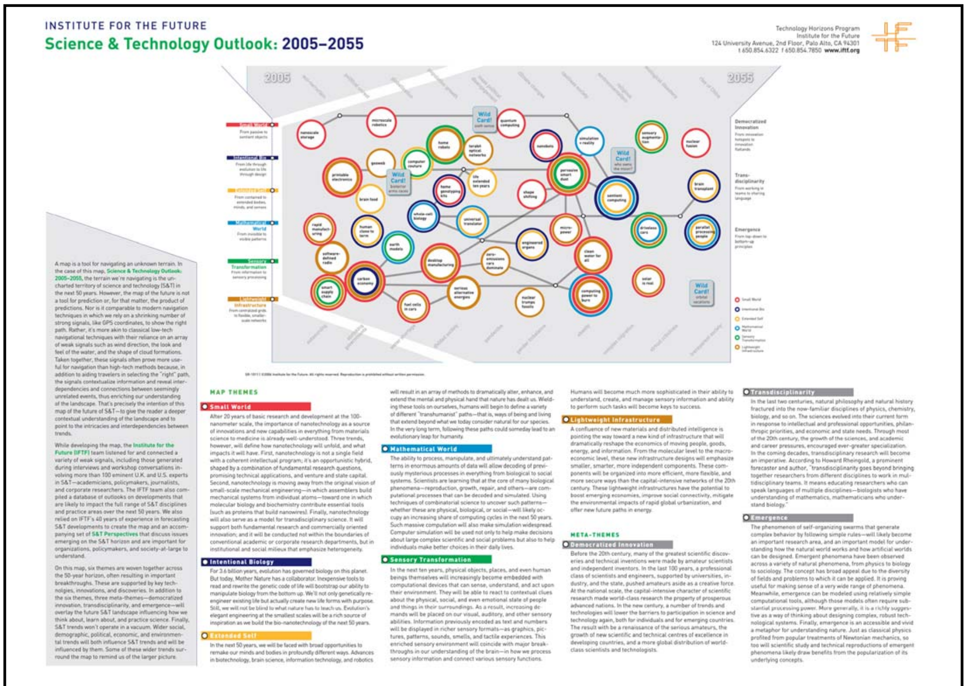
Impact of Air Travel on Global Spread of Infectious Diseases - Vittoria Colizza, Alessandro Vespignani - 2007



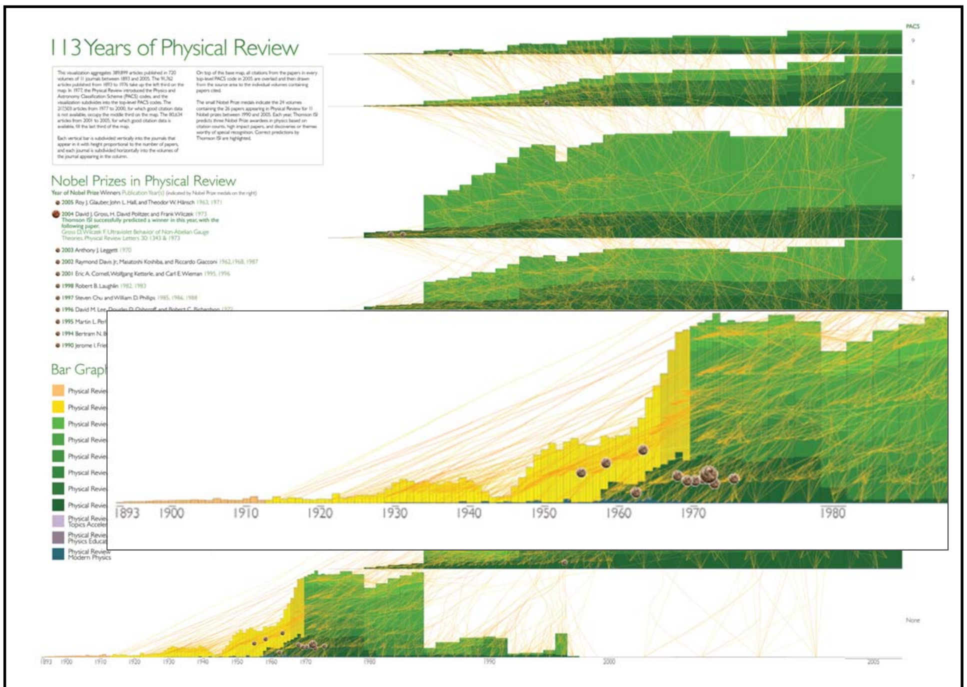
Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis - Sarah Williams, Carlo Ratti, and Riccardo Maria Pulselli - 2006



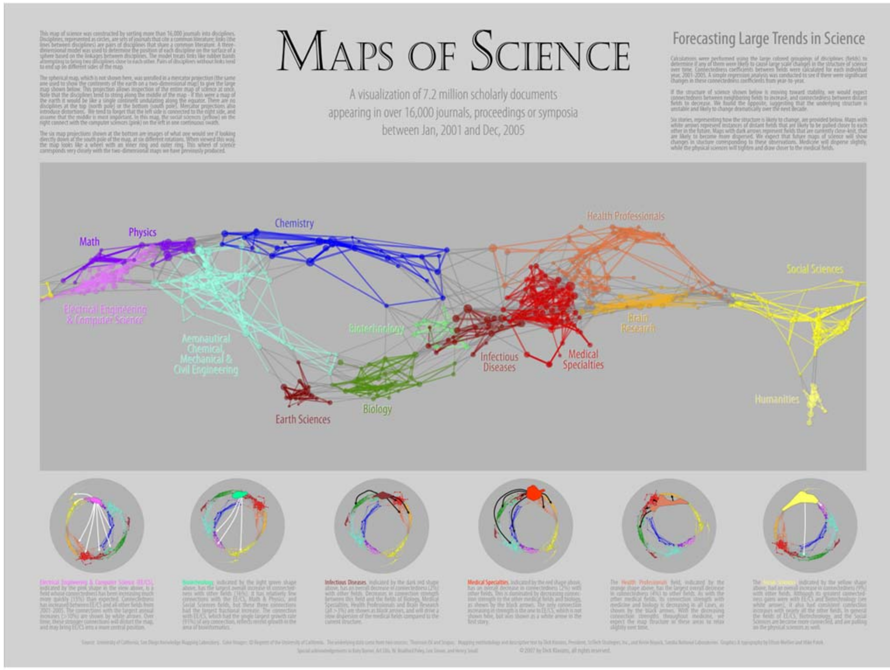
Pulse of the Nation - Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and James Niels Rosenquist - 2010



Science & Technology Outlook: 2005-2055 - Alex Soojung-Kim Pang, David Pescovitz, Marina Gorbis, Jean Hagan - 2006



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



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

Modelling Science

10

Part I: Foundations

1 An Introduction to Modeling Science: Basic Model Types, Key Definitions, and a General Framework for the Comparison of Process Models
Borner, Boyack, Milojevic & Morris

2 Mathematical Approaches to Modeling Science from an Algorithmic-Historiography Perspective by Lucio-Arias & Scharnhorst

Part II: Exemplary Model Types

3 Knowledge Epidemics and Population Dynamics Models for Describing Idea Diffusion by Vitanov & Ausloos

4 Agent-Based Models of Science by Payette

5 Evolutionary Game Theory and Complex Networks of Scientific Information by Hanauske

Part III: Exemplary Model Applications

6 Dynamic Scientific Co-Authorship Networks by Mali, Kronegger, Doreian & Ferligoj

7 Citation Networks by Radicchi, Fortunato & Vespignani

Part IV: Outlook

8 Science Policy and the Challenges for Modeling Science by van den Besselaar, Borner & Scharnhorst

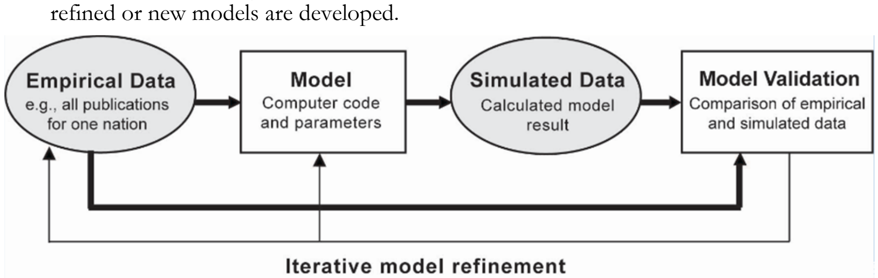
11

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.*

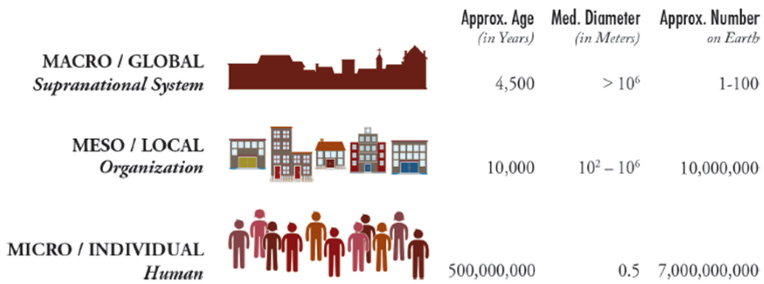
Modeling Process

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.



12

Sample Model #1
Modelling Co-Evolving Author-Paper Networks (MESO)



13

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.

14

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.

```
// 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;
}
```

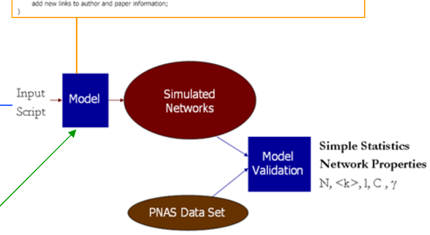
15

Table 3 Statistics for SIM data

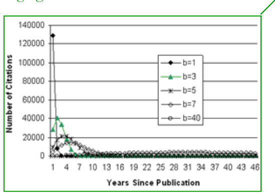
Year	#p	#a	#r	#c	a#ca
1981	1624	3953	0	756	8.21
1982	1040	5200	31200	112161	4
1983	1118	5590	33540	21397	4
1984	1197	5985	35910	10224	4
1985	1275	6375	38250	6184	4
1986	1353	6765	40590	4687	4
1987	1432	7160	42960	3573	4
1988	1510	7550	45300	2816	4
1989	1589	7945	47670	2219	4
1990	1667	8335	50010	1853	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**	

```
-----------------------------------------
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

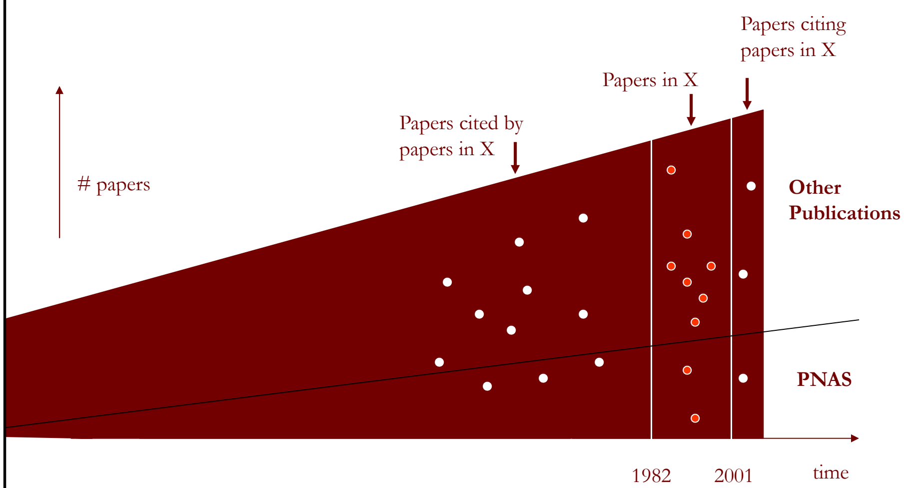
The properties of the networks generated by this model are validated against a 20-year data set (1982-2001) of documents of type article published in the Proceedings of the National Academy of Science (PNAS) – about 106,000 unique authors, 472,000 co-author links, 45,120 papers cited within the set,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

Table 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**	

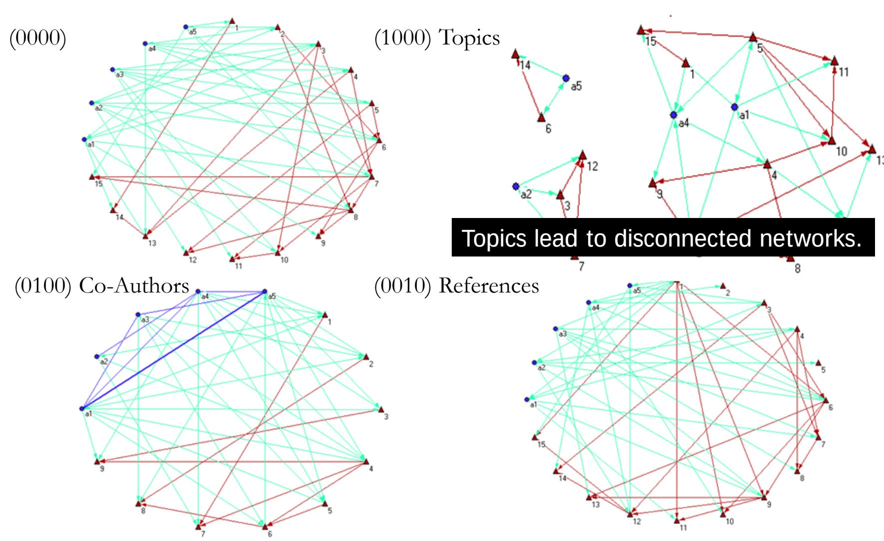
20-Year PNAS Dataset (1982-2001)

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



17

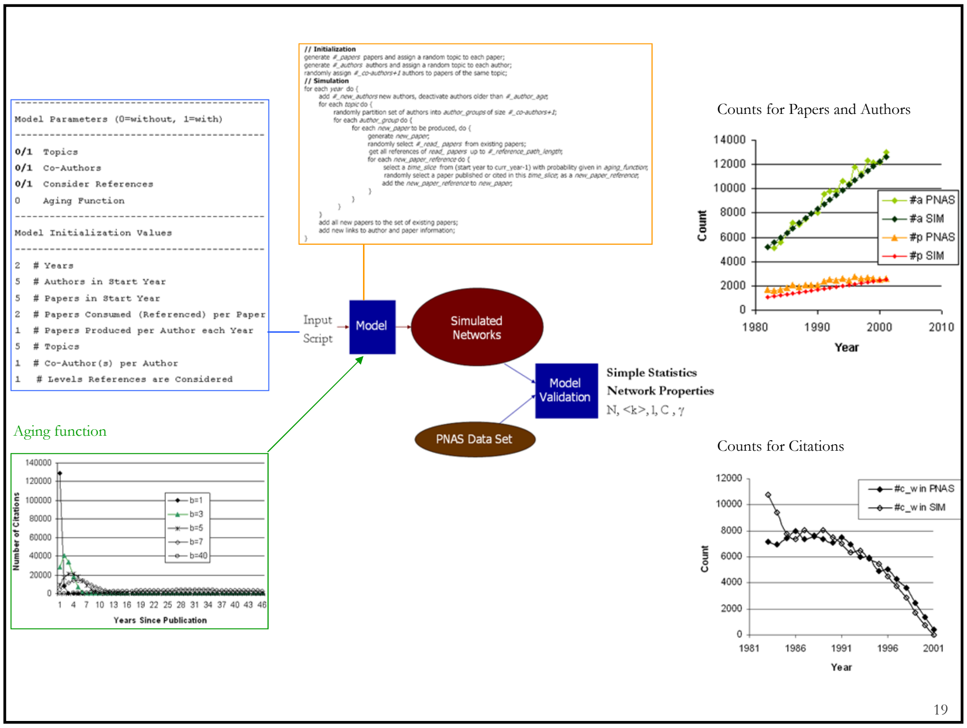
The TARL Model: The Effect of Parameters



Topics lead to disconnected networks.

Co-authoring leads to fewer papers.

18



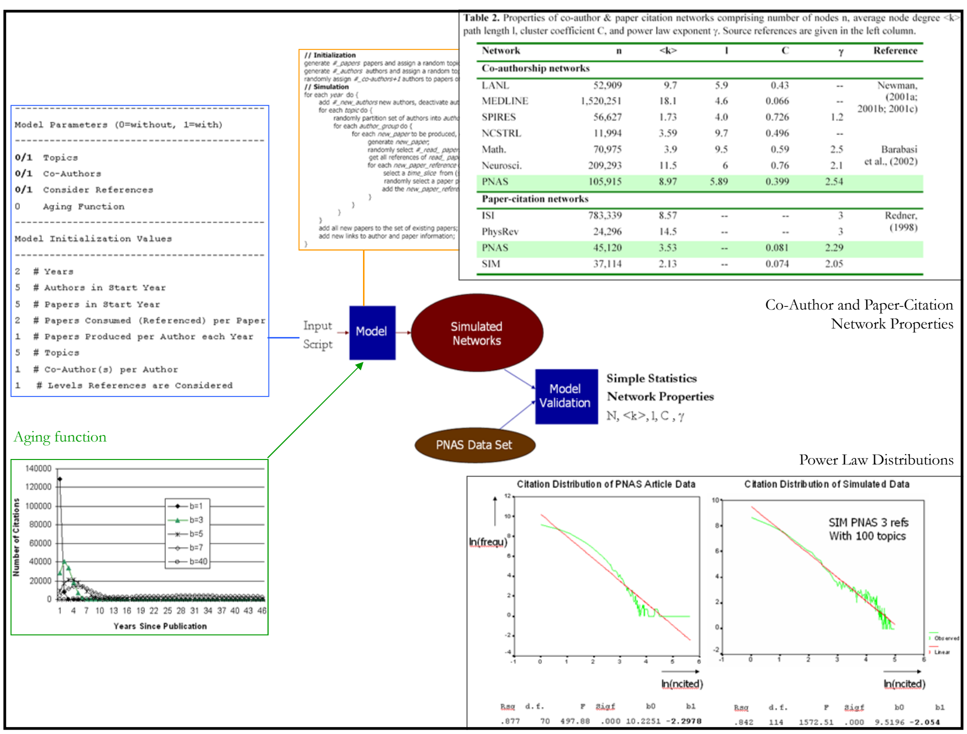
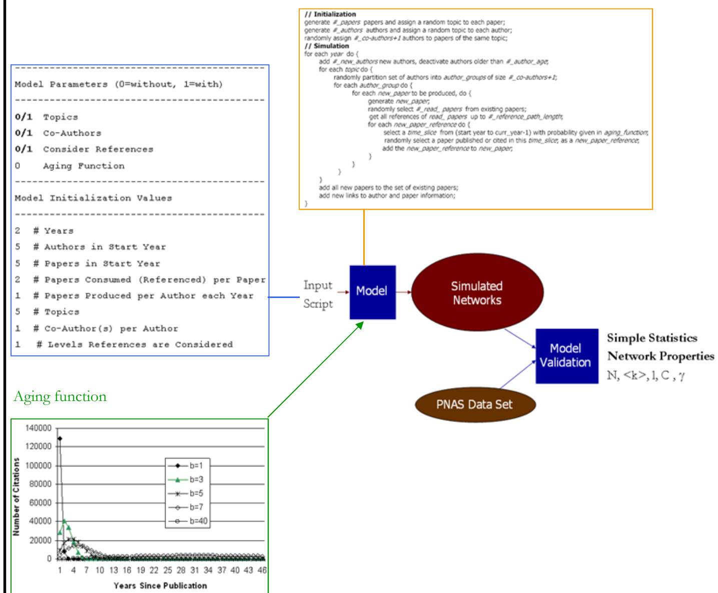


Table 2. Properties of co-author & paper citation networks comprising number of nodes n, average node degree <k> path length l, cluster coefficient C, and power law exponent γ. Source references are given in the left column.

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	



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.

21

Sample Model #2

Collective allocation of science funding as an alternative to peer review

(MESO)

	Approx. Age (in Years)	Med. Diameter (in Meters)	Approx. Number on Earth
MACRO / GLOBAL *Supranational System*	4,500	$> 10^6$	1-100
MESO / LOCAL *Organization*	10,000	$10^2 – 10^6$	10,000,000
MICRO / INDIVIDUAL *Human*	500,000,000	0.5	7,000,000,000

22

NEWSFOCUS

Making Every Scientist a Research Funder

When it comes to using peer review to distribute research dollars, Johan Bollen favors radical simplicity.

Over the years, many scientists have suggested that the current system could be improved by changing the composition of the review panels, tweaking the interactions among reviewers, or revising how the proposals are scored. But Bollen, a computer scientist at Indiana University, Bloomington, would simply award all eligible researchers a block grant—and then require them to give some of it away to colleagues they judge most deserving.

That radical step, described in a paper Bollen and four Indiana colleagues recently posted on *EMBO Reports*, retains peer review's core concept of tapping into the views of the most knowledgeable researchers. But it would eliminate the huge investment in time and money required to submit proposals and assemble panels to judge them.

Bollen's process would be almost instantaneous: In a version of expert-directed crowdsourcing, scientists would fill out a form once a year listing their favored researchers, and a predetermined portion of their annual grant money—a total of, say, 50%—would then be transferred to their choices.

"So many scientists spend so much time on peer review, and there's a high level of frustration," Bollen explains. "We already know who the best people are. And if you're doing good work, then you deserve to receive support."

Others are skeptical. "I've known Johan for a long time and have the highest regard for his ability as an out-of-the-box thinker," says Stephen Griffin, a retired National Science Foundation (NSF) program manager who's now a visiting professor of information sciences at the University of Pittsburgh in Pennsylvania. "But there are a number of issues he doesn't address."

Those sticking points include the likely mismatch between what researchers need and what their colleagues give them; the absence of any replacement for the overhead payments in today's grants, which support infrastructure at host institutions; and the dearth of public accountability for the billions of dollars that would flow from public coffers to individuals. "Scientists aren't really equipped to be a funding agency," Griffin notes.

Bollen acknowledges that the process would need safeguards to ensure that scientists don't reward their friends or punish their enemies. But his analysis suggests that the U.S. research landscape would not look all that different if his radical proposal were adopted.

Drawing upon citation data in 37 million papers over 20 years, the Indiana researchers conducted a simulation premised on the idea that scientists would reallocate their federal dollars according to how often they cited their peers. The simulation, he says, yielded a funding pattern "similar in shape to the actual distribution" at NSF and the National Institutes of Health for the past decade—at a fraction of the overhead required by the current system.

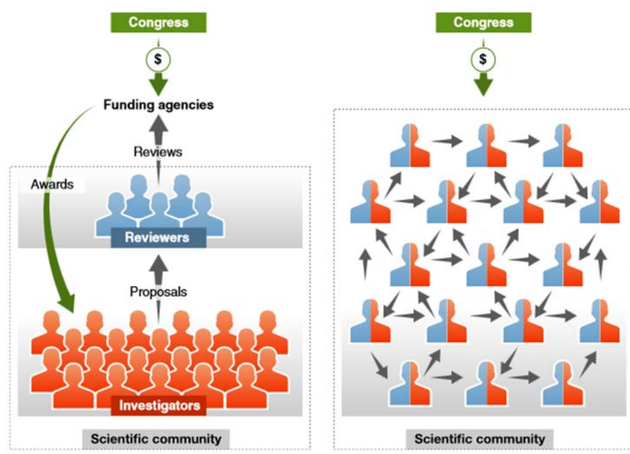
–JDM

February 7, 2014

Science 7 February 2014: Vol. 343 no. 6171 p. 598
DOI: 10.1126/science.343.6171.598
http://www.sciencemag.org/content/343/6171/598.full?sid=4f40a7f0-6ba2-4ad8-a181-7ab394fe2178

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

Bollen, Johan, David Crandall, Damion Junk, Ying Ding, and Katy Börner. 2014. EMBO Reports 15 (1): 1-121.



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.

24

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

Bollen, Johan, David Crandall, Damion Junk, Ying Ding, and Katy Börner. 2014. EMBO Reports 15 (1): 1-121.

Assume

Total funding budget in year y is t_y

Number of qualified scientists is n

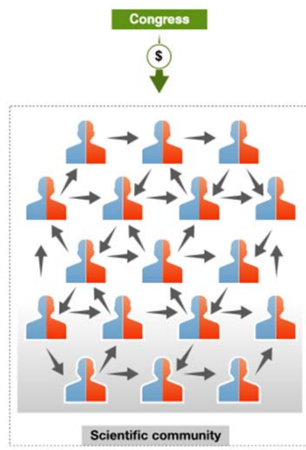
Each year,

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 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.



25

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

Bollen, Johan, David Crandall, Damion Junk, Ying Ding, and Katy Börner. 2014. EMBO Reports 15 (1): 1-121.

Example:

Total funding budget in year is 2012 NSF budget

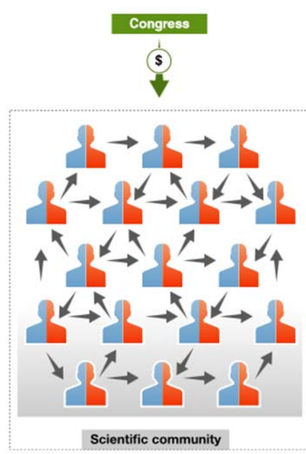
Given the number of NSF funded scientists, each receives a $100,000 basic grant.

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.



26

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

Bollen, Johan, David Crandall, Damion Junk, Ying Ding, and Katy Börner. 2014. EMBO Reports 15 (1): 1-121.

Model Run and Validation:

Model is presented in http://arxiv.org/abs/1304.1067

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

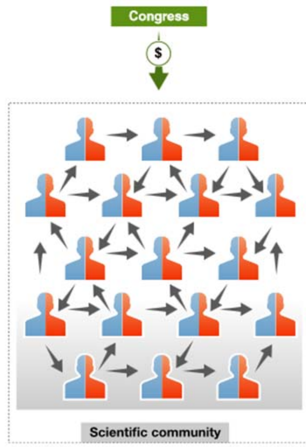
Using 37M articles from TR 1992 to 2010 Web of Science (WoS) database, we extracted **770M citations**. From the same WoS data, we also determined 4,195,734 unique author names and we took the **867,872 names** who had authored at least one paper per year in any five years of the period 2000–2010.

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.



27

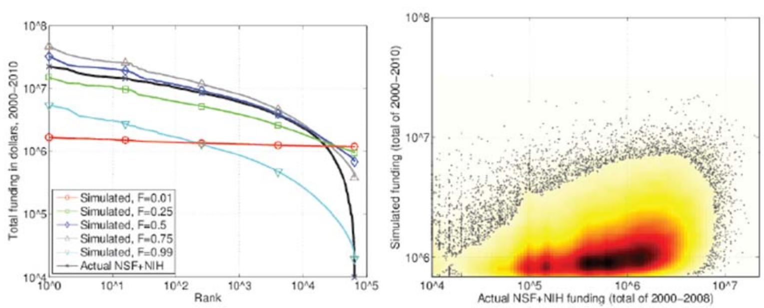


Fig. 2: Results of the distributed funding system simulation for 2000-2010. *(a):* The general shape of the funding distribution is similar to that of actual historical NSF and NIH funding distribution. The shape of the distribution can be controlled by adjusting F, the fraction of funds that scientists must give away each year. *(b):* On a per-scientist basis, simulated funding from our system (with F=0.5) is correlated with actual NSF and NIH funding (Pearson $R = 0.2683$ and Spearman $\rho = 0.2999$).

28

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

Bollen, Johan, David Crandall, Damion Junk, Ying Ding, and Katy Börner. 2014. EMBO Reports 15 (1): 1-121.

Model Efficiency:

Using data from the Taulbee Survey of Salaries Computer Science (http://cra.org/resources/taulbee) and the National Science Foundation (NSF) the following calculation is illuminating:

If four professors work four weeks full-time on a proposal submission, labor costs are about $30k. With typical funding rates below 20%, about five submission-review cycles might be needed resulting in a total expected labor cost of **$150k**.

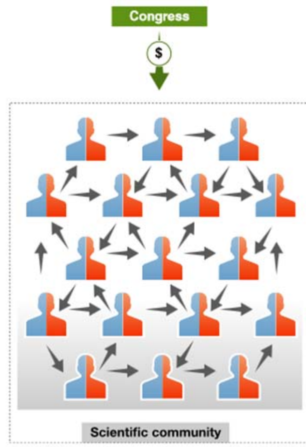
The average NSF grant is **$128k** per year.

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 research.

That is, U.S. universities should forbid professors to apply for grants—if they can afford to forgo the indirect dollars.

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



29

Sample Model #3

Monitoring, Modeling, and Forecasting Tools for Fostering an Innovative S&T Workforce (MICRO … MACRO)

	Approx. Age *(in Years)*	Med. Diameter *(in Meters)*	Approx. Number *on Earth*
MACRO / GLOBAL *Supranational System*	4,500	> 10^6	1-100
MESO / LOCAL *Organization*	10,000	$10^2 - 10^6$	10,000,000
MICRO / INDIVIDUAL *Human*	500,000,000	0.5	7,000,000,000

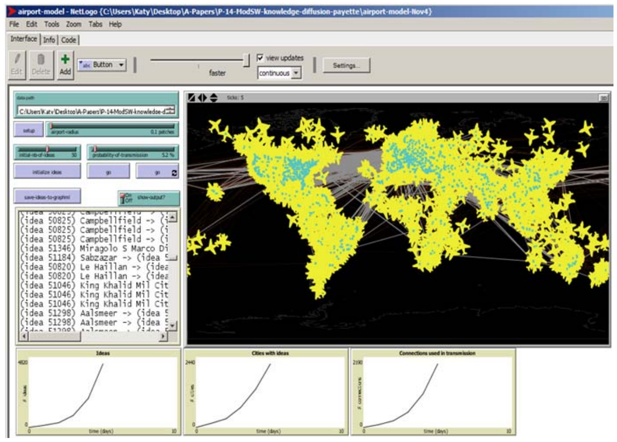
30

Monitoring, Modeling, and Forecasting Tools for Fostering an Innovative S&T Workforce
With Nicolas Payette. Work in progress.

This project aims to develop monitoring, modeling, and forecasting approaches and tools for fostering an innovative science and technology workforce.

Large-scale datasets of scholarly activity including funding, publications, patents, and job openings among others are analyzed and modeled.

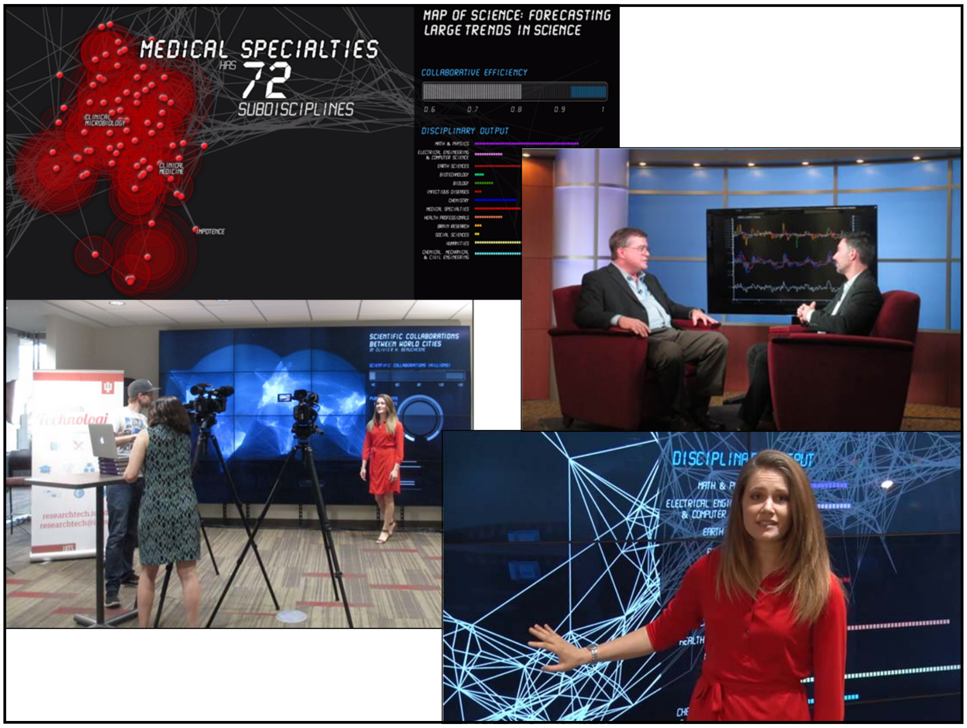
NetLogo is used to study the impact of transportation pathways (here airport traffic data) on co-authorships formed. We also model the diffusion of ideas via transportation and collaboration networks.



31

Broadcasting STI Model Results

32



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/

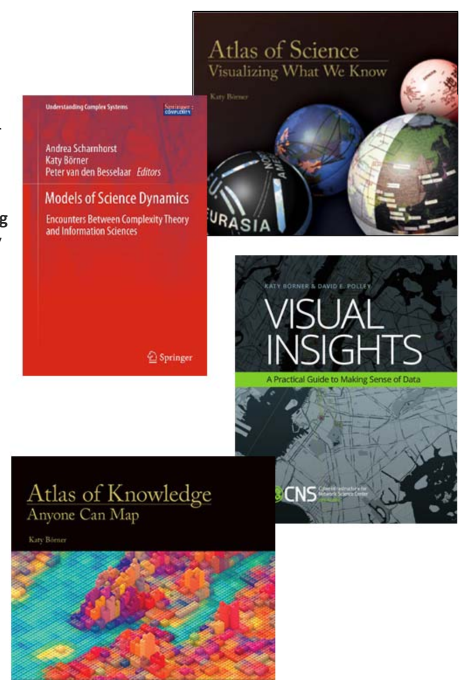
Börner, Katy (2010) **Atlas of Science: Visualizing What We Know**. The MIT Press. http://scimaps.org/atlas

Scharnhorst, Andrea, Börner, Katy, van den Besselaar, Peter (2012) **Models of Science Dynamics**. Springer Verlag.

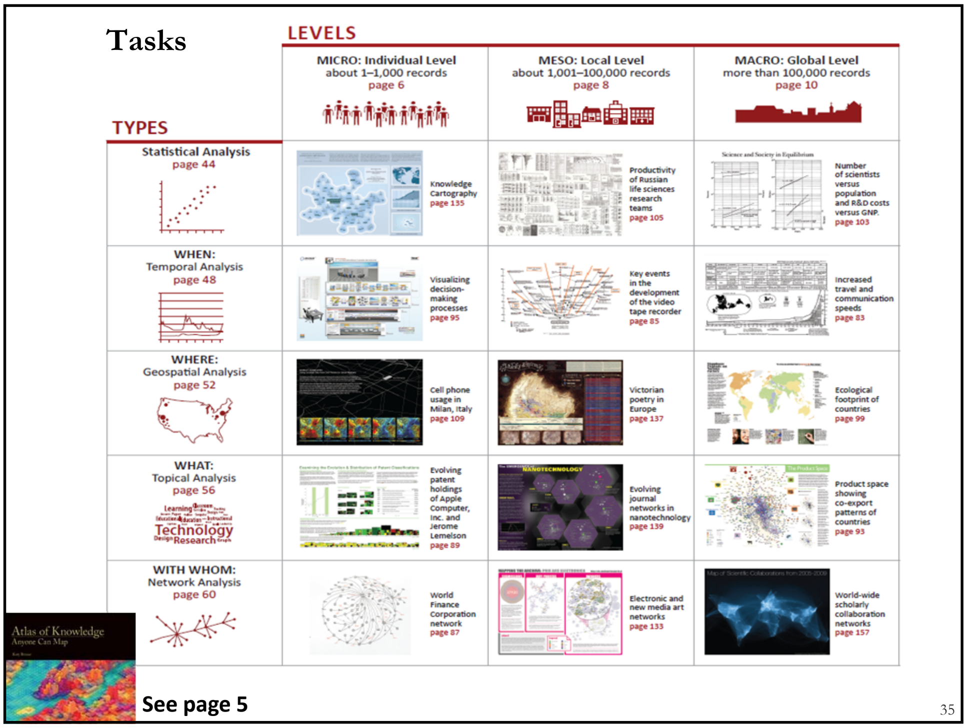
Katy Börner, Michael Conlon, Jon Corson-Rikert, Cornell, Ying Ding (2012) **VIVO: A Semantic Approach to Scholarly Networking and Discovery**. Morgan & Claypool.

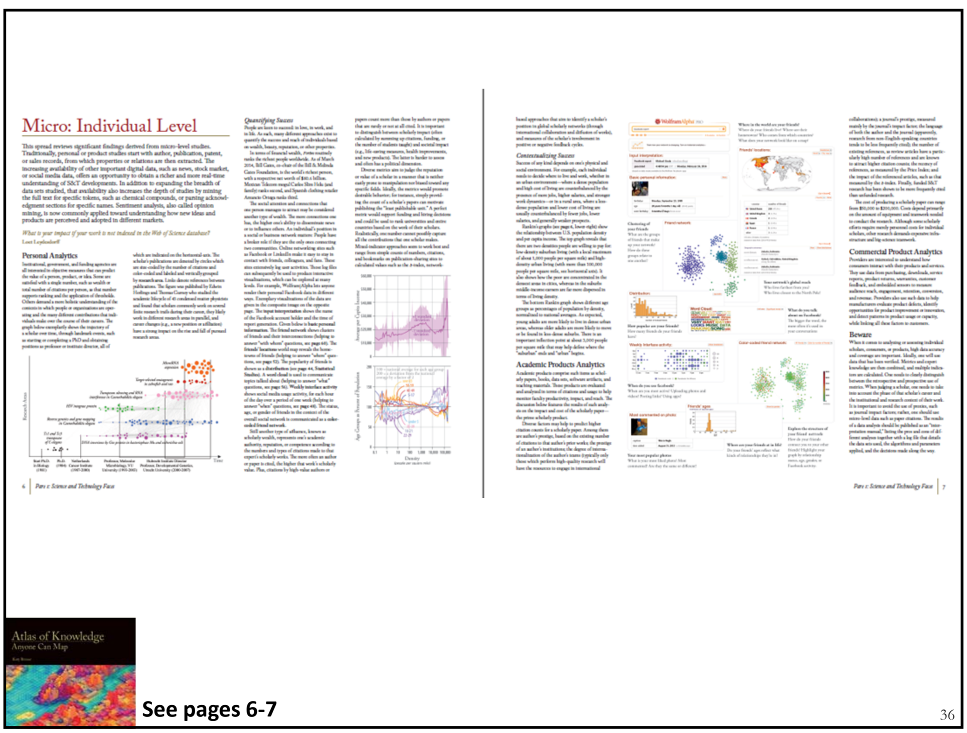
Katy Börner and David E Polley (2014) **Visual Insights: A Practical Guide to Making Sense of Data**. The MIT Press.

Börner, Katy (2015) **Atlas of Knowledge: Anyone Can Map**. The MIT Press. http://scimaps.org/atlas2

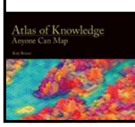


34


Tasks
LEVELS
MICRO: Individual Level
about 1–1,000 records
page 6
MESO: Local Level
about 1,001–100,000 records
page 8
MACRO: Global Level
more than 100,000 records
page 10
TYPES
Statistical Analysis
page 44
Knowledge Cartography
page 135
Productivity of Russian life sciences research teams
page 105
Number of scientists versus population and R&D costs versus GNP.
page 103
WHEN:
Temporal Analysis
page 48
Visualizing decision-making processes
page 95
Key events in the development of the video tape recorder
page 85
Increased travel and communication speeds
page 83
WHERE:
Geospatial Analysis
page 52
Cell phone usage in Milan, Italy
page 109
Victorian poetry in Europe
page 137
Ecological footprint of countries
page 99
WHAT:
Topical Analysis
page 56
Evolving patent holdings of Apple Computer, Inc. and Jerome Lemelson
page 89
Evolving journal networks in nanotechnology
page 139
Product space showing co-export patterns of countries
page 93
WITH WHOM:
Network Analysis
page 60
World Finance Corporation network
page 87
Electronic and new media art networks
page 133
World-wide scholarly collaboration networks
page 157
Atlas of Knowledge
Anyone Can Map
See page 5
35

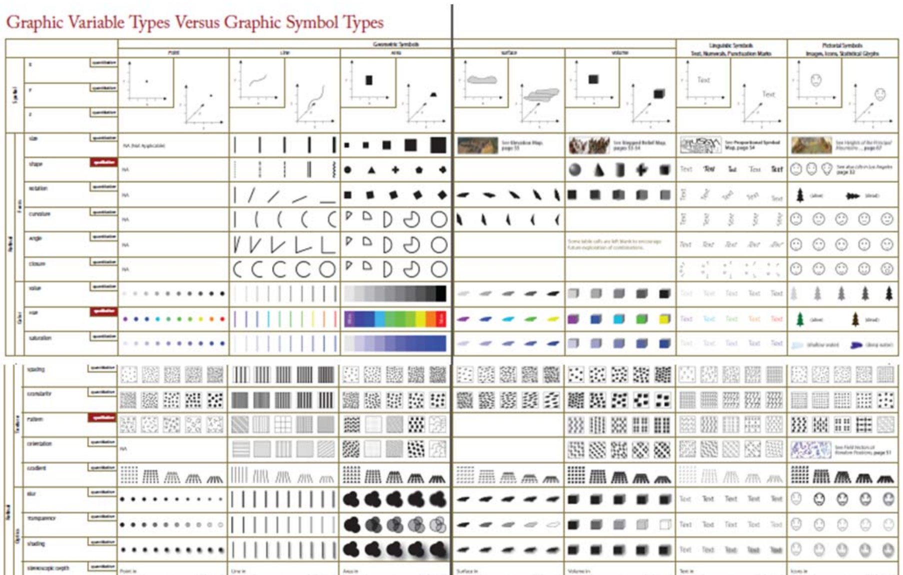

Micro: Individual Level
Atlas of Knowledge
Anyone Can Map
See pages 6-7
36

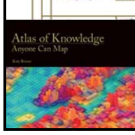
Insight Need Types page 26	Data Scale Types page 28	Visualization Types page 30	Graphic Symbol Types page 32	Graphic Variable Types page 34	Interaction Types page 26
• categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships	• nominal • ordinal • interval • ratio	• table • chart • graph • map • network layout	• geometric symbols point line area surface volume • linguistic symbols text numerals punctuation marks • pictorial symbols images icons statistical glyphs	• spatial position • retinal form color optics motion	• overview • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion



See page 24

Graphic Variable Types Versus Graphic Symbol Types





See pages 36-39

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

39