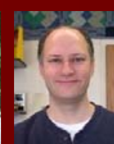
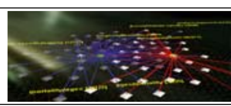
Modeling the Co-Evolution of Scholarly Networks

Katy Börner
School of Library and Information Science
INDIANA UNIVERSITY
BLOOMINGTON
katy@indiana.edu

This is joint work with
Jeegar T. Maru, Computer Science, jmaru@indiana
Robert L. Goldstone, Psychology, rgoldsto@indiana

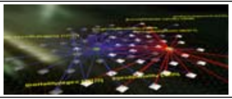




Overview

1. Mapping Scientific Structure and Evolution
   - Descriptive Models vs. Process Models
   - Isolated Networks vs. Network Ecologies
   - The Influence of Time and Semantics on (Scientific) Network Evolution
2. The TARL Model
   - Model Design
   - Model Validation Using a 20 Year PNAS Data Set
   - The Influence of Model Parameters
3. Discussion & Future Work

Dominated by research in biology



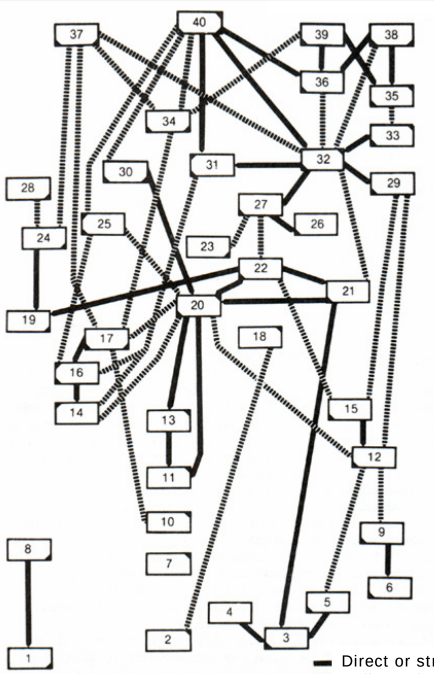
1. Mapping Scientific Structure and Evolution

To answer questions such as:

- What are the major research areas, experts, institutions, regions, nations, grants, publications, journals in xx research?
- Which areas are most insular?
- What are the main connections for each area?
- What is the relative speed of areas?
- Which areas are the most dynamic/static?
- What new research areas are evolving?
- Impact of xx research on other fields?
- How does funding influence the number and quality of publications?

Answers are needed by funding agencies, companies, and us researchers.

Katy Börner, Jeegar T. Maru & Robert L. Goldstone: Modeling the Co-Evolution of Scholarly Networks.



Direct or strongly implied citation
Indirect citation

Figure 6.3 Historiograph of DNA development.

KEY

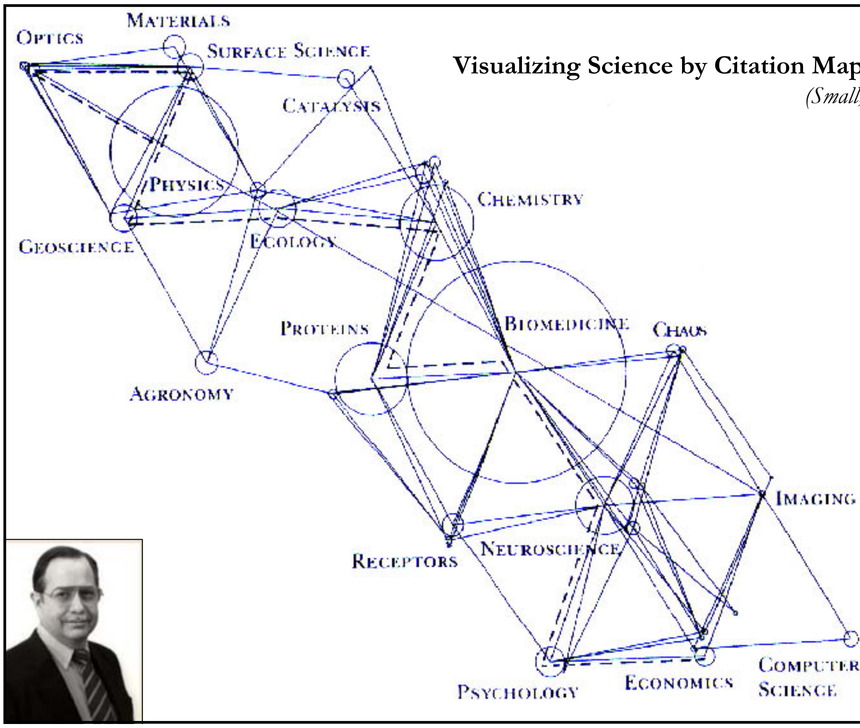
1. Braconnot 1820
2. Mendel 1865
3. Miescher 1871
4. Flemming 1879
5. Kossel 1886
6. Fischer and Piloty 1891
7. DeVries 1900
8. Fischer 1907
9. Levene and Jacobs 1909
10. Muller 1926
11. Griffith 1928
12. Levene with Mori and London 1929
13. Alloway 1932
14. Stanley 1935
15. Levene and Tipson 1935
16. Bawden and Pirie 1936-1937
17. Caspersson and Schultz 1938-1939
18. Beadle and Tatum 1941
19. Martin and Synge 1943-1944
20. Avery, MacLeod, and McCarty 1944
21. Chargaff 1947
22. Chargaff 1950
23. Pauling and Corey 1950-1951
24. Sanger 1951-1953
25. Hershey and Chase 1952
26. Wilkins 1953
27. Watson and Crick 1953
28. DuVigneaud 1953
29. Todd 1955
30. Palade 1954-1956
31. Fraenkel-Conrat 1955-1957
32. Ochoa 1955-1956
33. Kornberg 1956-1957
34. Hoagland 1957-1958
35. Jacob and Monod 1960-1961
36. Hurwitz 1960
37. Dintzis 1961
38. Novelli 1961-1962
39. Allfrey and Mirsky 1962
40. Nirenberg and Matthaei 1961-1962

Historiograph of DNA Development

(Garfield, Sher, & Torpie, 1964)

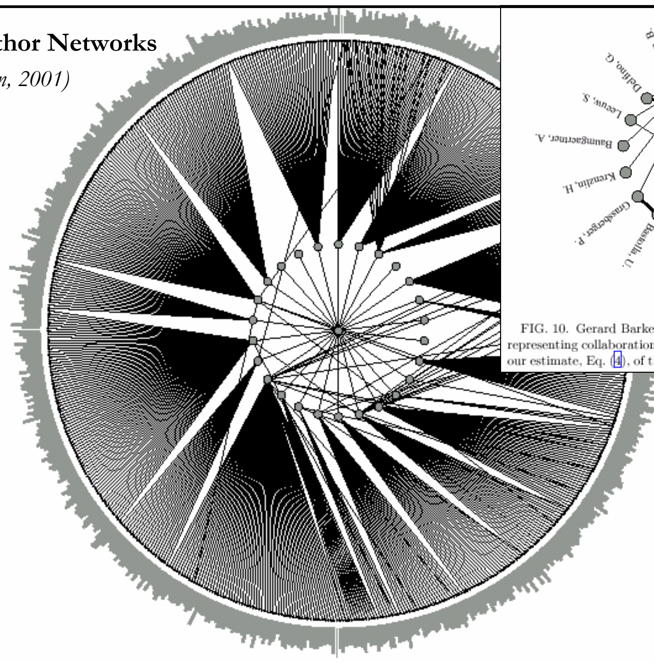
Visualizing Science by Citation Mapping

(Small, 1999)



Co-author Networks

(Newman, 2001)



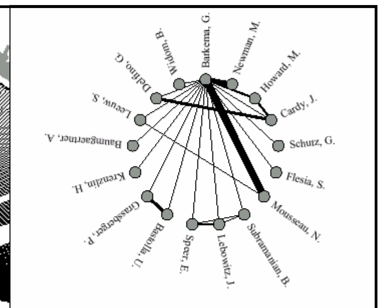


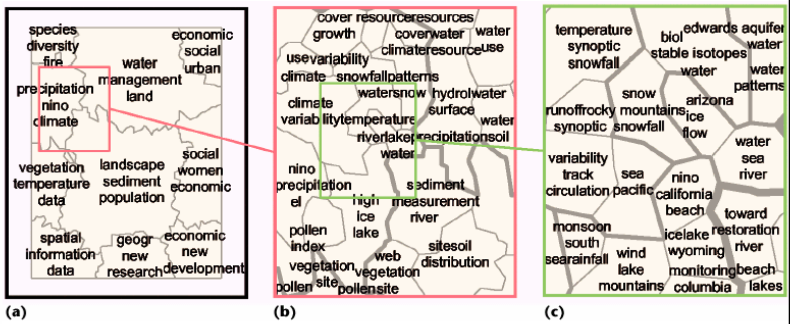
FIG. 10. Gerard Barkema and his collaborators, with lines representing collaborations whose thickness is proportional to our estimate, Eq. (4), of the strength of the corresponding tie.

FIG. 6. The point in the center of the figure represents the author of the paper you are reading, the first ring his collaborators, and the second ring their collaborators. Collaborative ties between members of the same ring, of which there are many, have been omitted from the figure for clarity.

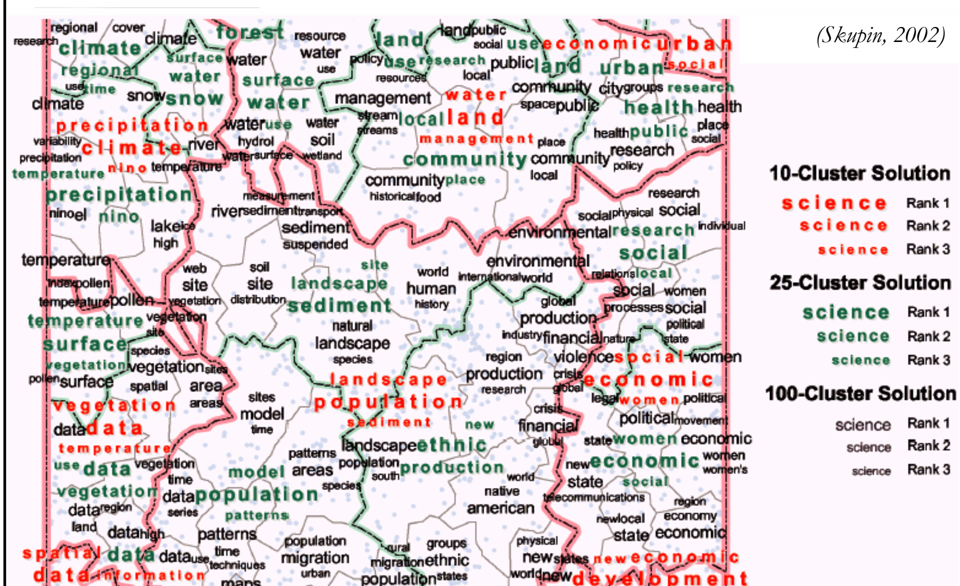
Cartographic Information Visualization

(Skupin, 2002)

6 Three different zoom levels in a visualization of conference abstracts: (a) complete map shown in a 10-cluster solution and map portions for (b) a 100-cluster and (c) 800-cluster solution. Higher level boundaries are accentuated to provide context during zoom operations.



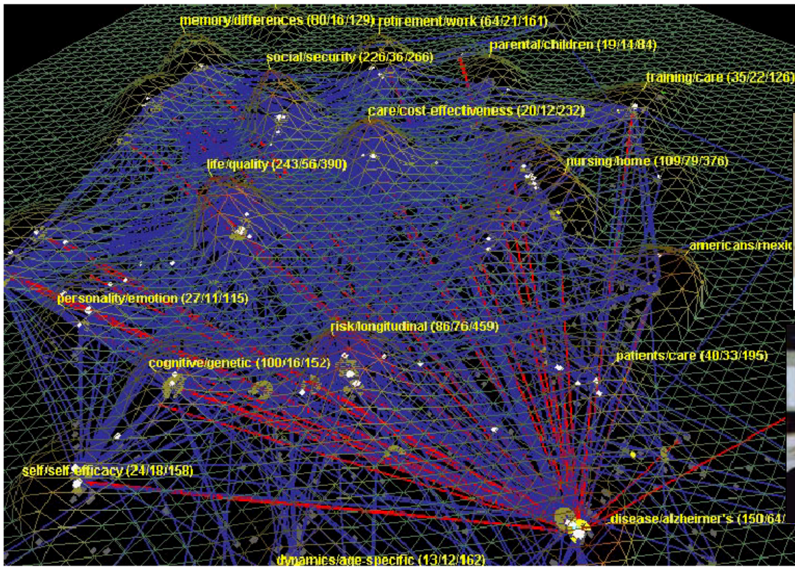
(Skupin, 2002)



7 Visualization of 2,220 conference abstracts with simultaneous overlay of three levels of a hierarchical clustering tree of SOM neurons: 10-cluster solution (red); 25-cluster solution (green); and 100-cluster solution (black). Cluster labels are scaled according to rank within the respective cluster.

Indicator-Assisted Evaluation and Funding of Research

Visualizing the influence of grants on the number and citation counts of research papers (Boyack & Börner, 2003)



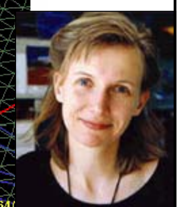


Figure 7. Author supplied linkage patterns from grants to publications with links highlighted in red for grant 01 P50 AG11715-01.

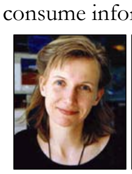
Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions

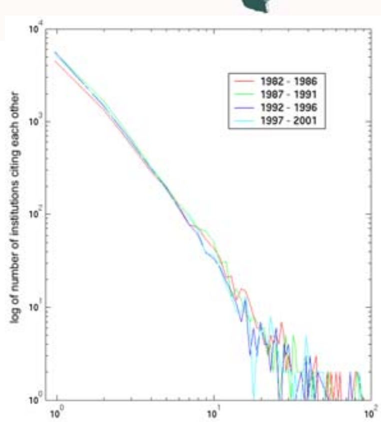
Börner & Penumarthy. (2005) Scientometrics Conference.

Does Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research instructions?

Analysis of top 500 most highly cited U.S. institutions.

Each institution is assumed to produce and consume information.

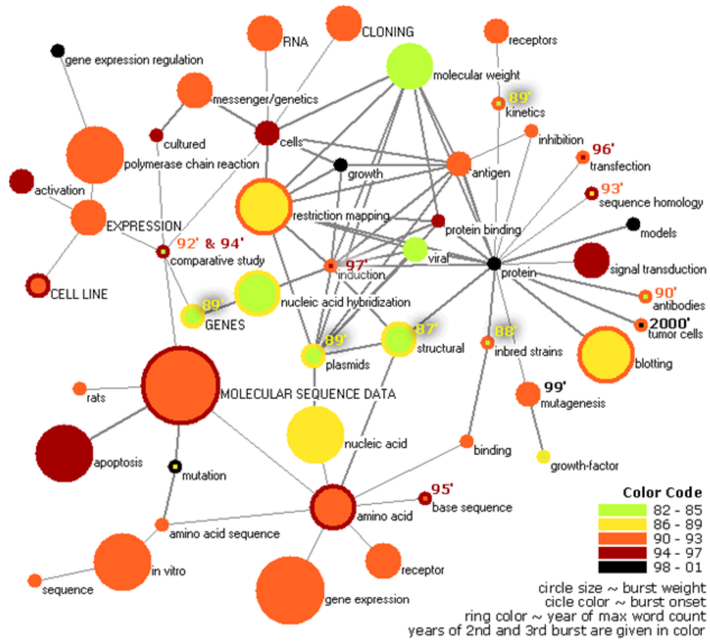


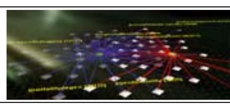


$\gamma_{82\text{-}86} = 1.94$ (R^2=91.5%)
$\gamma_{87\text{-}91} = 2.11$ (R^2=93.5%)
$\gamma_{92\text{-}96} = 2.01$ (R^2=90.8%)
$\gamma_{97\text{-}01} = 2.01$ (R^2=90.7%)

Mapping Topic Bursts in PNAS
(*Mane & Börner, 2004)*

Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.





a) Descriptive Models vs. Process Models

Descriptive Models

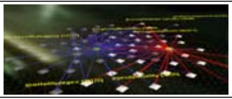
Aim to describe the major features of a (typically static) data set, e.g., statistical patterns of article citation counts, networks of citations, individual differences in citation practice, the composition of knowledge domains, and the identification of research fronts as indicated by new but highly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



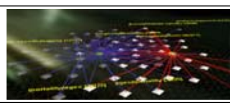
Process Models

Can be used to predict the effects of

- Different publishing mechanisms, e.g., E-journals vs. books on co-authorship, speed of publication, etc.
- Large collaborations vs. single author research on information diffusion.
- Interdisciplinary collaboration on the amount of duplication or the quality of (deep) science.
- Many small vs. few large grant on # publications, Ph.D. students, etc.
- …

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!

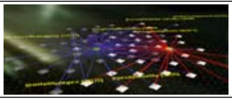


Process Models

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/nodes.

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).

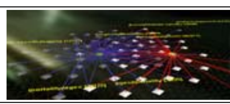


Models for Evolving Networks

Excellent Review Articles

- Albert & Barabási (2002). Statistical mechanics of complex networks.
- Dorogovtsev, S. N., & Mendes, J. F. F. (2002). Evolution of networks.
- Newman, M. E. J. (2001). Scientific collaboration networks. I. Network construction and fundamental results.
- Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality.

Scale Free Networks are typically simulated by processes of *incremental growth, rewiring, and preferential attachment.*



Preferential Attachment

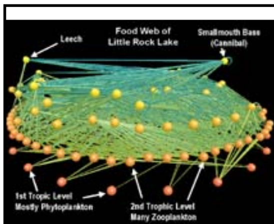
Preferential attachment supports a *rich get richer* phenomenon also known as the *Mathew effect* (Merton, 1973), or *cumulative advantage* (Price, 1976).

- Papers with many citations will attract even more citations …
- Authors with many co-authors will attract even more co-authors …
- Authors with many papers will produce even more papers …

Preferential attachment models link (new) papers/authors to highly connected (cited) papers/authors.

But, even experts in a field do not have an overview of the connectivity of papers and/or authors. Each author interacts directly only with a rather limited number of other authors and papers and makes local decisions based on his/her position in the author-paper network.

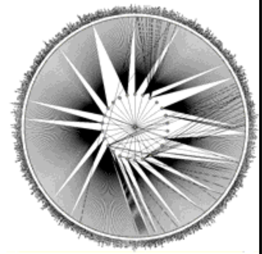
Katy Börner, Jeegar T. Maru & Robert L. Goldstone: Modeling the Co-Evolution of Scholarly Networks.



b) Isolated Networks vs. Network Ecologies

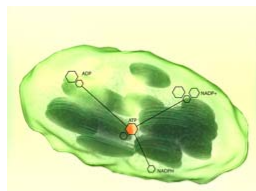
Food Web
(Williams & Martinez, 2000)

Internet
(Faloutsos, 1999)

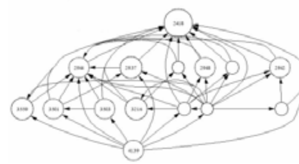


Co-Authorship
(Newman, 2001)

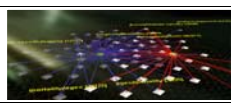
Sex Web
(Liljeros et al., 200



Metabolic Networks
(Jeong, et al., 2000)

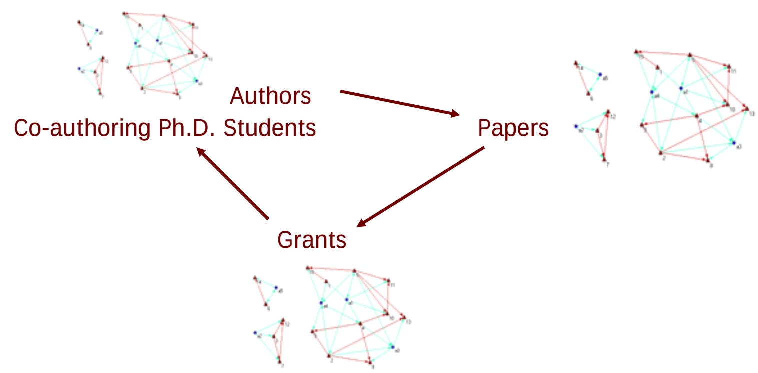


Paper Citations
(Redner, 1998)



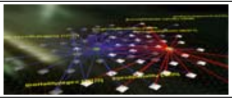
Network Ecologies

Most real world networks exist within a delicate ecology of networks.



To fully understand, e.g., the knowledge diffusion among authors via their papers, different networks need to be considered simultaneously. Here we will concentrate on co-author and paper-citation networks.

Katy Börner, Jeegar T. Maru & Robert L. Goldstone: Modeling the Co-Evolution of Scholarly Networks.



c) The Influence of Time and Semantics

Aging

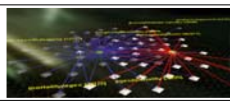
is an antagonistic force to preferential attachment. Even highly connected nodes typically stop receiving links after time has passed.

Aging cluster papers and authors temporally.

Topics

By dividing science into separate fields, the global rich-get-richer effect is broken down into many local rich-get-richer effects, leading to a more egalitarian distribution of received citations.

Topics cluster papers and authors semantically.



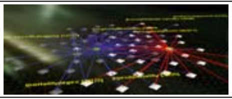
2. The TARL Model (Topics, Aging, and Recursive Linking)

The TARL model 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.

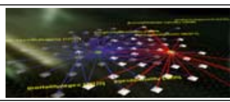
Börner, Katy, Maru, and Jeegar Goldstone, Robert. (2004) The Simultaneous Evolution of Author and Paper Networks. PNAS, 101(Suppl_1):5266-5273. Also available as cond-mat/0311459.



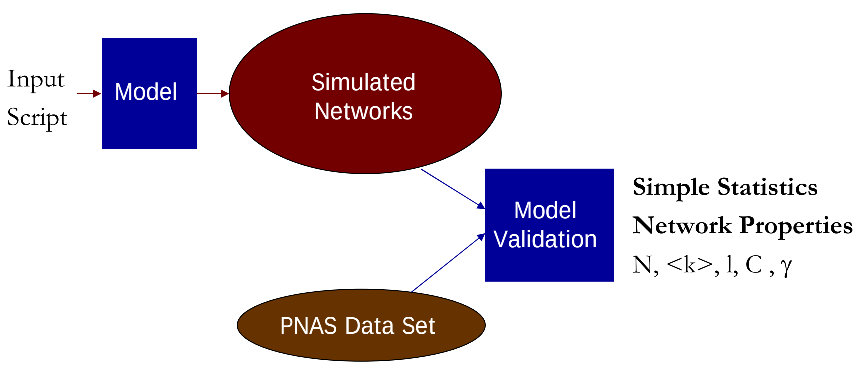
The TARL Model: Basic Assumptions

- 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.

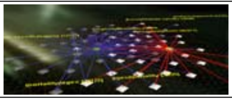
- Preferential attachment is modeled as an emergent property of the elementary, local networking activity of authors reading and citing papers, but also the references listed in papers. Analogously, authors may consider collaborating with co-authors of their co-authors, linking to web pages linked from web pages they read, etc.



The TARL Model: Run & Validation



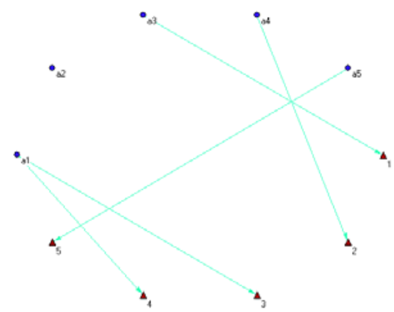
Subsequently, I will give an intuitive explanation of the modeling process, an explanation for engineers, one for computer scientists, as well as formulas for math/physics folks.



The TARL Model: Initialization

Input Script

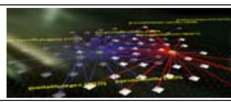
- Parameters (topics, co-authors, reference path length, aging function)
- # papers, #authors, # topics, aging function
- # years, papers consumed (referenced) per paper, # papers produced per author each year, # co-author(s) per author, # levels references are considered, age of authors, the number of their active years, and the increase in the number of authors over the years.



Example:

5 authors, 5 papers, no topics

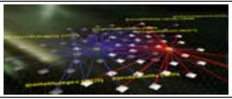
Each paper has a randomly selected set of authors but no references.



The TARL Model: Sample Input Script File

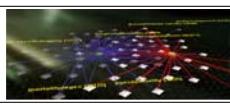
```
------------------------------------------
Model Parameters (0=without, 1=with)
------------------------------------------
0  Topics
0  Co-Authors
0  Consider References
0  Aging Function
------------------------------------------
Model Initialization Values
------------------------------------------
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
```

Not shown are parameters that define the age of authors, the number of their active years, and the increase in the number of authors over the years.



The TARL Model: Pseudo Code

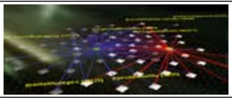
```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate 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;
}
```

The TARL Model: Sample Input Script File

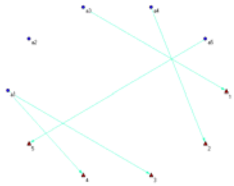
```
------------------------------------------
Model Parameters (0=without, 1=with)
------------------------------------------
0  Topics
0  Co-Authors
0  Consider References
0  Aging Function
------------------------------------------
Model Initialization Values
------------------------------------------
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
```

Katy Börner, Jeegar T. Maru & Robert L. Goldstone: Modeling the Co-Evolution of Scholarly Networks.

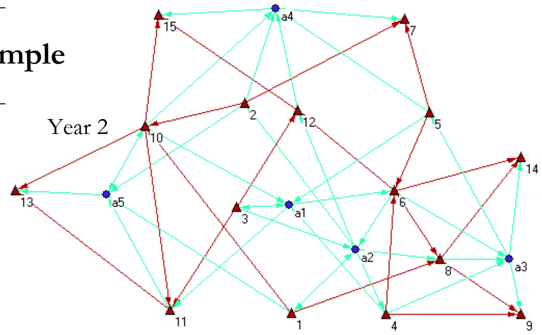


Example

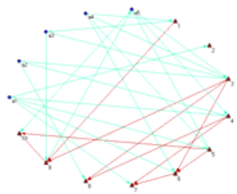
Year 0 - Initialization



Year 2



Year 1



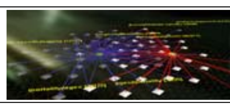
Initial setup, first year, and second year topology of a simple author-paper network.

Authors a1, a2 ,… are represented by blue circles
Papers 1, 2, … are denoted by red triangles

Red arrows indicate the information flow (via citation links) from older papers to more recent papers.

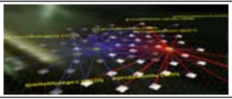
Green arrows denote **consumed** and **produced** paper-author relationships.

Arrows denote flow of information.

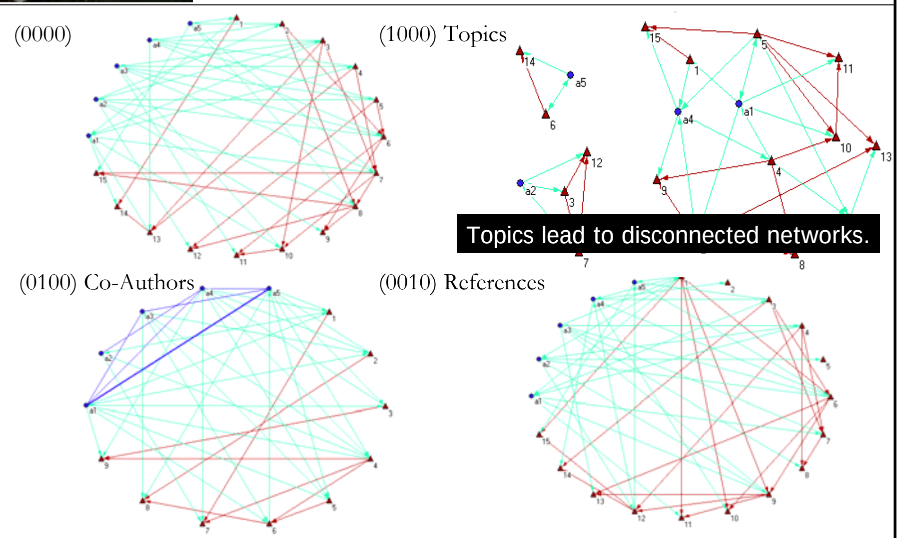


The TARL Model: Sample Input Script File

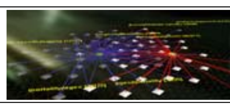
```
-------------------------------------------
Model Parameters (0=without, 1=with)
-------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
-------------------------------------------
Model Initialization Values
-------------------------------------------
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
```



The TARL Model: The Effect of Parameters

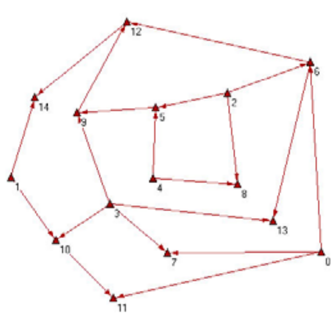


Co-authoring leads to fewer papers.

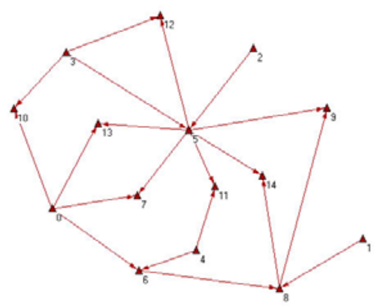


The TARL Model: Reading References

Init + 2 year paper citation networks



without considering references (0000)



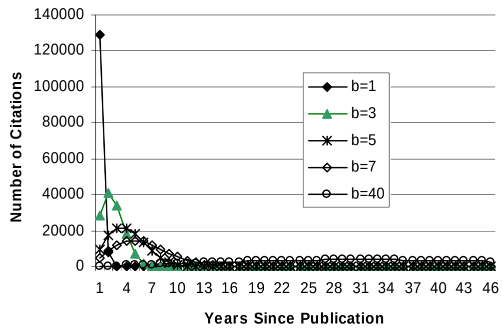
with reading references (0010)

The TARL Model: Aging Function

Citation probability functions observed in paper citation networks can be fit by a Weibull distribution of the form

$$W(t) = cab^{-a} t^{(a-1)} e^{-\left(\frac{t}{b}\right)^{a}}$$

where *c* is a scaling factor, *a* controls the variability of distribution, and *b* controls the rightward extension of the curve.



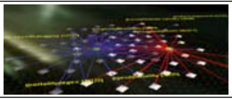
The TARL Model: Probability of Receiving Citations

If references as well as aging are considered, then the probability of Paper *y* being cited, *P(y),* corresponds to the normalized sum of the aging dependent probability for each of its tokens,

$$P(y) = \frac{\sum_{t=1}^{n} \sum_{i \in P_{r,t} \wedge i=y} W(t)}{\sum_{t=1}^{n} \sum_{i \in P_{r,t}} W(t)}$$

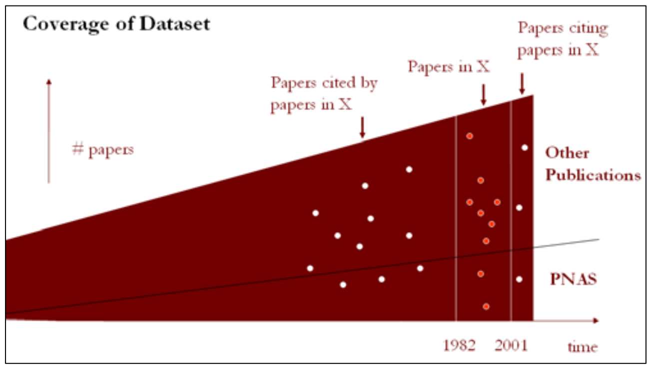
where n is the total number of years considered.

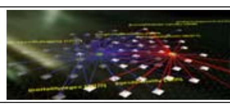
Hence a paper that was published in Year y and received 4 citations in Year y+1 and 2 citations in Year y+2 has 7 tokens that are weighted by the probability value for each year.



The TARL 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.





The TARL Model: Validation

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**	

Young papers did not garner many citations yet.

Katy Börner, Jeegar T. Maru & Robert L. Goldstone: Modeling the Co-Evolution of Scholarly Networks.

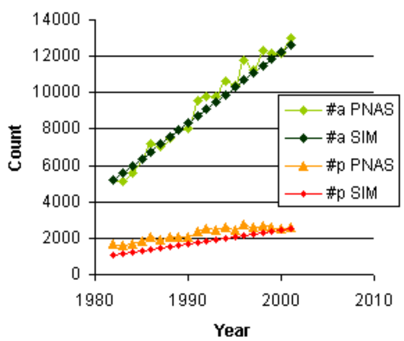
PNAS Simulation Input Script File

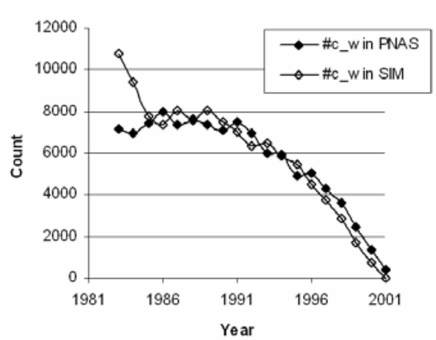
```
-------------------------------------------
Model Parameters (0=without, 1=with)
-------------------------------------------
0   Topics
1   Co-Authors
1   Consider References
1   Aging Function (Weibull with b=3)
-------------------------------------------
Model Initialization Values
-------------------------------------------
21     # Years
4809   # Authors in Start Year
1624   # Papers in Start Year
392    # Additional Authors per Year
3      # Papers Referenced per Paper
1      # Papers Produced per Author each Year
4      # Co-Authors
1      # Levels References are Considered
```

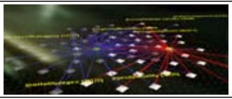
First year is used for initialization

Comparison PNAS & SIM

Total number of papers (#p), authors (#a), received citations (#c) and references (#r) for years 1982 through 2001.







Network Properties

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.

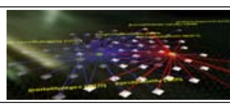
Network	n	<k>	l	C	γ	Reference
Co-authorship networks						
LANL	52,909	9.7	5.9	0.43	--	Newman, (2001a; 2001b; 2001c)
MEDLINE	1,520,251	18.1	4.6	0.066	--	
SPIRES	56,627	1.73	4.0	0.726	1.2	
NCSTRL	11,994	3.59	9.7	0.496	--	
Math.	70,975	3.9	9.5	0.59	2.5	Barabasi et al., (2002)
Neurosci.	209,293	11.5	6	0.76	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						

Source:

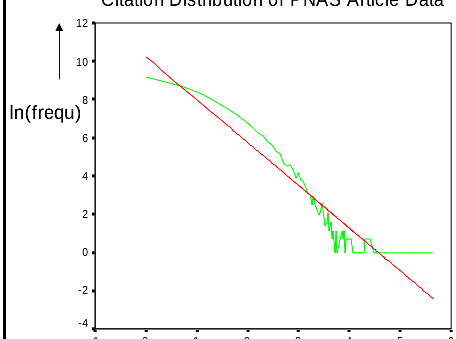
Albert, R., & Barabá

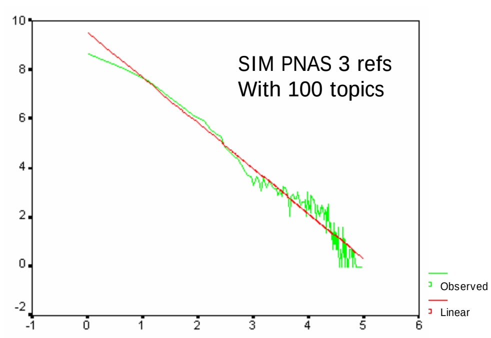
Reviews of Modern Phy

For undirected co-author networks, the in-degree of a node equals its out-degree and hence the exponents for both distributions are identical. For directed paper citation networks, the number of references is rather small and constant. Only the in-degree distribution (received citations) are considered.



Power Law Distribution Exponents





Rsq	d.f.	F	Sigf	b0
.877	70	497.88	.000	10.2251

If topics are considered, the distribution shows the same systematic deviations from a power law as observed for PNAS article data set: The least-cited and most-cited papers are cited less often than predicted by a power-law, and the moderately-cited papers are cited more often than predicted.

Power Law with Exponential Cutoff provides a better fit

Networks in which aging occurs, e.g., actor networks or friendship networks, show a connectivity distribution that has a power law regime followed by an exponential or Gaussian decay or have an exponential or Gaussian connectivity distribution (Amaral et al., 2000). Newman showed that connectivity distributions of co-author networks can be fitted by a power-law form with an exponential cutoff (Newman, 2001c).

Following this lead, we fit a power law with exponential cutoff of the form

$$f(x) = Ax^{-B}e^{\frac{x}{C}}$$

This function provided an excellent fit to the PNAS paper citation network with values of A=13,652, B= .49, and C=4.21 (R^2=1.00).

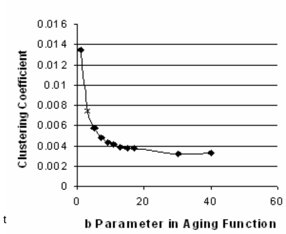
The TARL Model: Influence of Parameters

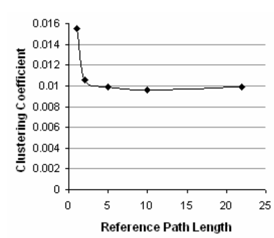
Topics: The number of topics is linearly correlated with the clustering coefficient of the resulting network: C= 0.000073 * #topics. Increasing the number of topics increases the power law exponent as authors are now restricted to cite papers in their own topics area.

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.





3. Discussion & Future Work

- TARL model grows author and paper networks simultaneously.
- Uses the reading and citing of paper references as a grounded mechanism to simulate preferential attachment.
- The number of topics is linearly correlated with the clustering coefficient of the resulting network and can be determined from the cluster coefficient observed in real world networks.
- The model incorporates aging, i.e., a bias for authors to cite recent papers and hence papers are not only clustered by topic, but also in time.

For the sake of simplicity we fixed the *number of papers* produced by each author per year and fixed the *number of co-authors*. To model the rich get richer effect for co-author networks, *recursive linking* can be applied so that authors co-author with co-authors of their co-authors.

Clearly, further validation of the model with <u>different parameter settings</u> and <u>other data sets</u> is necessary.

Modeling and Studying Feedback Cycles in Network Ecologies

The productivity of an author may depend not only from his/her position in the author-paper network but also on research funds, facilities, and students. Hence, grant support will be modeled as a third network to capture the positive feedback cycle observed in real world network ecologies.

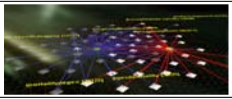
Network Structure & Network Usage

In how far does the usage of a network (e.g., via search engines) influence its structure?

Visualizing the Evolution of Scientific Fields and Knowledge Diffusion

How to map the structure and evolution of all of science.

Katy Börner, Jeegar T. Maru & Robert L. Goldstone: Modeling the Co-Evolution of Scholarly Networks.



Acknowledgements

This work greatly benefited from discussions with and comments from Kevin Boyack, Albert-László Barabási, Mark Newman, Olaf Sporns, Filippo Menczer, and the anonymous reviewers. Mark Newman made code available to determine the small world properties of networks. Nidhi Sobti was involved in the analysis of the influence of model parameter values. Batagelj & Mrvar's Pajek program was used to generate the network layouts.

This work is supported by a National Science Foundation CAREER Grant under IIS-0238261 to the first author, and a National Science Foundation grant 0125287 to the third author.