Multi-Level Science Models

Katy Börner
CNS, School of Informatics and Computing, Indiana University, USA
Royal Netherlands Academy of Arts and Sciences (KNAW), The Netherlands
Science, Technology and Innovation Visiting Research Fellow at OECD, France

Workshop on Simulating the Social Processes of Science
Lorentz Center, Leiden, Netherlands
April 7-11, 2014

Science and Society in Equilibrium

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

113 Years of Physical Review - Bruce W. Herr II, Russell Duhon, Katy Bomer, Elisha Hardy, Shashikant Penumarthi - 2007
Impact of Communication and Transportation Speeds

Shrinking of Our Planet - R. Buckminster Fuller, John McHale - 1965

Table:

<table>
<thead>
<tr>
<th>Year</th>
<th>2016</th>
<th>2026</th>
<th>2036</th>
<th>2046</th>
<th>2056</th>
<th>2066</th>
<th>2076</th>
<th>2086</th>
<th>2096</th>
<th>2106</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heat of evolution</td>
<td>2.5°C</td>
<td>5°C</td>
<td>7°C</td>
<td>9°C</td>
<td>11°C</td>
<td>13°C</td>
<td>15°C</td>
<td>17°C</td>
<td>19°C</td>
<td>21°C</td>
</tr>
<tr>
<td>Communication speeds</td>
<td>10 Gbps</td>
<td>100 Gbps</td>
<td>1 Tbps</td>
<td>10 Tbps</td>
<td>100 Tbps</td>
<td>1 Pbps</td>
<td>10 Pbps</td>
<td>100 Pbps</td>
<td>1 Ebps</td>
<td>1 Zbps</td>
</tr>
</tbody>
</table>

Graph:

- The relative size of the world as travel time decreases.
- Shrinkage of the planet as communication speeds increase.

Caption:

Shrinking of Our Planet - R. Buckminster Fuller, John McHale - 1965
Impact of Air Travel on Global Spread of Infectious Diseases

Forecasts of the Next Pandemic Influenza

Map of Scientific Paradigms

Science can be thought of as a combination of theories and paradigms. There are areas of current research, while paradigms comprise the dominant sets of ideas and existing knowledge that are used by today's researchers. This map shows 35 major paradigms in science, which are represented by white and yellow circles. The largest circles represent the most influential paradigms, while the smallest circles represent the least influential paradigms. The size of each circle is proportional to the number of papers that cite papers associated with that paradigm. The color of each circle is determined by the publication year of the first reference associated with that paradigm. The yellow color represents newer paradigms, while the white color represents older paradigms.

The map of scientific paradigms and the reference system can be used to explore the current state of research in various fields. Researchers can use the map to identify areas of research that are currently active and areas that are not. This information can be useful for planning future research efforts and for identifying potential new areas of study.

Map of Scientific Paradigms - Kevin W. Boyack, Richard Klavans - 2006
Model Types

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


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.

Sample Model #1
PNAS Co-Evolving Author-Paper Networks (MESO)
Modeling the Co-Evolving Author-Paper Networks


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.

// Initialization
generate n_papers papers and assign a random topic to each paper;
generate n_authors authors and assign a random topic to each author;
randomly assign n_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add n_new_authors new authors, deactivate authors older than n_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size n_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select n_read_papers from existing papers;
                get all references of read_papers up to n_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;
        }
    }
}
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.

### 20-Year PNAS Dataset (1982-2001)

Coverage in terms of time span, total number of papers, and complete author's work
The TARL Model: The Effect of Parameters

Co-authoring leads to fewer papers.

Topics lead to disconnected networks.
Aging function

**Aging function**

### Topics
- The number of topics is linearly correlated with the clustering coefficient of the resulting network: \( C = 0.000073 \times \#\text{topics} \). 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.
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.
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 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 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.
From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review

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 $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.

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 ρ = 0.2683 and Spearman ρ = 0.2999).
From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review


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.

References


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