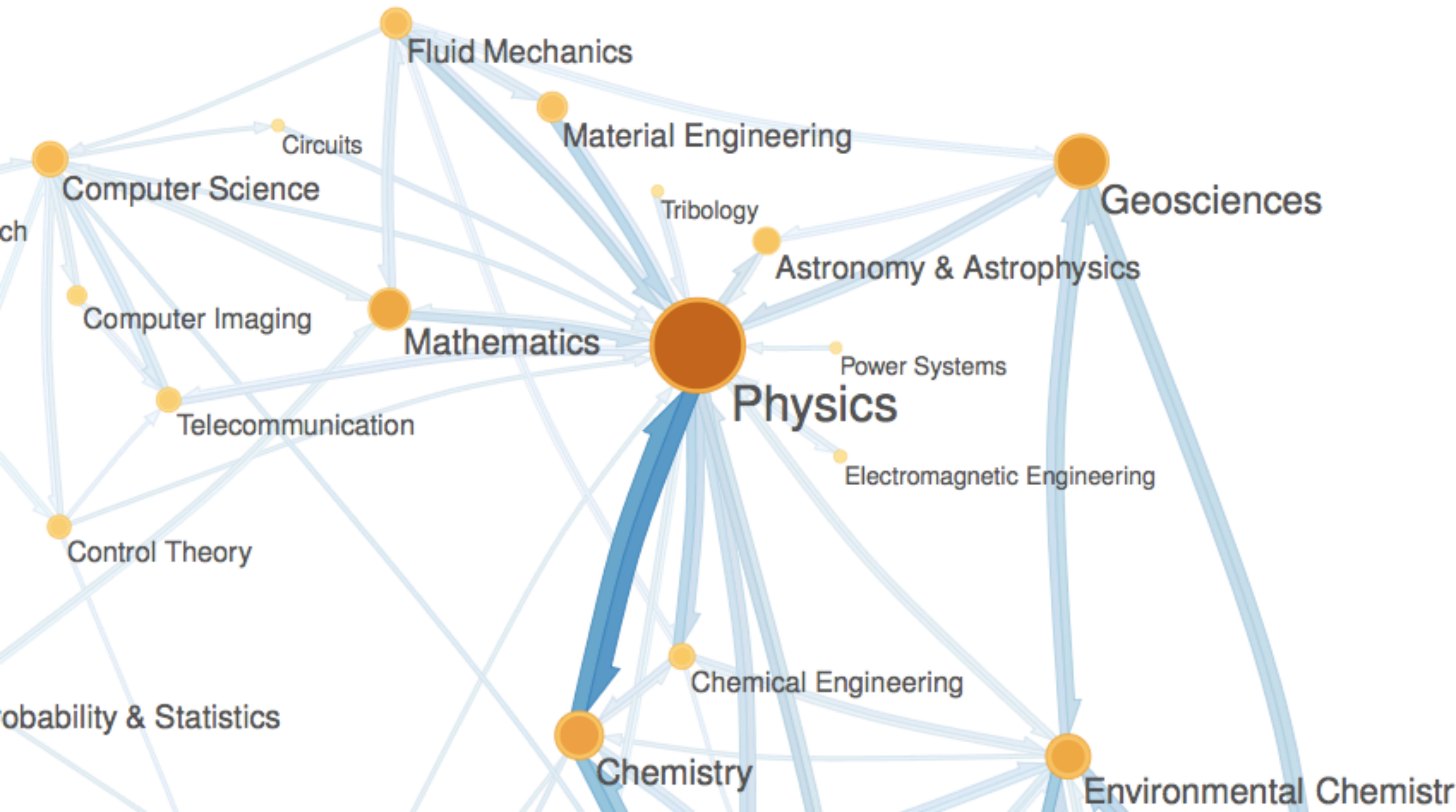
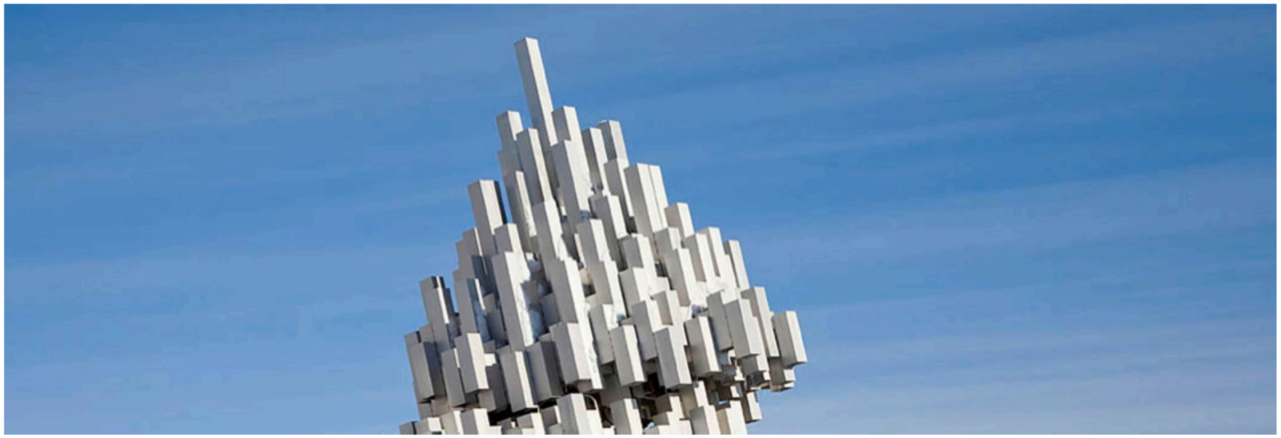
Facilitating discovery with zoomable maps

Jevin West, Information School, University of Washington



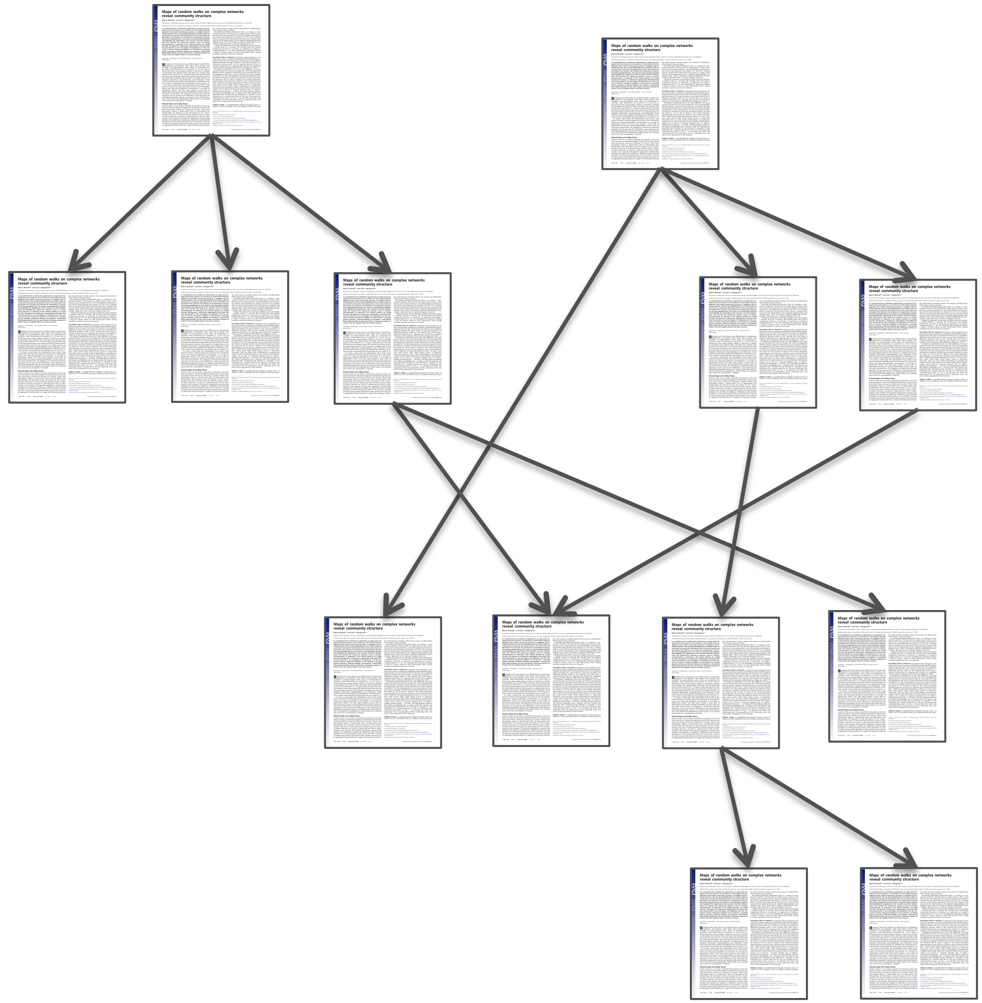
compressing ⟷ finding patterns

5.8 MB (TIFF) ⟶ 0.9 MB (TIFF + LZW)



5.8 MB (TIFF) ⟶ 2.8 MB (TIFF + LZW)

Citations form a vast network



de Solla Price, *Science* (1965)

The Scholarly Graph

NSF

JSTOR

arXiv.org

SSRN

Microsoft Academic Search Beta

ads NASA

THOMSON REUTERS

ProQuest Start here.

PatentVector™

WIKIPEDIA The Free Encyclopedia

PLOS

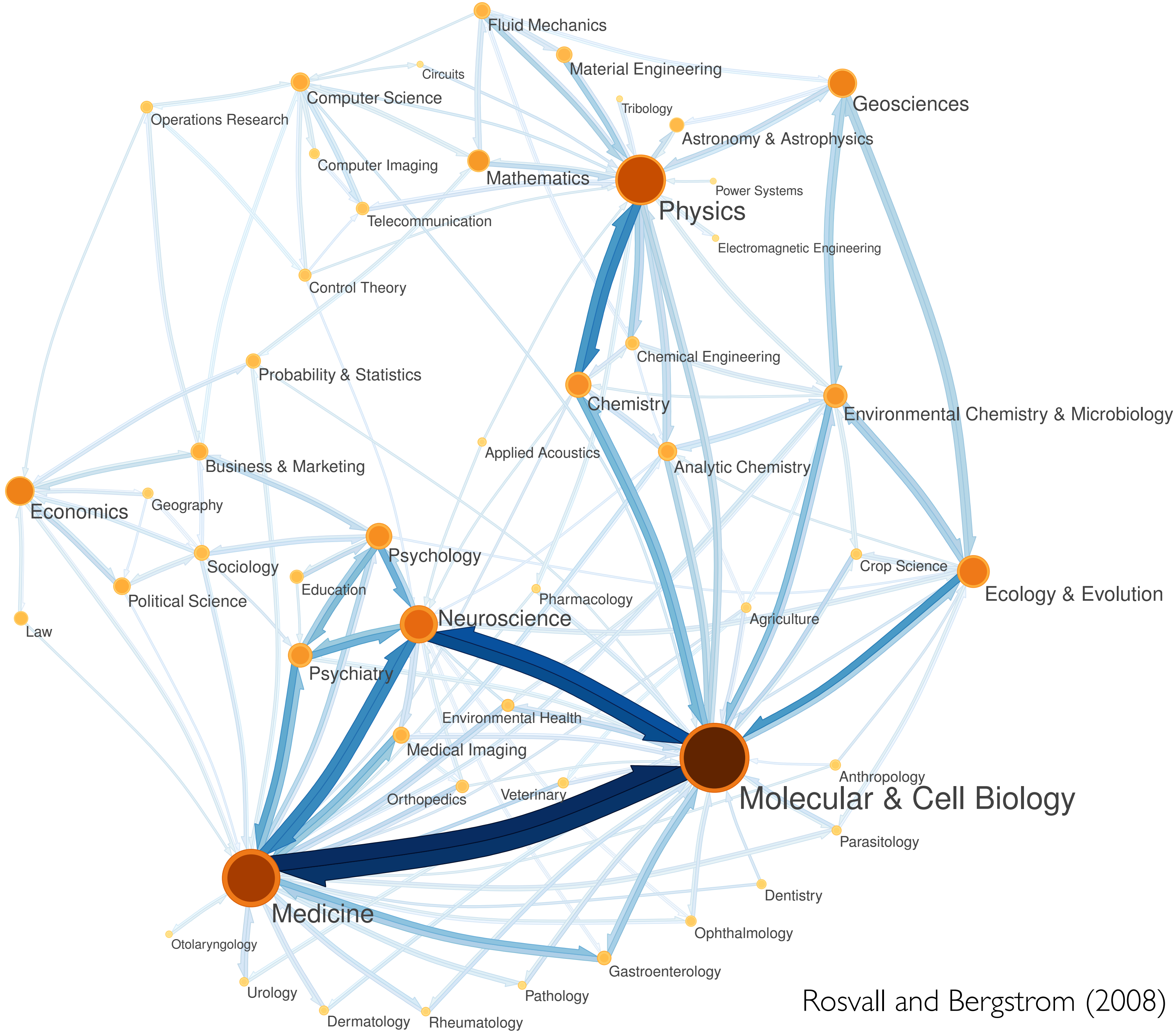
PNAS

IEEE

ELSEVIER

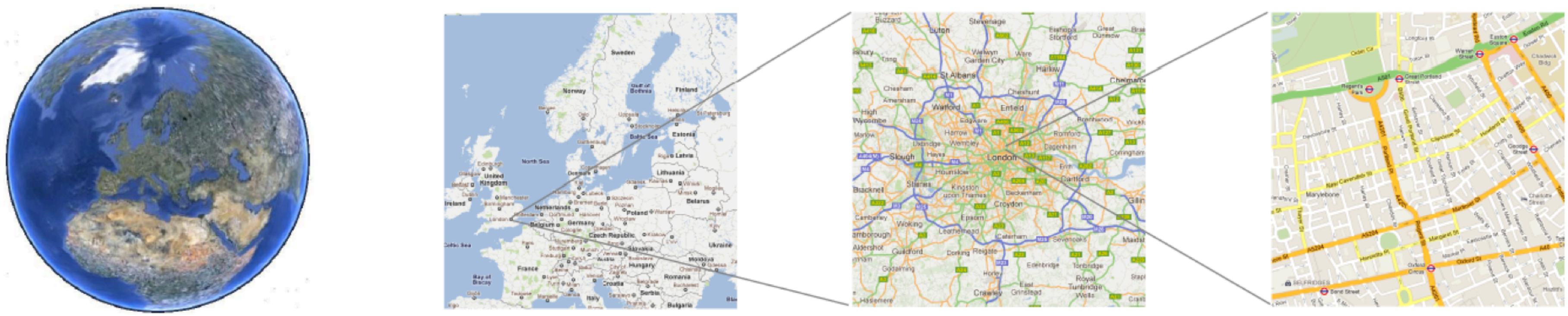
dblp computer science bibliography

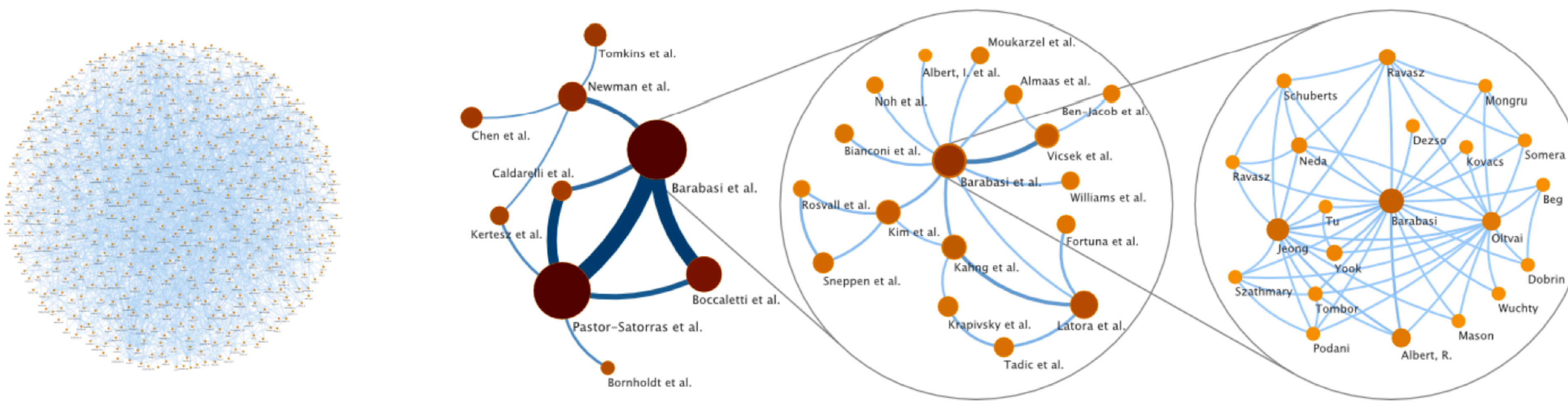
PubMed

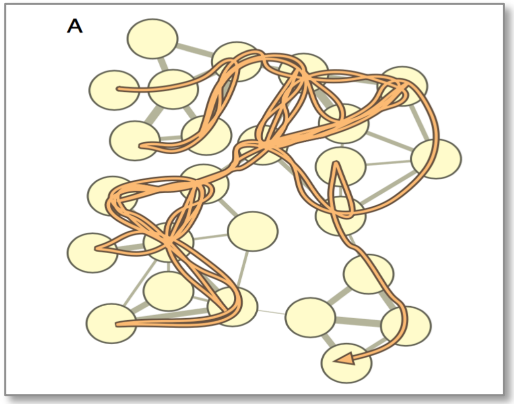


Rosvall and Bergstrom (2008) PNAS

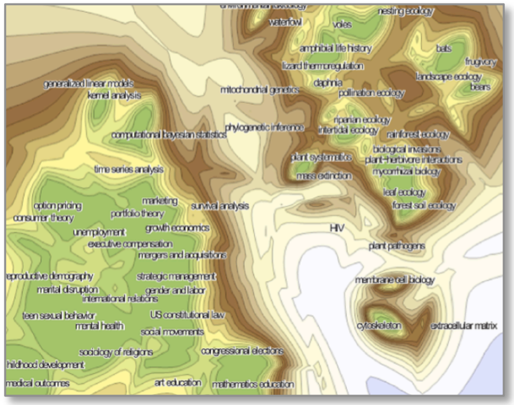
Facilitating discovery with zoomable maps







Science of Mapping



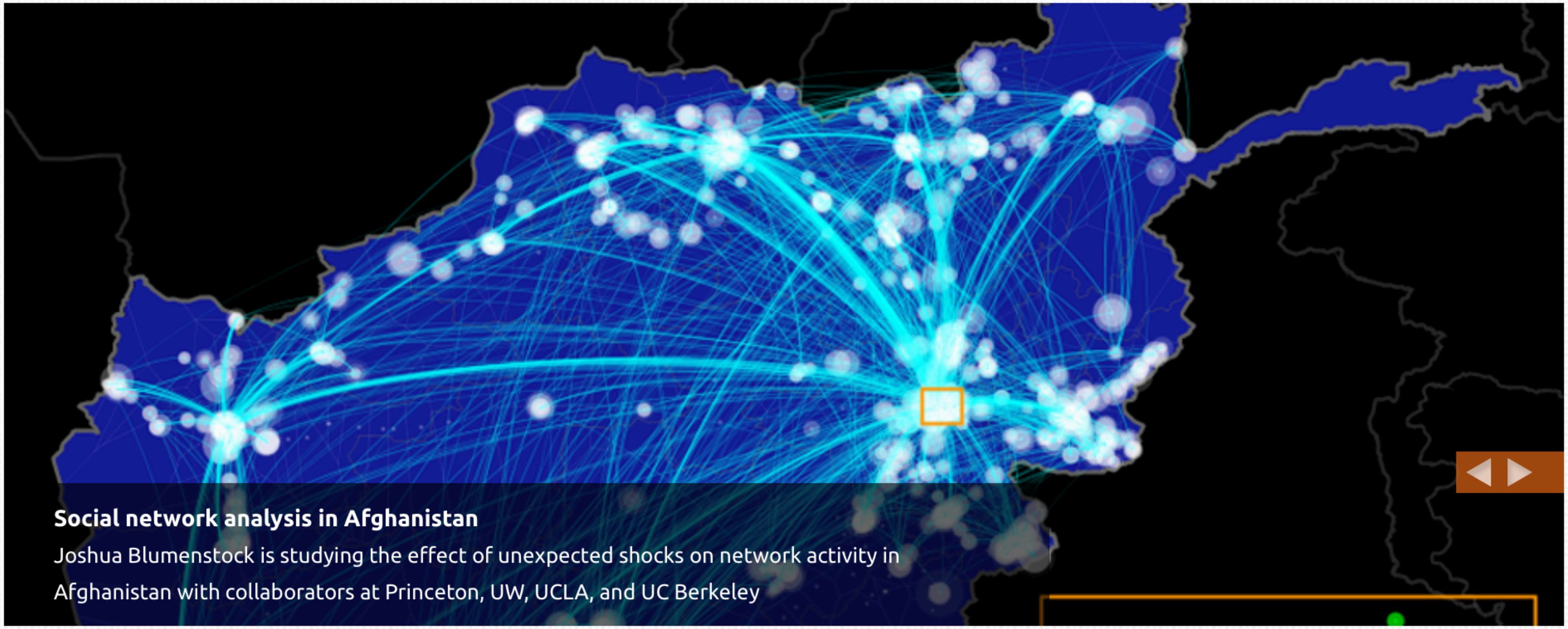
Mapping of Science

W DataLab
Data Science and Analytics Lab

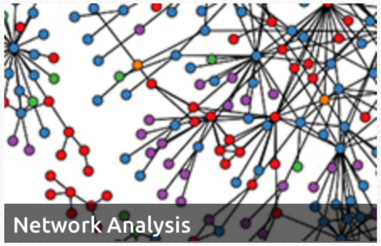
University of Washington Information School

Projects | Publications | People | About

Search



Research Focus Areas



News and Updates

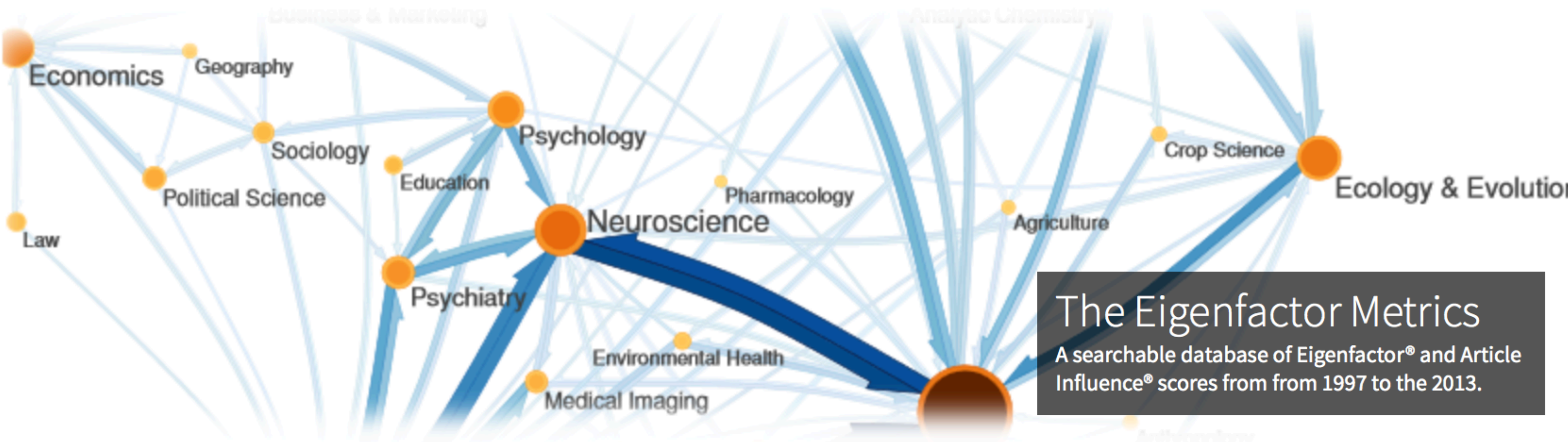
28 **Blumenstock at Population Association of America**

What we do

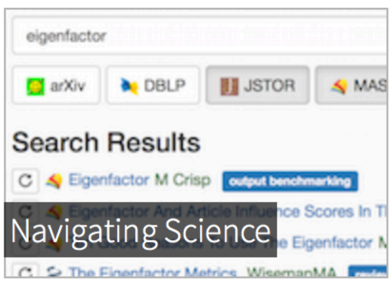
The DataLab is the nexus for research on Data Science and Analytics at the UW iSchool. We study **large-scale, heterogeneous human data** in an

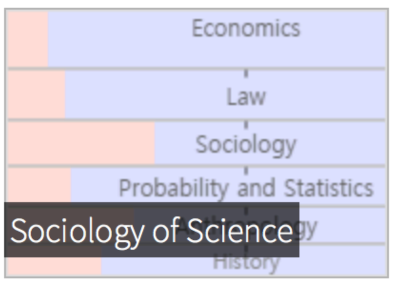
EIGENFACTOR.org

HOME | PROJECTS | PAPERS | ABOUT



RESEARCH AREAS





NEWS

23 Nov.

JEVIN WEST ON MEGAJOURNALS IN THE *CHRONICLE OF HIGHER EDUCATION*

Jevin West discusses the rise of the megajournal and our open access cost effectiveness tool in the *Chronicle of Higher Education*.

---

23 Nov.

EIGENFACTOR TEAM PLACES SECOND IN MICROSOFT RESEARCH'S WSDM CUP

The WSDM Cup Challenge asked teams to use 30GB of data from the Microsoft Academic Graph to rank the

Babel Home About API

babel.eigenfactor.org

oren etzioni

Ian Wesley-Smith

v | DBLP | JSTOR | MAS | PLOS | PubMed

s

al Methods For Analyzing Speedup Learning Experiments O Etzioni satisfaction programs

-Face And Computer-Mediated Communities Amitai Etzioni, Oren Etzioni 1998 resources sustained

cument Clustering O Zamir document clustering

Communities: Virtual Vs. Real A Etzioni 1996 implications internet

al Methods For Analyzing Speedup Learning Experiments. O Etzioni 1993 scheduling problems

al Methods For Analyzing Speedup Learning Experiments O Etzioni 1993 generating abstractions

Get Related Web Document Clustering: A Feasibility Demonstration O Zamir 1997 document clustering

Get Related Web Document Clustering: A Feasibility Demonstration. O Zamir 1997 browsing large

Get Related Sound And Efficient Closed- World Reasoning O Etzioni proving problem

Get Related Appears In Comm. OfACM O Etzioni scalable comparison-shopping

« Previous 1 2 3 4 5 6 7 8 9 10 Next »

Papers related to Statistical Methods For Analyzing Speedup Learning Experiments O Etzioni satisfaction programs

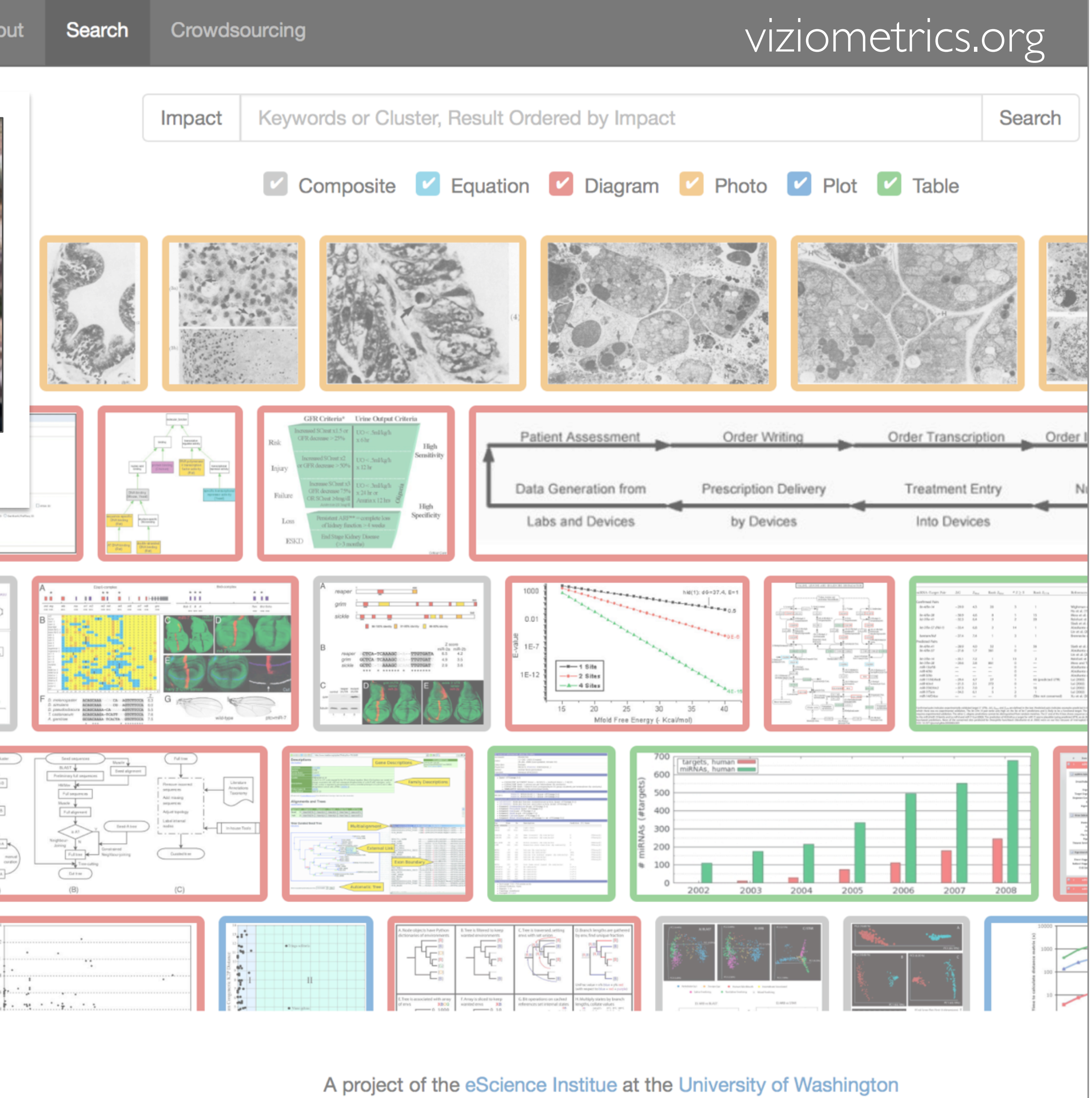
Get Related Automatically Configuring Constraint Satisfaction Programs: A Case Study S Minton 1995 satisfaction programs

Get Related Abstraction Via Approximate Symmetry T Ellman 1992 satisfaction programs

Get Related Integrating Heuristics For Constraint Satisfaction Problems: A Case Study S Minton 1992 satisfaction programs

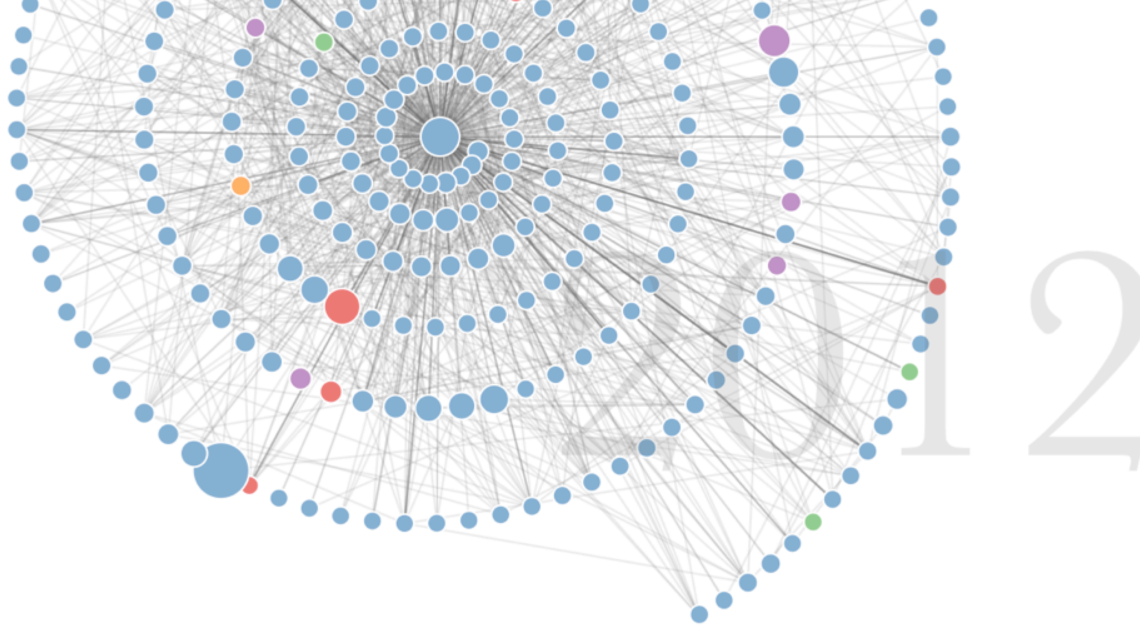
Get Related An Analytic Learning System For Specializing Heuristics S Minton 1992 satisfaction programs

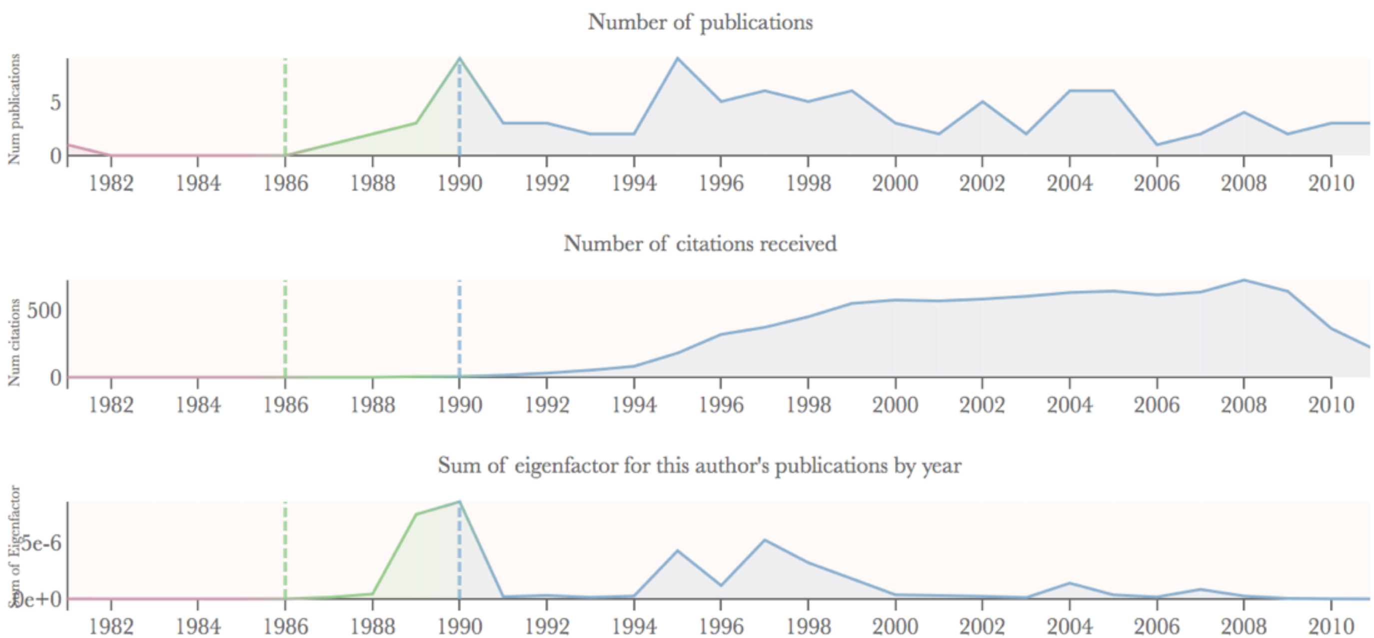
Get Related Automated Synthesis Of Constrained Generators W Braudaway 1988 satisfaction programs viziometrics.org



Poshen Lee

Jason Portenoy



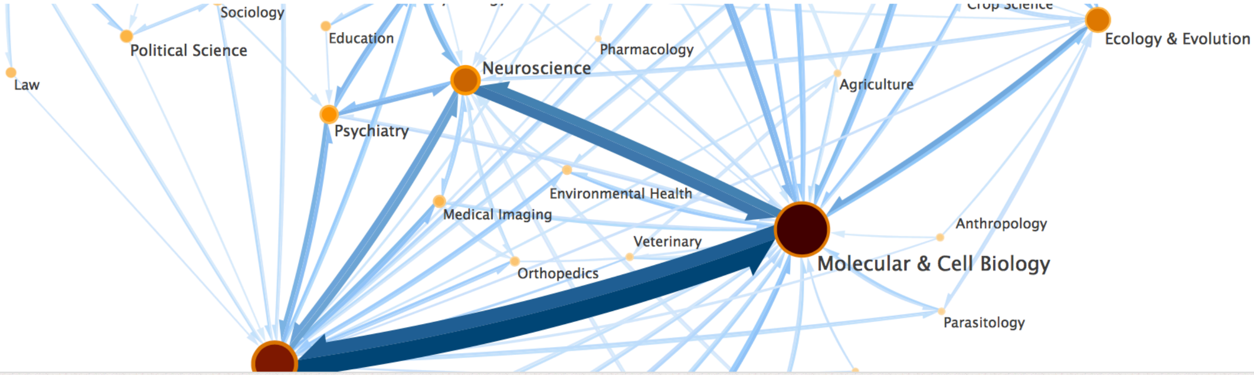


scholar.eigenfactor.org

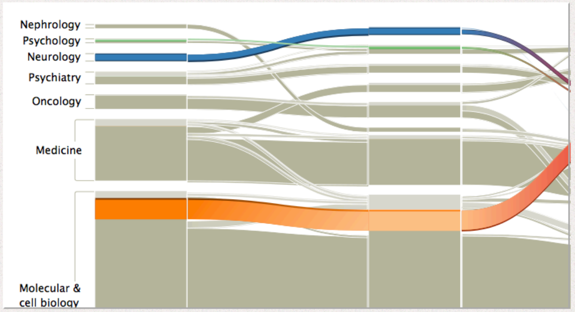
MapEquation

apps code publications about

Simplify and highlight important structures in complex networks



Apps »



Code »

```
using infomath::plogp;
for (unsigned int i = 0; i < numNodes; ++i)
{
enter_log_enter += plogp(m_moduleFlowData[i].enter
exit_log_exit += plogp(m_moduleFlowData[i].exitFlo
flow_log_flow += plogp(m_moduleFlowData[i].exitFlo
enterFlow += m_moduleFlowData[i].enterFlow;
}
enterFlow += exitNetworkFlow;
enterFlow_log_enterFlow = plogp(enterFlow);
```

Publications »

Maps of information flow reveal community structure in complex networks

Martin Rosvall and Carl T. Bergstrom

PNAS **105**, 1118 (2008). [arXiv:0707.0609]

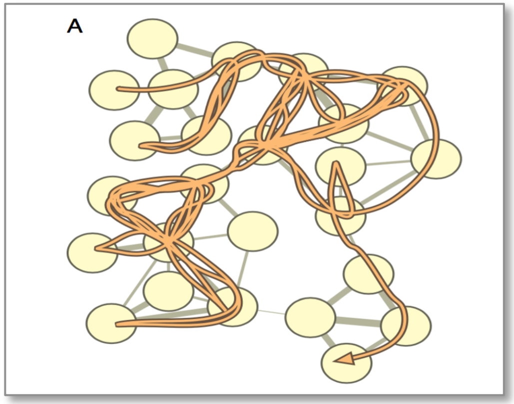
To comprehend the multipartite organization of large-scale biological and social systems, we introduce a new information-theoretic approach to reveal community structure in

News

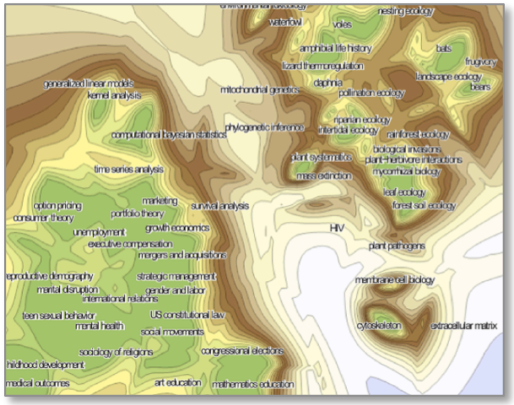
August 13, 2015 Interactive storyboard — Multilevel network sampling — infer network modes from multiple samples

August 13, 2015 Interactive storyboard — Higher-order Markov models — identify flows on memory and multilayer networks

July 23, 2015 Source code — Infomap — updates to memory and multilayer algorithms



Science of Mapping



Mapping of Science

Data

Compressing ⟷ Finding patterns

If we can find a good code for describing flow on a network, we will have solved the dual problem of finding the important structures with respect to that flow.

The map equation

frequency of inter-module movements

frequency of movements within module *i*

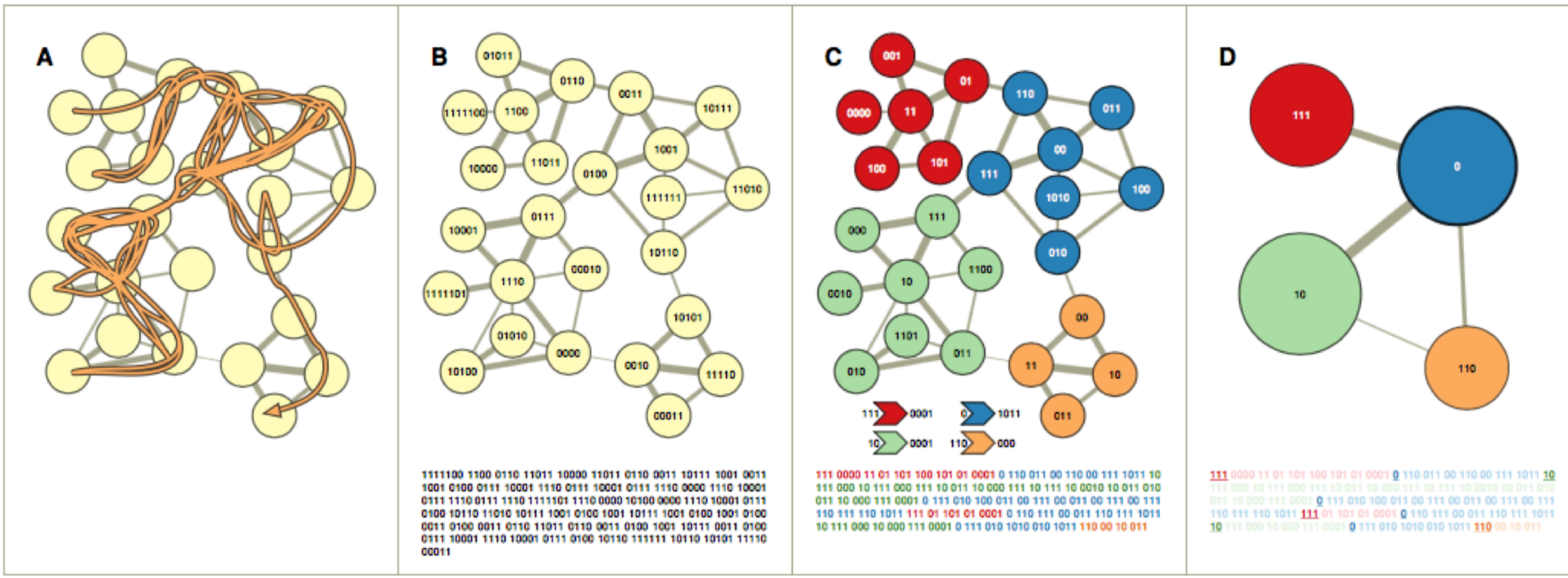
$$L(\mathsf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i)$$

code length of module names

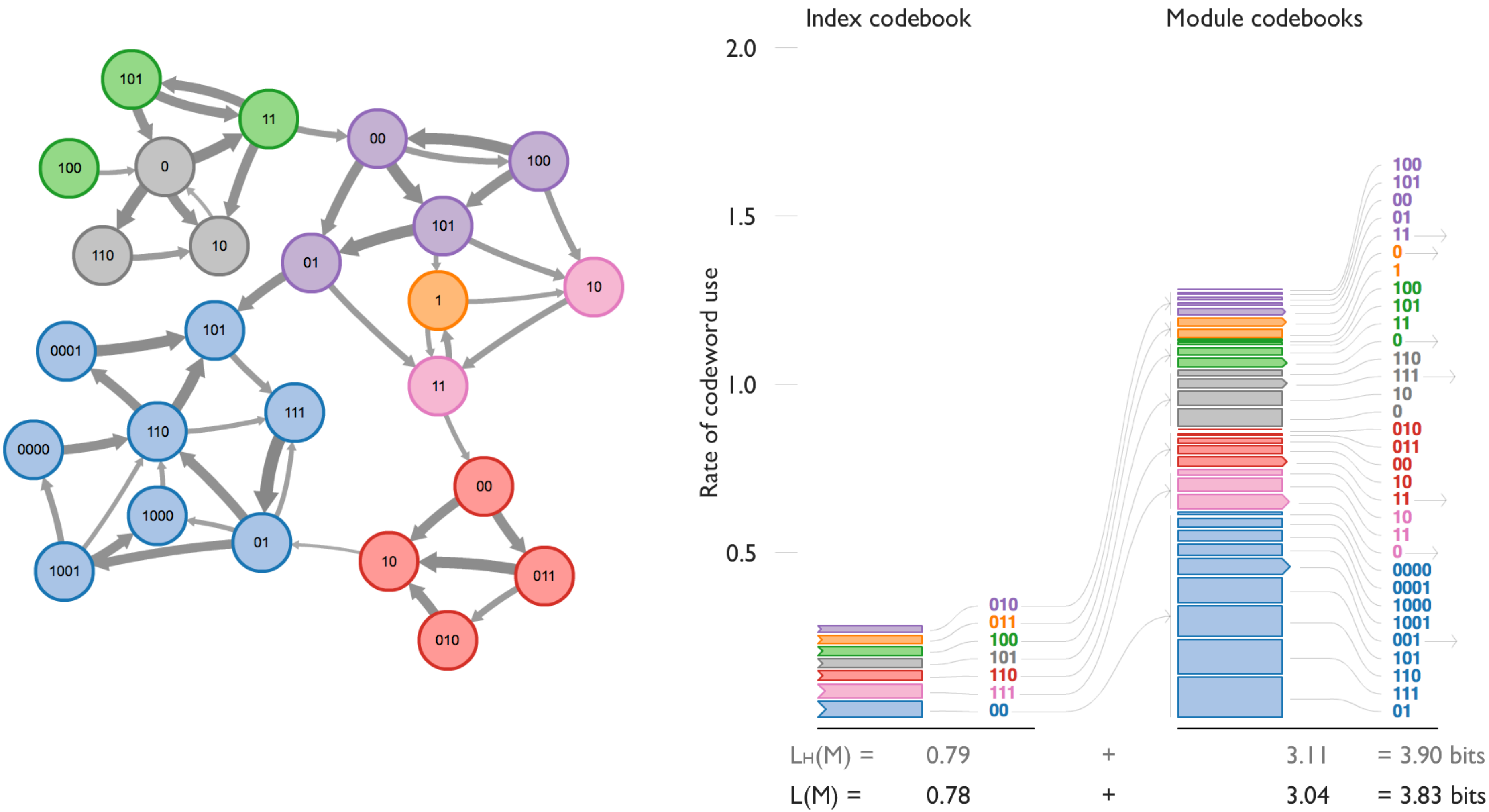
code length of node names in module *i*

Rosvall and Bergstrom (2008) *PNAS*

Finding regularities in the dynamics on networks



Rosvall and Bergstrom (2008) PNAS



Mapequation.org, Daniel Edler

Performance

PHYSICAL REVIEW E **80**, 056117 (2009)

Community detection algorithms: A comparative analysis

Andrea Lancichinetti[1,2] and Santo Fortunato[1]

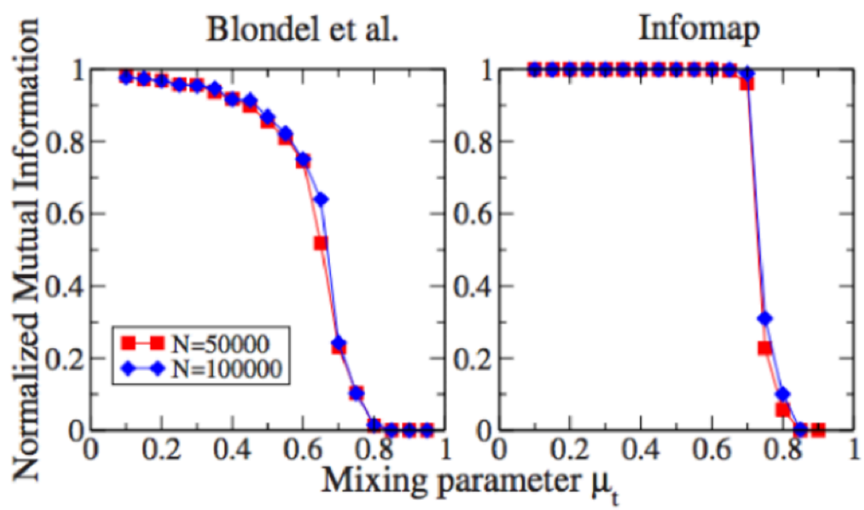
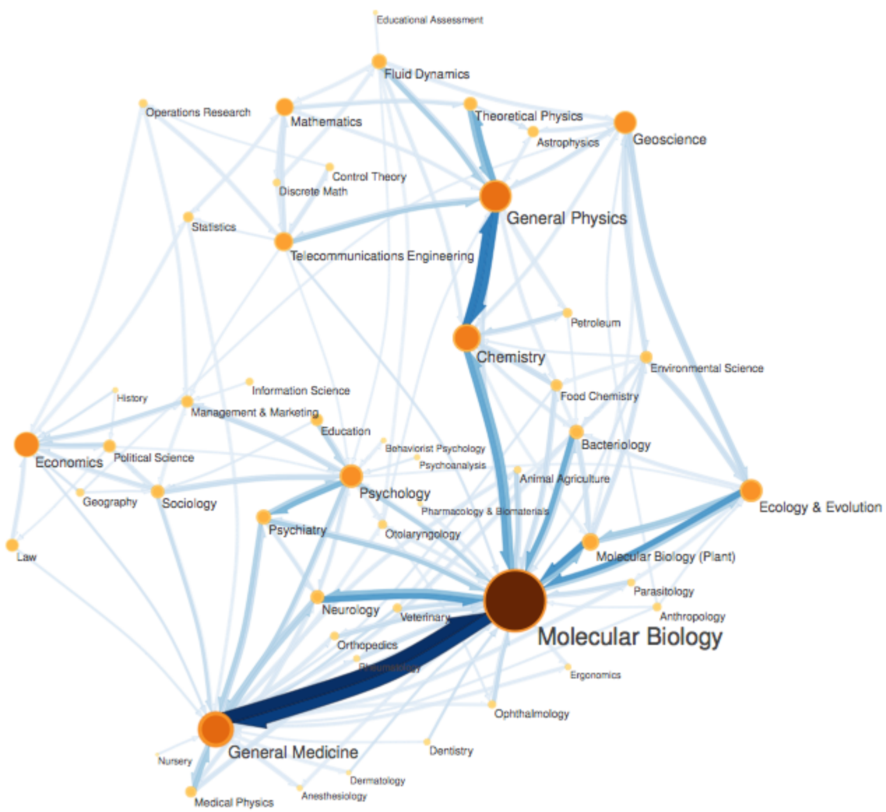


FIG. 3. (Color online) Tests of the algorithm by Blondel *et al.* and Infomap on large LFR benchmark graphs with undirected and unweighted links.

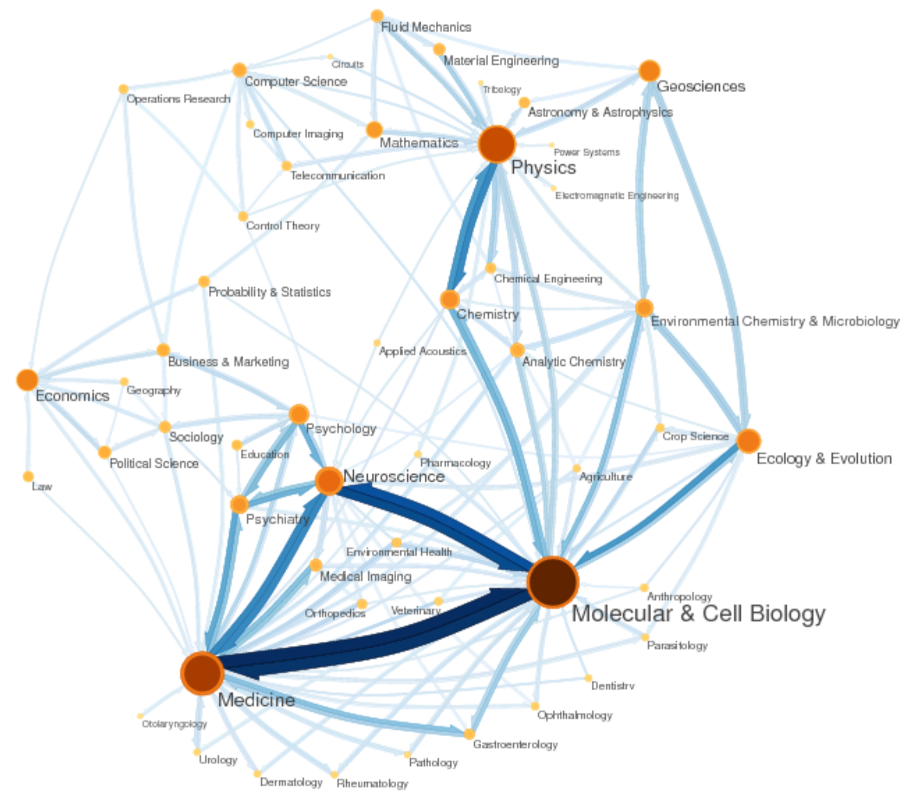
DOI: 10.1103/PhysRevE.80.056117

PACS number(s): 89.75.Hc

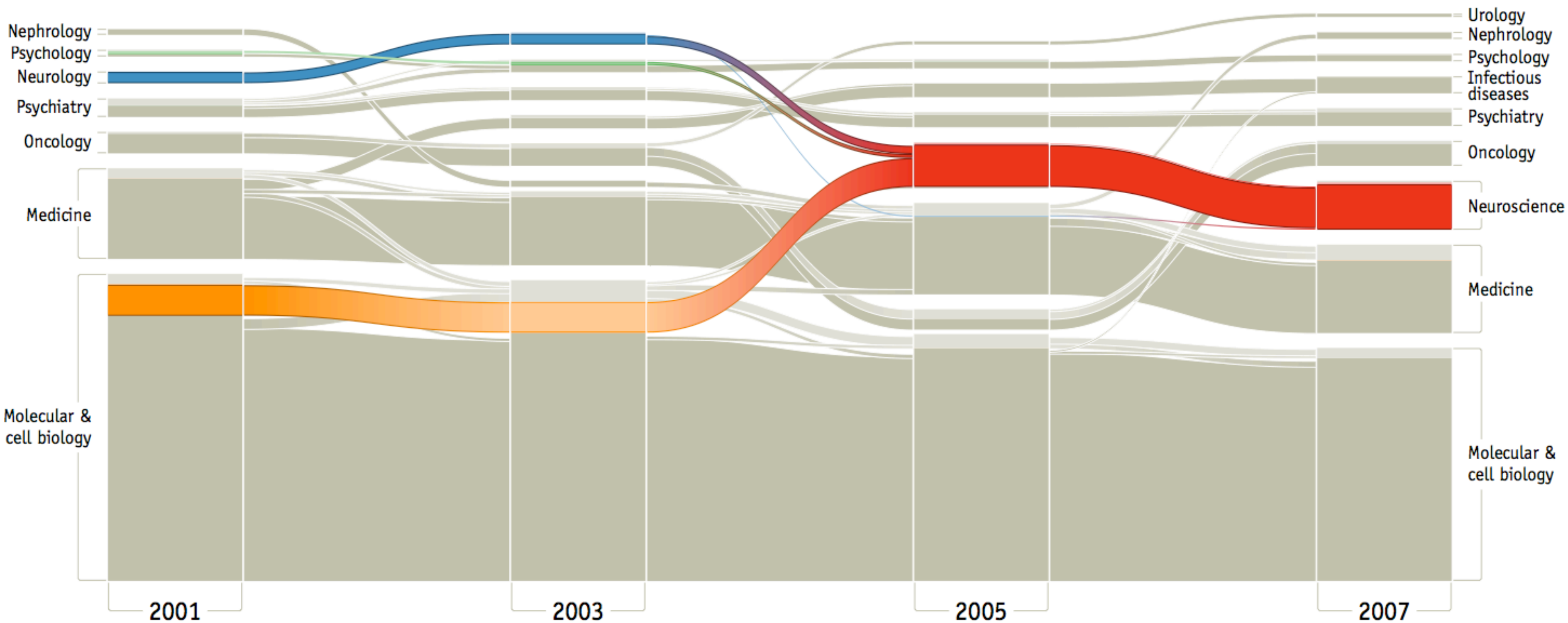
1995


Educational Assessment
Fluid Dynamics
Operations Research
Mathematics
Theoretical Physics
Astrophysics
Geoscience
Control Theory
Discrete Math
General Physics
Statistics
Telecommunications Engineering
Petroleum
Chemistry
Environmental Science
Information Science
Food Chemistry
History
Management & Marketing
Education
Bacteriology
Behaviorist Psychology
Psychoanalysis
Economics
Political Science
Animal Agriculture
Psychology
Geography
Sociology
Pharmacology & Biomaterials
Ecology & Evolution
Psychiatry
Otolaryngology
Law
Molecular Biology (Plant)
Parasitology
Neurology
Veterinary
Anthropology
Molecular Biology
Orthopedics
Ergonomics
Ophthalmology
General Medicine
Dentistry
Nursery
Dermatology
Anesthesiology
Medical Physics

2004


Fluid Mechanics
Circuits
Material Engineering
Computer Science
Geosciences
Tribology
Operations Research
Astronomy & Astrophysics
Computer Imaging
Mathematics
Power Systems
Physics
Telecommunication
Electromagnetic Engineering
Control Theory
Chemical Engineering
Probability & Statistics
Environmental Chemistry & Microbiology
Chemistry
Applied Acoustics
Analytic Chemistry
Business & Marketing
Economics
Geography
Psychology
Crop Science
Sociology
Education
Pharmacology
Ecology & Evolution
Political Science
Neuroscience
Agriculture
Law
Psychiatry
Environmental Health
Medical Imaging
Anthropology
Orthopedics
Veterinary
Molecular & Cell Biology
Parasitology
Dentistry
Medicine
Ophthalmology
Otolaryngology
Gastroenterology
Urology
Pathology
Dermatology
Rheumatology

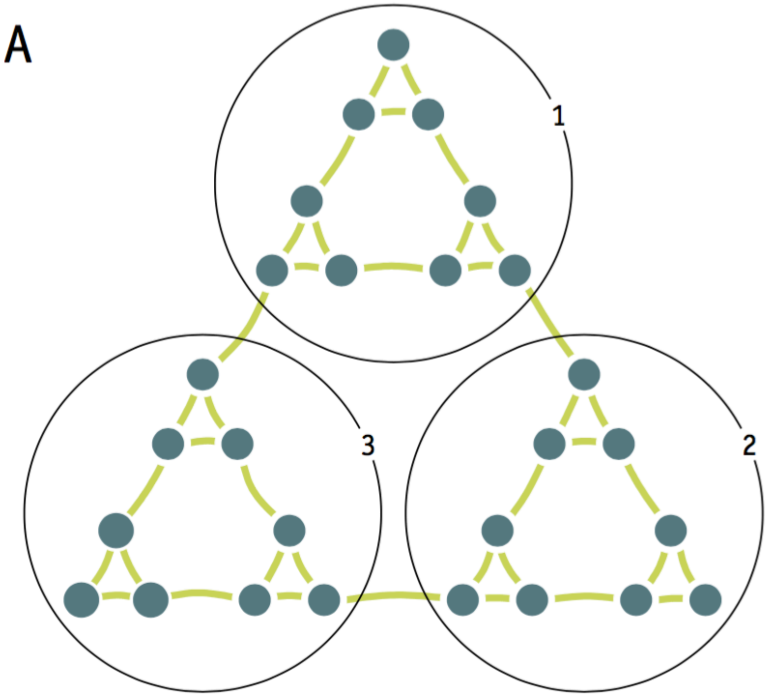
The emergence of Neuroscience



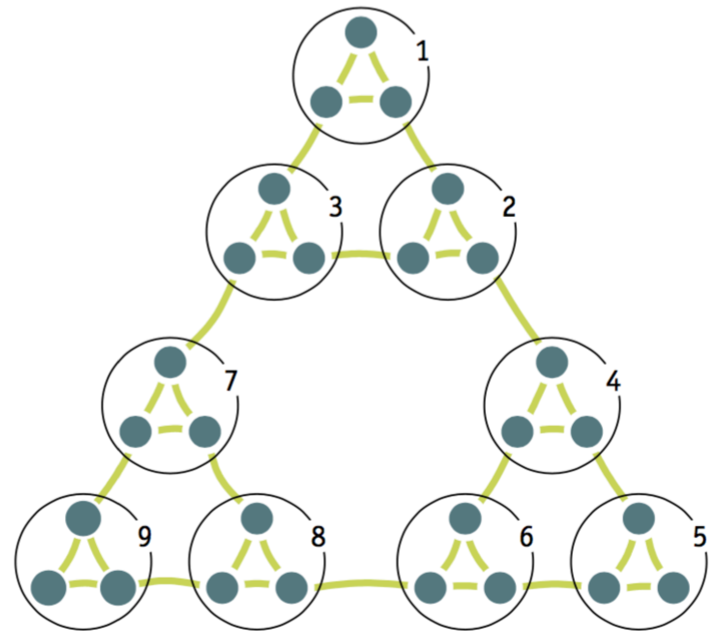
Rosvall and Bergstrom (2010) PLoS One

Two-level partitions with the map equation

A



B



$$L(\mathsf{M}) = \underbrace{q_{\curvearrowleft} H(\mathcal{Q})}_{0.12\text{ bits}} + \underbrace{\left\{\begin{matrix} p^1_{\circlearrowright} H(\mathcal{P}^1) \\ p^2_{\circlearrowright} H(\mathcal{P}^2) \\ p^3_{\circlearrowright} H(\mathcal{P}^3) \end{matrix}\right.}_{3.56\text{ bits}} = 3.68 \text{ bits.}$$

$$L(\mathsf{M}) = \underbrace{q_{\curvearrowleft} H(\mathcal{Q})}_{0.97\text{ bits}} + \underbrace{\left\{\begin{matrix} p^1_{\circlearrowright} H(\mathcal{P}^1) \\ p^2_{\circlearrowright} H(\mathcal{P}^2) \\ p^3_{\circlearrowright} H(\mathcal{P}^3) \\ p^4_{\circlearrowright} H(\mathcal{P}^4) \\ p^5_{\circlearrowright} H(\mathcal{P}^5) \\ p^6_{\circlearrowright} H(\mathcal{P}^6) \\ p^7_{\circlearrowright} H(\mathcal{P}^7) \\ p^8_{\circlearrowright} H(\mathcal{P}^8) \\ p^9_{\circlearrowright} H(\mathcal{P}^9) \end{matrix}\right.}_{2.60\text{ bits}} = 3.57 \text{ bits.}$$

Advantages of InfoMap

- Outperforms most other competing algorithms in accuracy benchmark tests
- Hierarchical version from first principles
- Fast, scalable codebase in C++
- Multiplex versions
- Continual development
- Visual tools associated with the code
- Simple information theoretic framework

Dynamics

Journal Ranking

$$P = \alpha H + (1 - \alpha)\ a.e^T$$

Matrix representing the random walk over citations

Probability of not teleporting

Cross-citation Matrix dictating the structure of the citation network

Probability of teleporting to completely new journal weighted by the number of articles in that journal

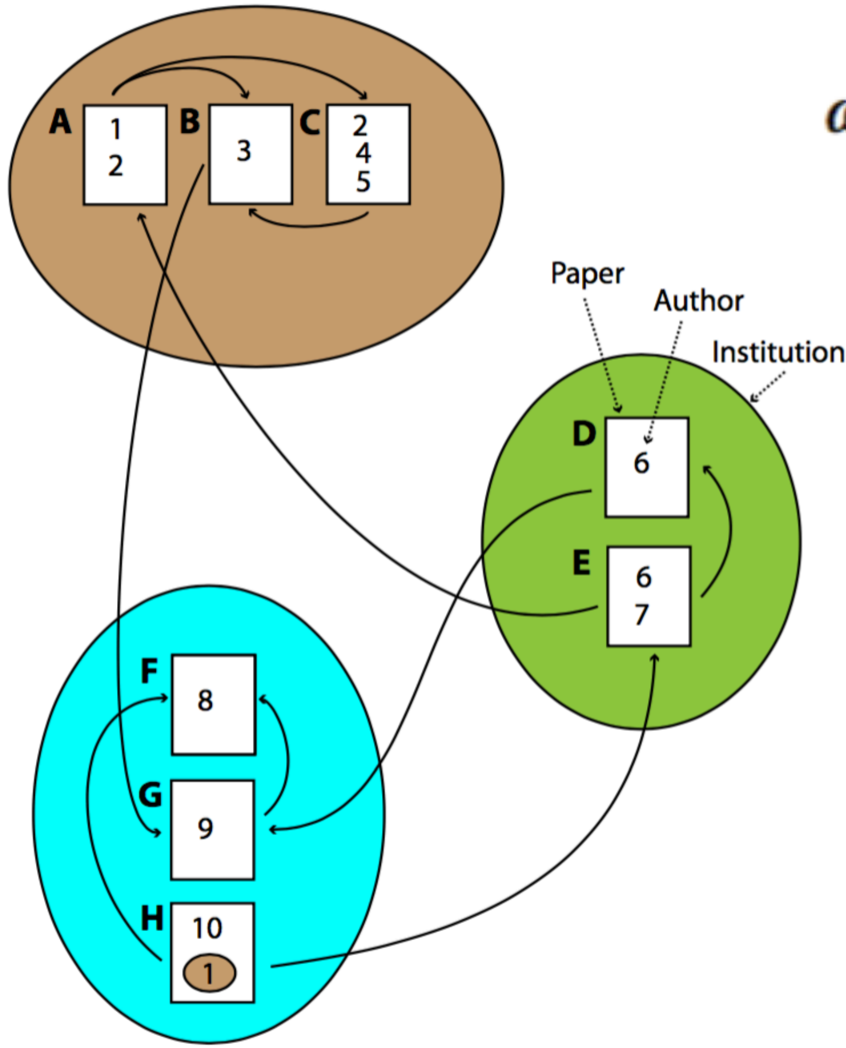
$$EF = 100 \frac{H\pi}{\sum_i [H\pi]_i}$$

Leading eigenvector of the random walk matrix P.

Normalization

West, JD et al. (2010) *College of Research Libraries*

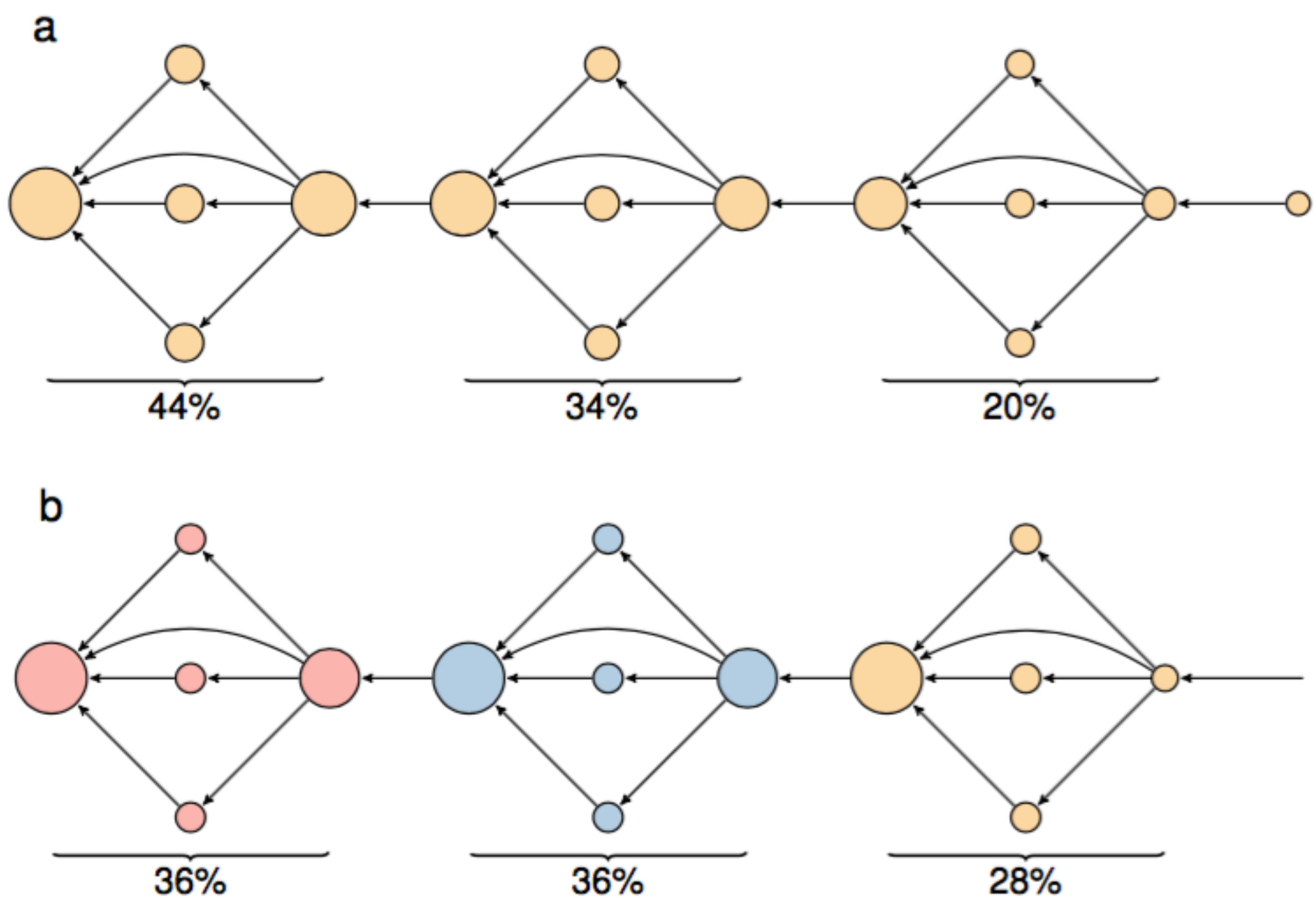
Author-level Ranking



$$\omega = \frac{1}{c(X)}\frac{1}{m}\frac{1}{n}$$

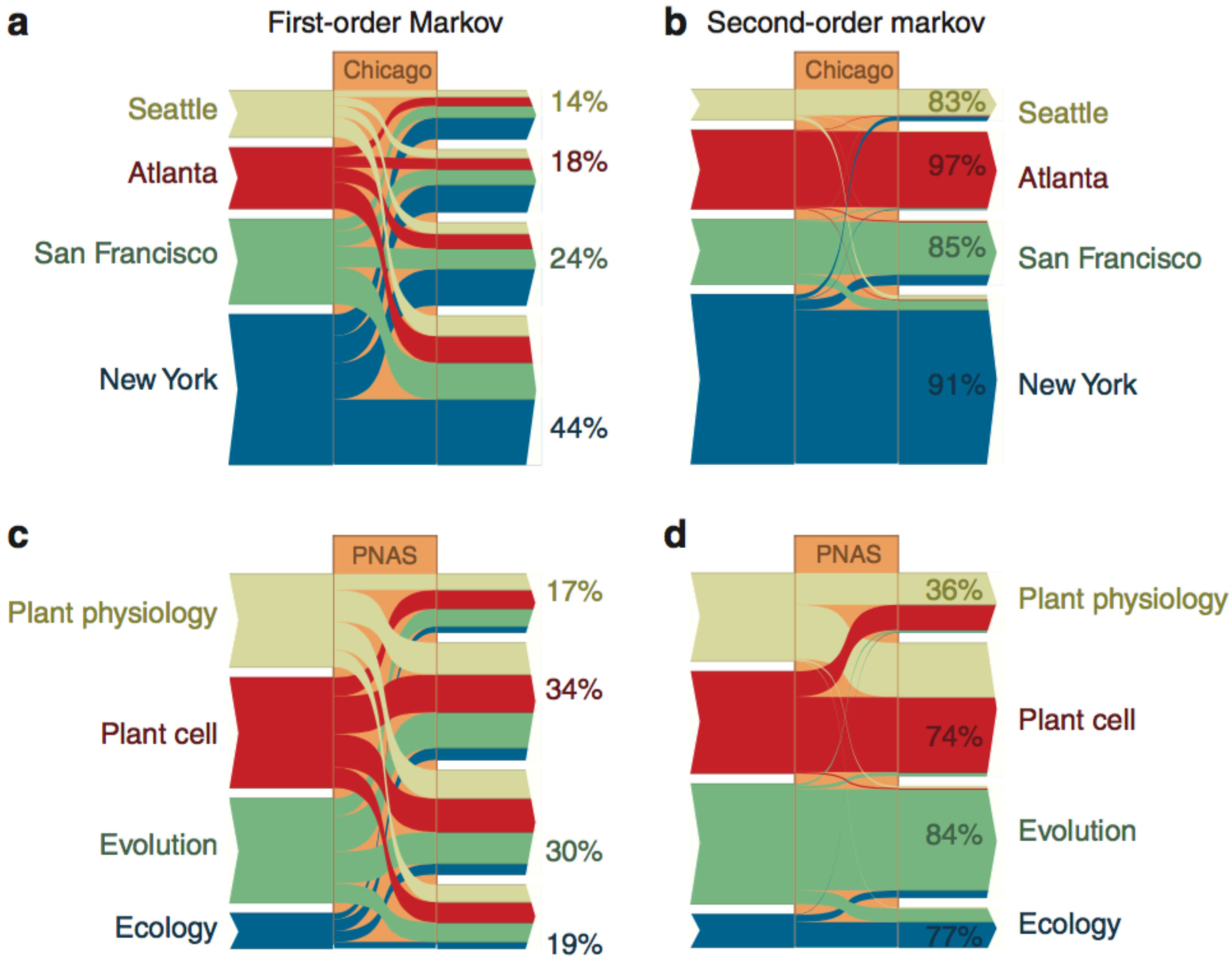
West et al. (2013) Author-level Eigenfactor metrics. *JASIST*

Article-level Ranking



West et al. (2016) Ranking and mapping article-level citation networks. *in prep.*

Higher Order Dynamics



Rosvall et al. (2014) Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*

The Graph

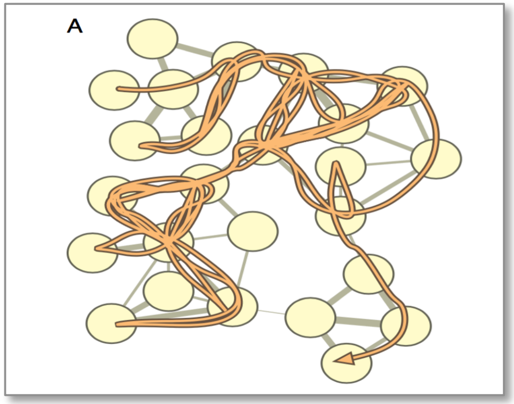
The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between publications, as well as authors, institutions, journal and conference "venues," and fields of study.

The Data

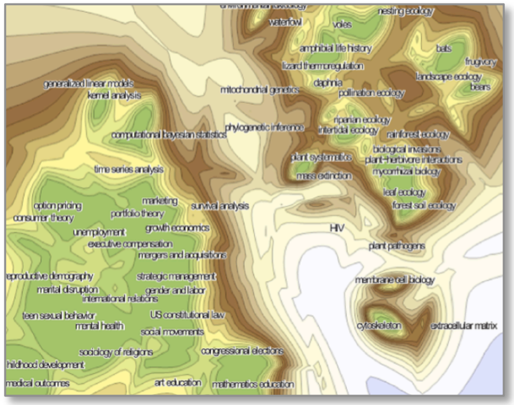
This data is available as a set of zipped text files stored in Microsoft Azure blob storage and available via HTTP. The file size (zipped) is ~30GB and may be downloaded here.

The Challenge

The goal of the Ranker Challenge is to assess the query-independent importance of scholarly articles, using data from the Microsoft Academic Graph.



Science of Mapping



Mapping of Science

What can you do with citation maps?

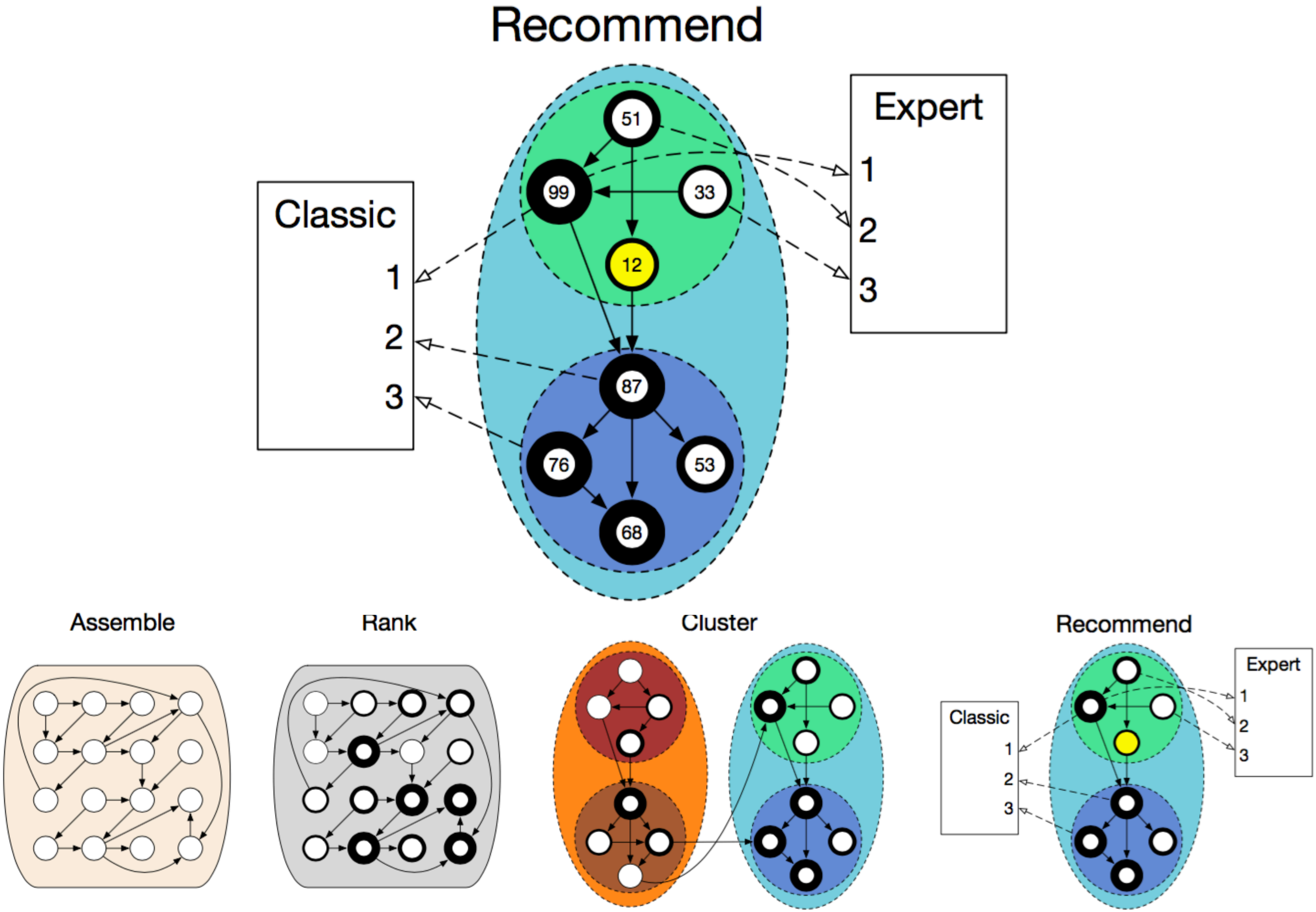
Recommend Articles

Study the Science of Science

Author Evaluation

Search and Discovery

Visual Interfaces



West, Wesley-Smith, Bergstrom (2016) A recommendation system based on hierarchical clustering of an article-level citation network. *in review*

Babel Home About API

babel.eigenfactor.org

oren etzioni

Ian Wesley-Smith

v DBLP JSTOR MAS PLOS PubMed

s

al Methods For Analyzing Speedup Learning Experiments O Etzioni satisfaction programs

-Face And Computer-Mediated Communities Amitai Etzioni, Oren Etzioni 1998 resources sustained

cument Clustering O Zamir document clustering

Communities: Virtual Vs. Real A Etzioni 1996 implications internet

al Methods For Analyzing Speedup Learning Experiments. O Etzioni 1993 scheduling problems

al Methods For Analyzing Speedup Learning Experiments O Etzioni 1993 generating abstractions

Get Related Web Document Clustering: A Feasibility Demonstration O Zamir 1997 document clustering

Get Related Web Document Clustering: A Feasibility Demonstration. O Zamir 1997 browsing large

Get Related Sound And Efficient Closed- World Reasoning O Etzioni proving problem

Get Related Appears In Comm. OfACM O Etzioni scalable comparison-shopping

« Previous 1 2 3 4 5 6 7 8 9 10 Next »

Papers related to Statistical Methods For Analyzing Speedup Learning Experiments O Etzioni satisfaction programs

Get Related Automatically Configuring Constraint Satisfaction Programs: A Case Study S Minton 1995 satisfaction programs

Get Related Abstraction Via Approximate Symmetry T Ellman 1992 satisfaction programs

Get Related Integrating Heuristics For Constraint Satisfaction Problems: A Case Study S Minton 1992 satisfaction programs

Get Related An Analytic Learning System For Specializing Heuristics S Minton 1992 satisfaction programs

Get Related Automated Synthesis Of Constrained Generators W Braudaway 1988 satisfaction programs

Scholar Mirrors

Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics
in Google Scholar Citations, ResearcherID, Researchgate, Mendeley, and Twitter

HOME | ABOUT | METHODOLOGY | OUR TEAM | OTHER PROJECTS

AUTHORS | DOCUMENTS | JOURNALS | PUBLISHERS

General overview

Displaying core authors 1-20 of 398. Sorted by GS citations (last 5 years), decreasingly.

☐ Check to display related authors as well

Search an author

Name	Online presence	Google Scholar +		ResearcherID +		ResearchGate +		Mendeley +		Twitter +	
		Citations	H Index	Citations	H Index	RG Score	Downloads	Readers	Followers	Tweets	Followers
Loet Leydesdorff		26484	73	6444	44	45.14	32165	0	11	84	375
Eugene Garfield*		22622	55	8790	153	-	-	-	-	-	-
Mike Thelwall		13840	61	3593	32	42.64	24989	7423	36	85	522
Derek J. de Solla Price		13263	33	-	-	-	-	-	-	-	-
Francis Narin		11297	45	-	-	32.38	795	-	-	-	-
Wolfgang Glänzel		10796	54	4924	38	41.16	10572	-	-	-	-
Ronald Rousseau		9570	42	NA	NA	42.75	8066	-	-	-	-
Chaomei Chen		9512	43	1740	20	34.65	31579	965	3	67	65
Anthony (Ton) F.J. van Raan		9200	53	-	-	38.47	6014	-	-	58	166
Ben R Martin		8975	39	-	-	-	-	-	-	-	-
András Schubert		8655	45	4121	31	39.24	1962	-	-	-	-
Peter Ingwersen		8356	35	NA	NA	30.64	8600	-	-	-	-
Henk F. Moed		8256	46	-	-	-	-	-	-	-	-
Blaise Cronin		7347	43	-	-	33.9	1891	-	-	-	-
Henry Small		7307	32	3360	23	-	-	-	-	-	-

scholar.eigenfactor.org

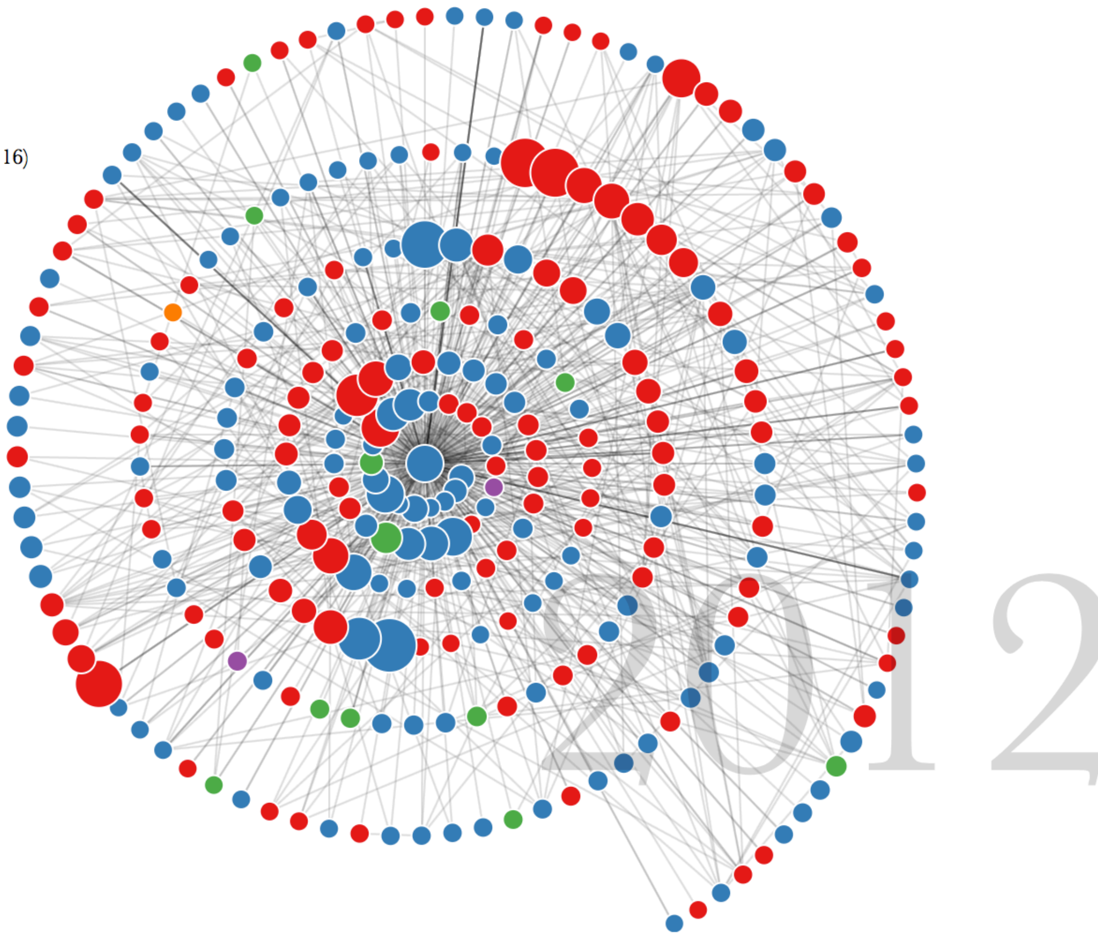
Visualizing Scholarly Influence Over Time

Influence of Pew Scholars

Roberta A. Gottlieb

Learn More

- Papers in category "Medicine" (domain 6)
- Papers in category "Biology" (domain 4)
- Papers in category "Chemistry" (domain 5)
- Papers in category "Unknown" (domain 0)
- Papers in category "Agriculture Science" (domain 16)

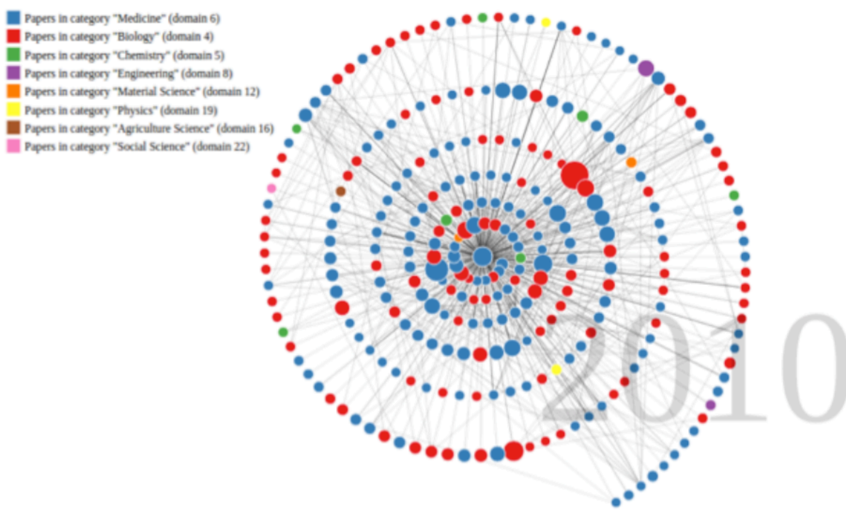


Roberta A. Gottlieb

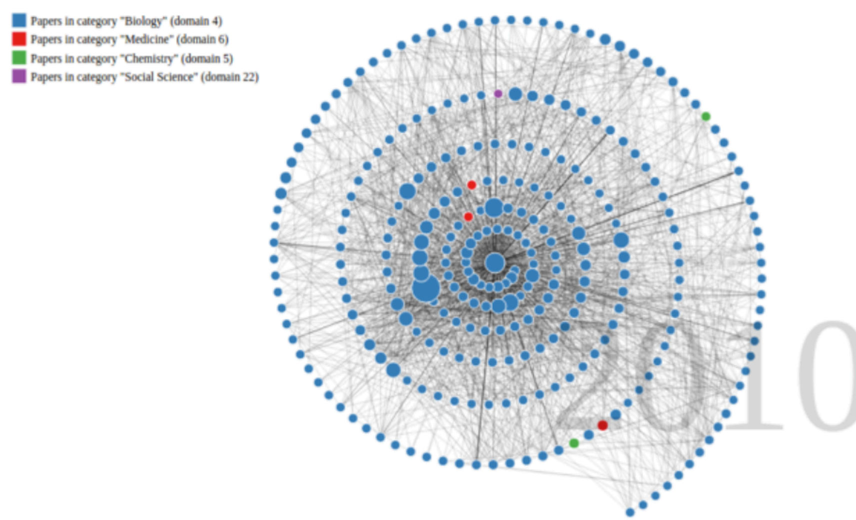
Pew Scholar
1997

Portenoy et al. (2016) Leveraging Citation Networks to Visualize Scholarly Influence Over Time. *in review*

Comparing Authors



A denser network means that the papers that cite the central author also tend to cite each other.



A more sparse network indicates fewer citations between papers shown in the network. This could be a result of the central scholar having impact across a wider set of academic communities.

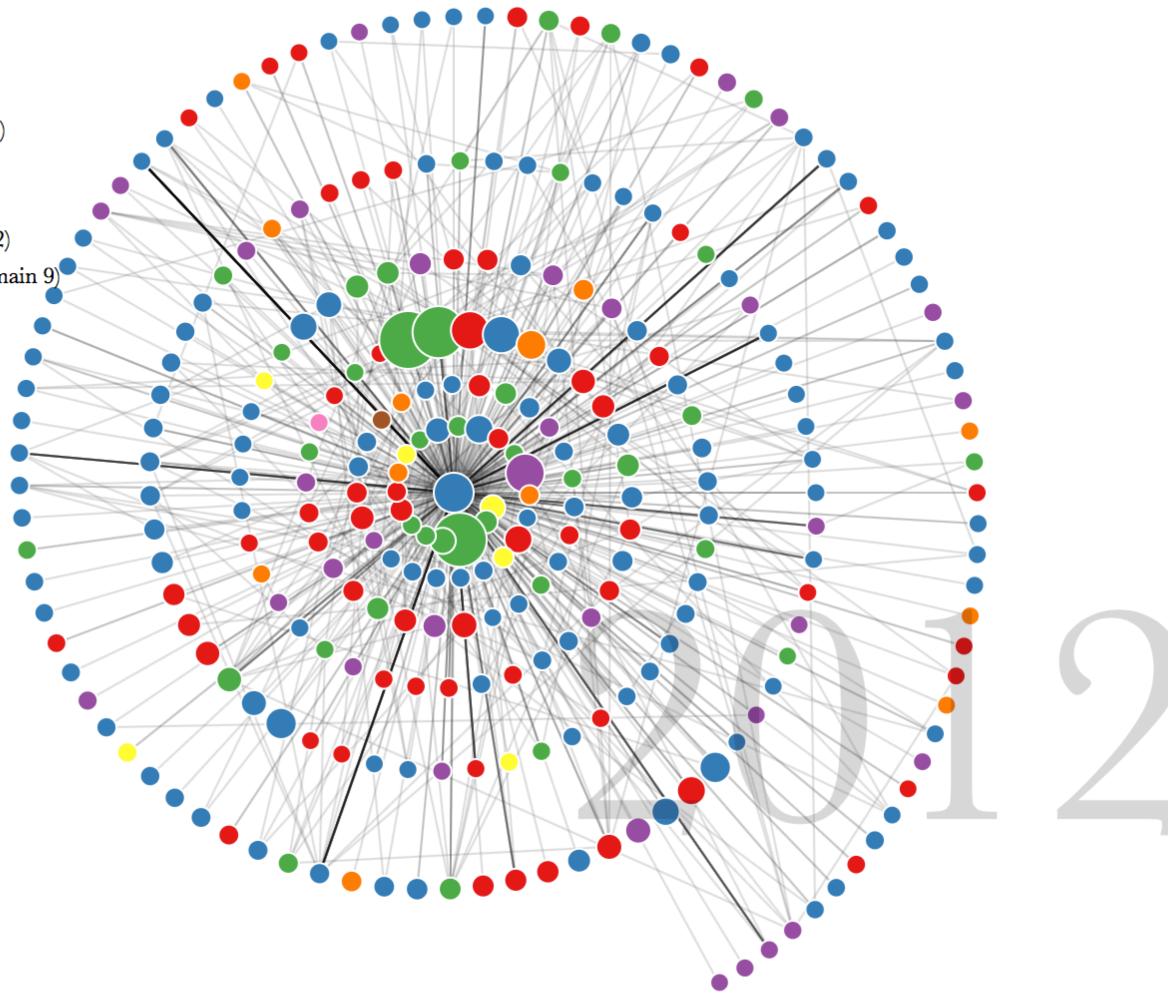
Visualizing Scholarly Influence Over Time

Influence of Pew Scholars

Mark W. Grinstaff

[Learn More]

Papers in category "Chemistry" (domain 5)
Papers in category "Medicine" (domain 6)
Papers in category "Biology" (domain 4)
Papers in category "Material Science" (domain 12)
Papers in category "Engineering" (domain 8)
Papers in category "Physics" (domain 19)
Papers in category "Computer Science" (domain 2)
Papers in category "Environmental Sciences" (domain 9)

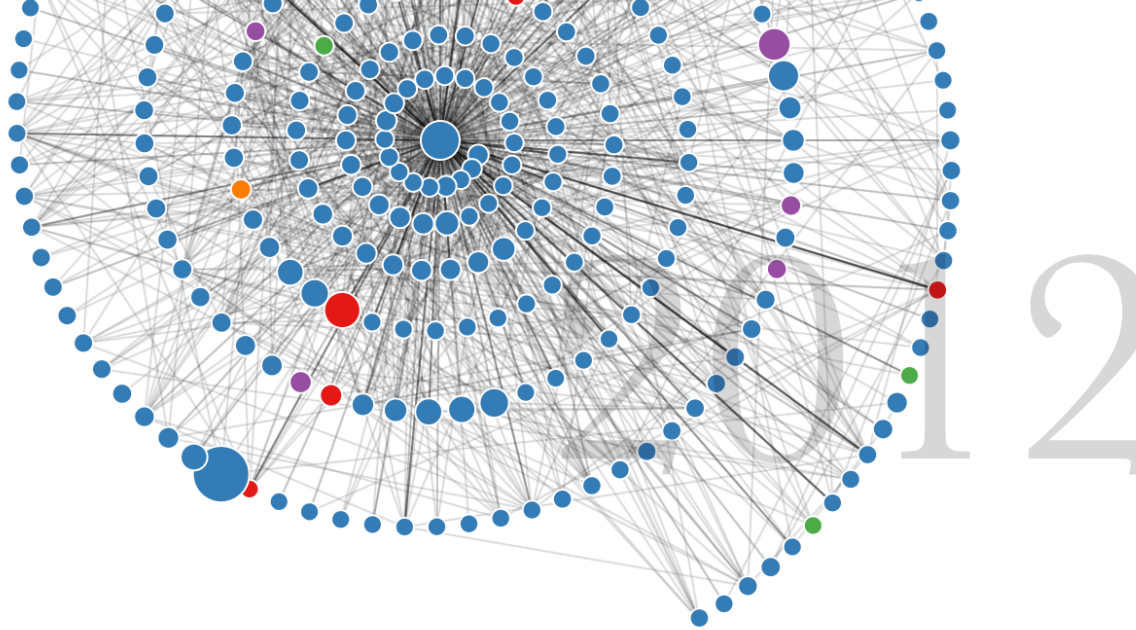


Mark W. Grinstaff

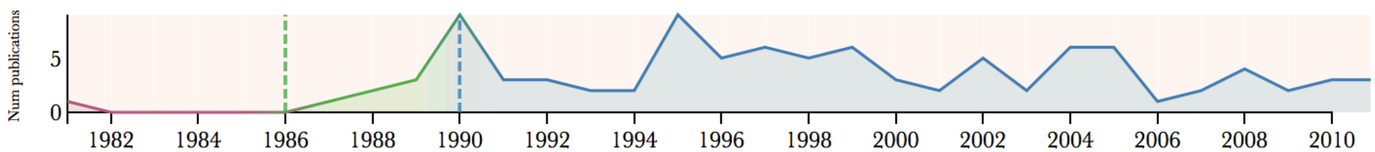
Pew Scholar
1999

Philip A.
Hieter

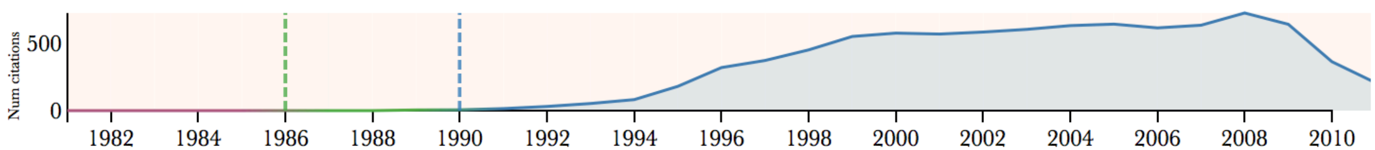
Pew Scholar
1986



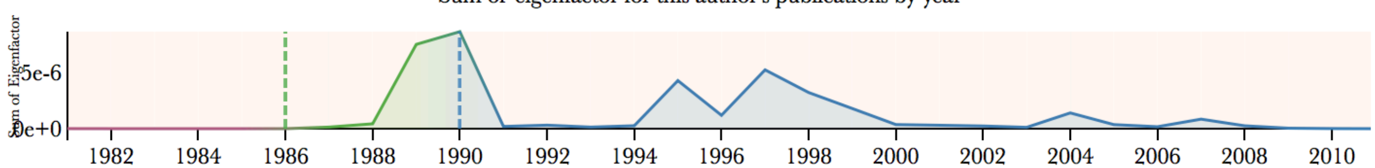
Number of publications



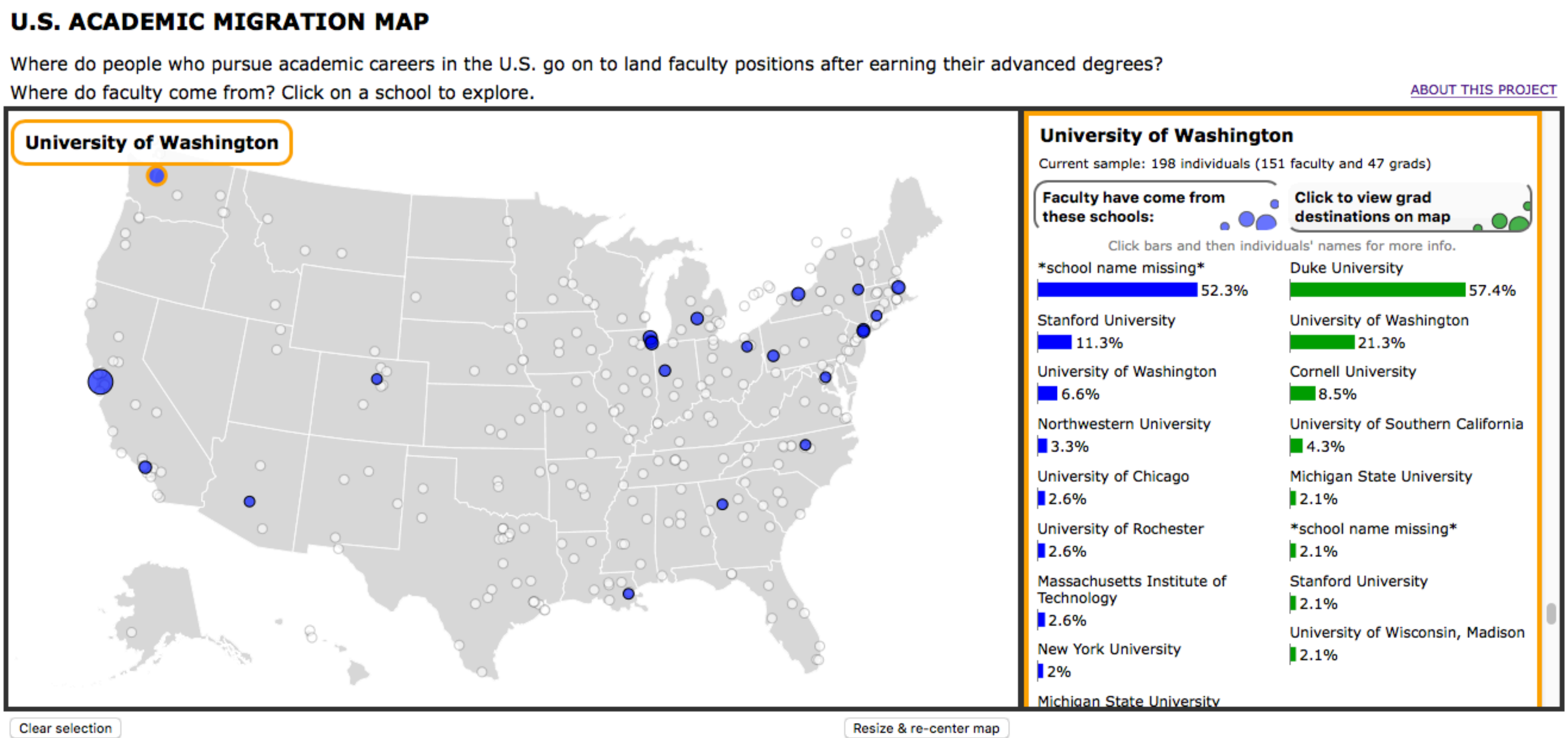
Number of citations received



Sum of eigenfactor for this author's publications by year



Academic Migration



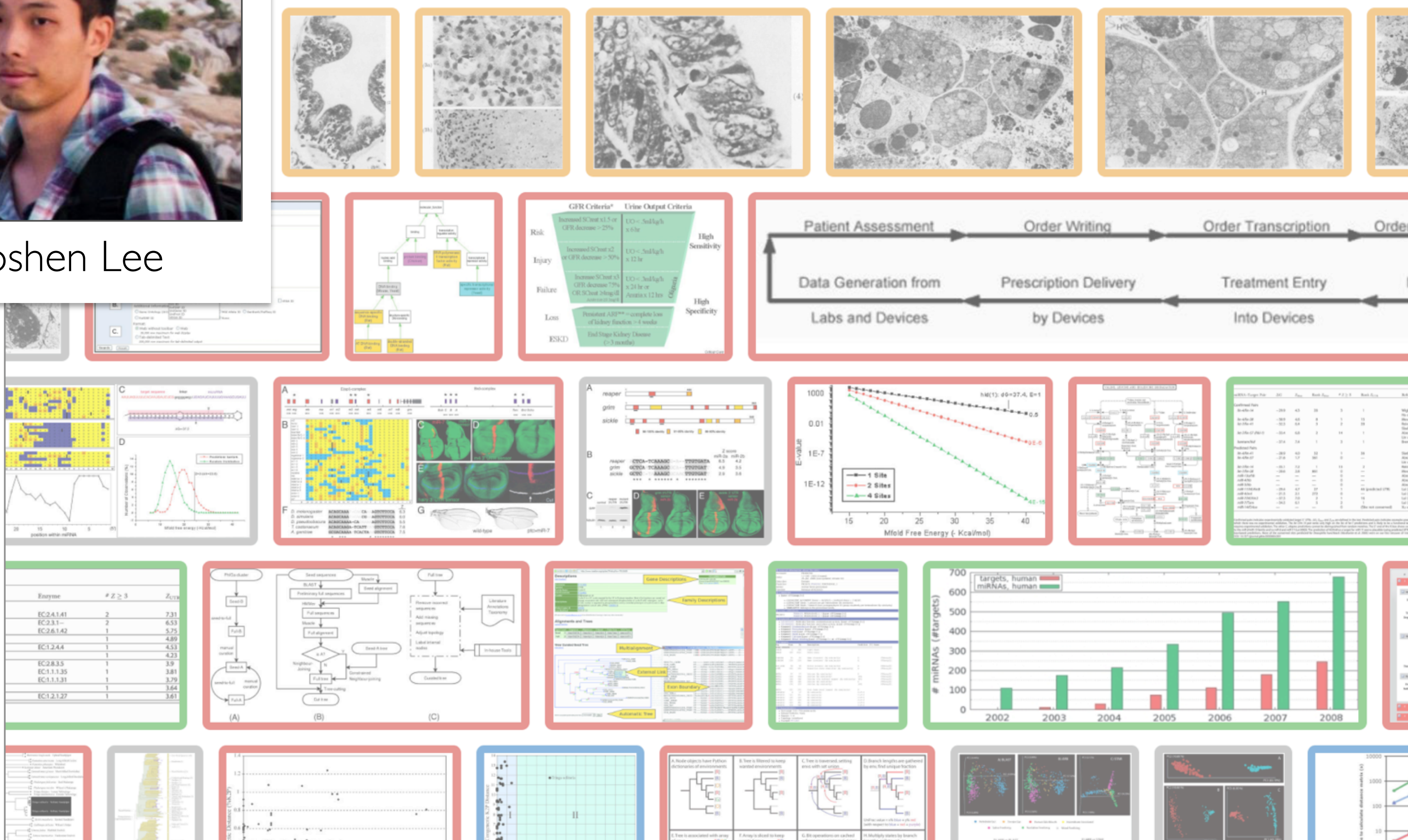
VizioMetrix About Search Crowdsourcing

viziometrics.org

Poshen Lee

Impact | Keywords or Cluster, Result Ordered by Impact | Search

Composite Equation Diagram Photo Plot Table



A project of the eScience Institue at the University of Washington

Questions

- How do patterns of encoding visual information in the literature vary across disciplines?
- How have patterns of encoding visual information in the literature evolved over time?
- Is there any link between patterns of encoding visual information and scientific impact?

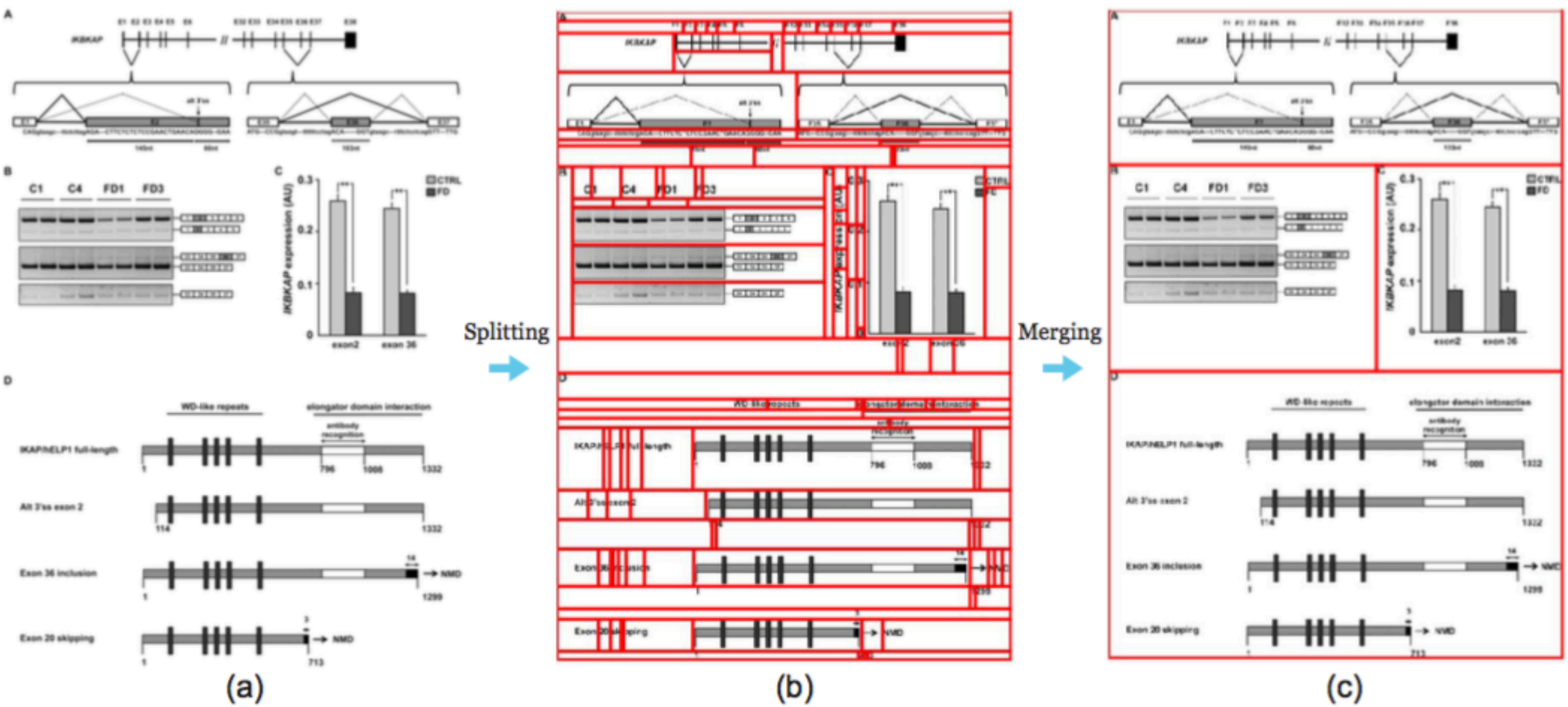
How can we better utilize visual information in the search and navigation process?

PubMed

650,000 papers

5 million images

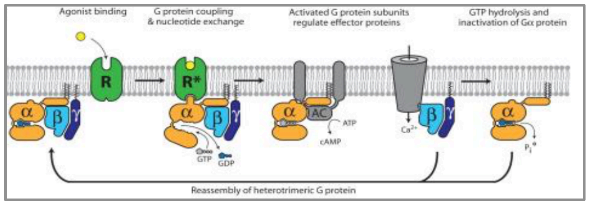
Composite Figure Dismantling



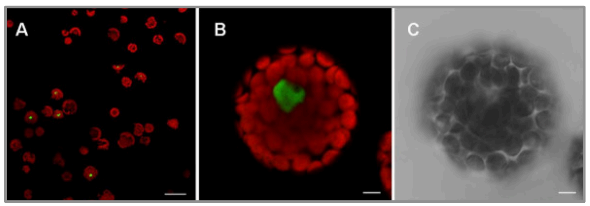
Lee et al (2015)

$$w_i = \sum_j^n (Z_{ij} + Z_{ij}^T)$$

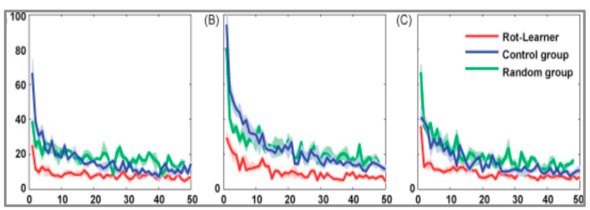
Equations (394)



Schematics (769)



Photos (782)

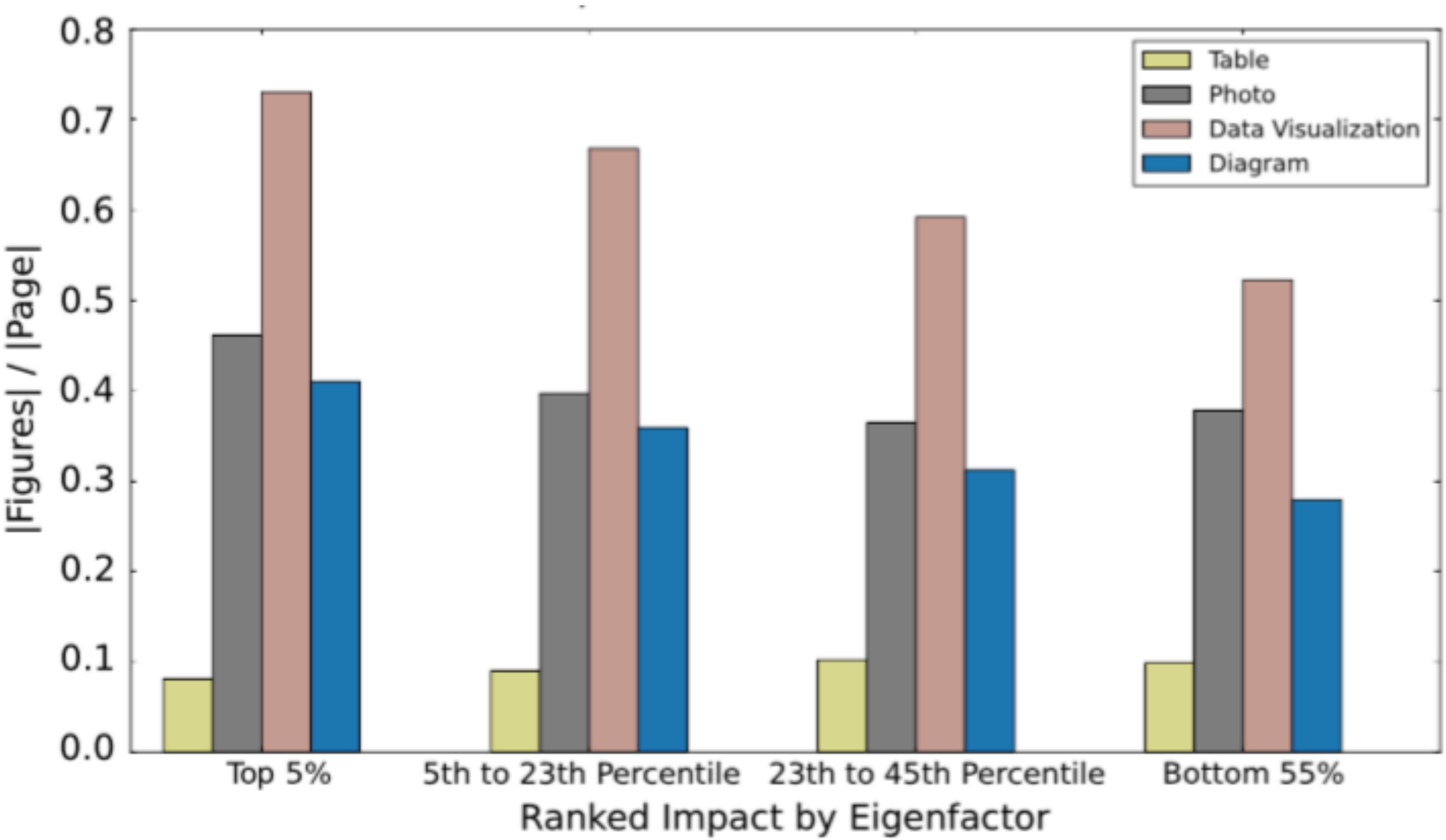


Plots (890)

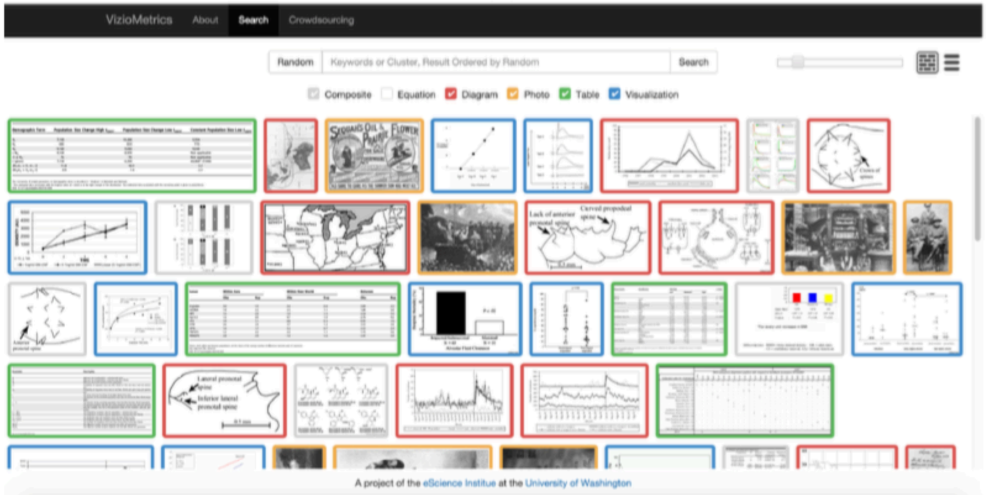
	PW reading	PW reading	PW reading	W reading	W reading
	W RT	*PW RT*	*CTL*	*W RT*	*PW RT*
MOG → LOT	0.28	0.18	***0.58***	***−0.70***	***−0.50***
MOG → LP	−0.22	***−0.52***	−0.04	0.27	−0.03
LOT → LP	0	0.10	0.24	***−0.56***	***−0.60***
LOT → IFG	0.38	0.17	0.40	0.43	0.13
LP → IFG	0.26	0.05	0.31	0.03	−0.03

Tables (436)

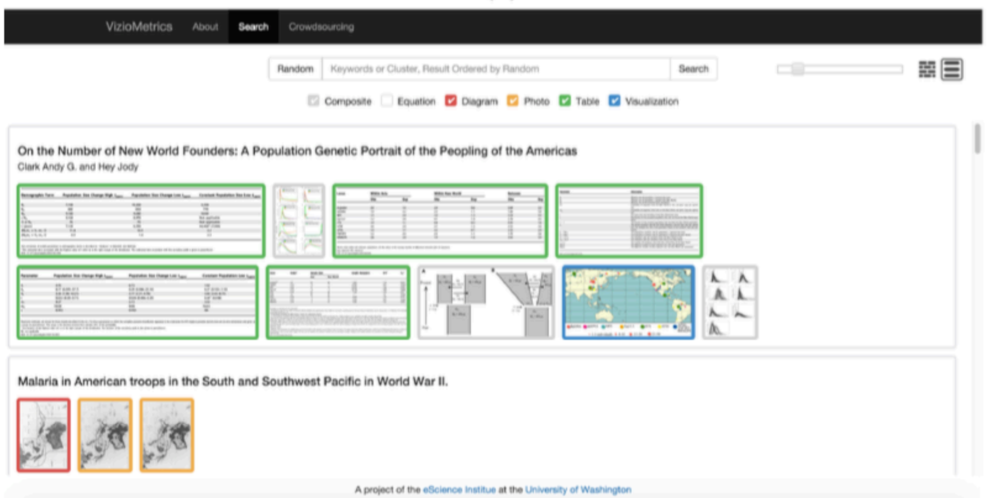
Impact versus Figure Density

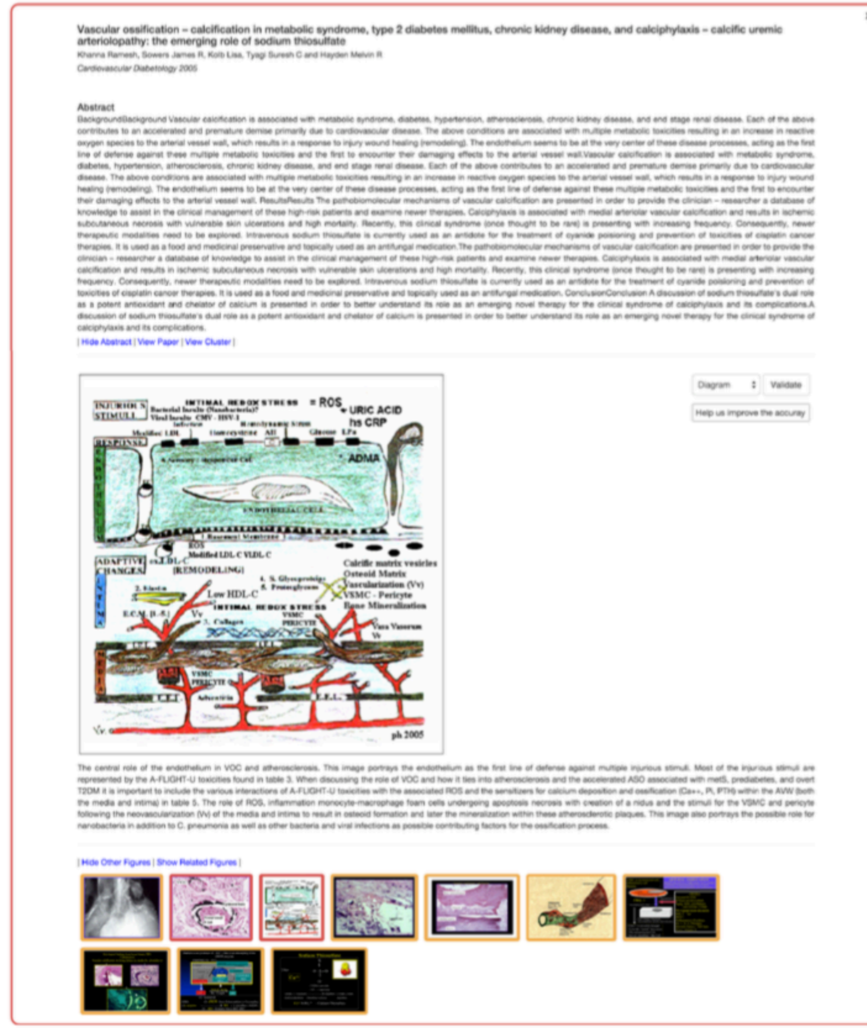


viziometrics.org

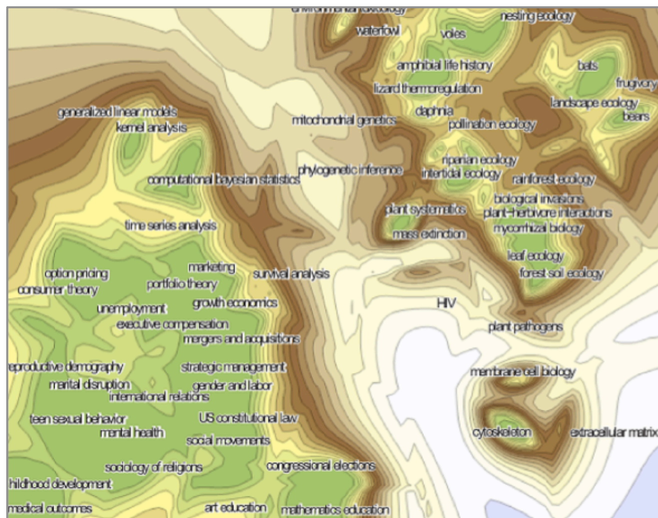


(b)



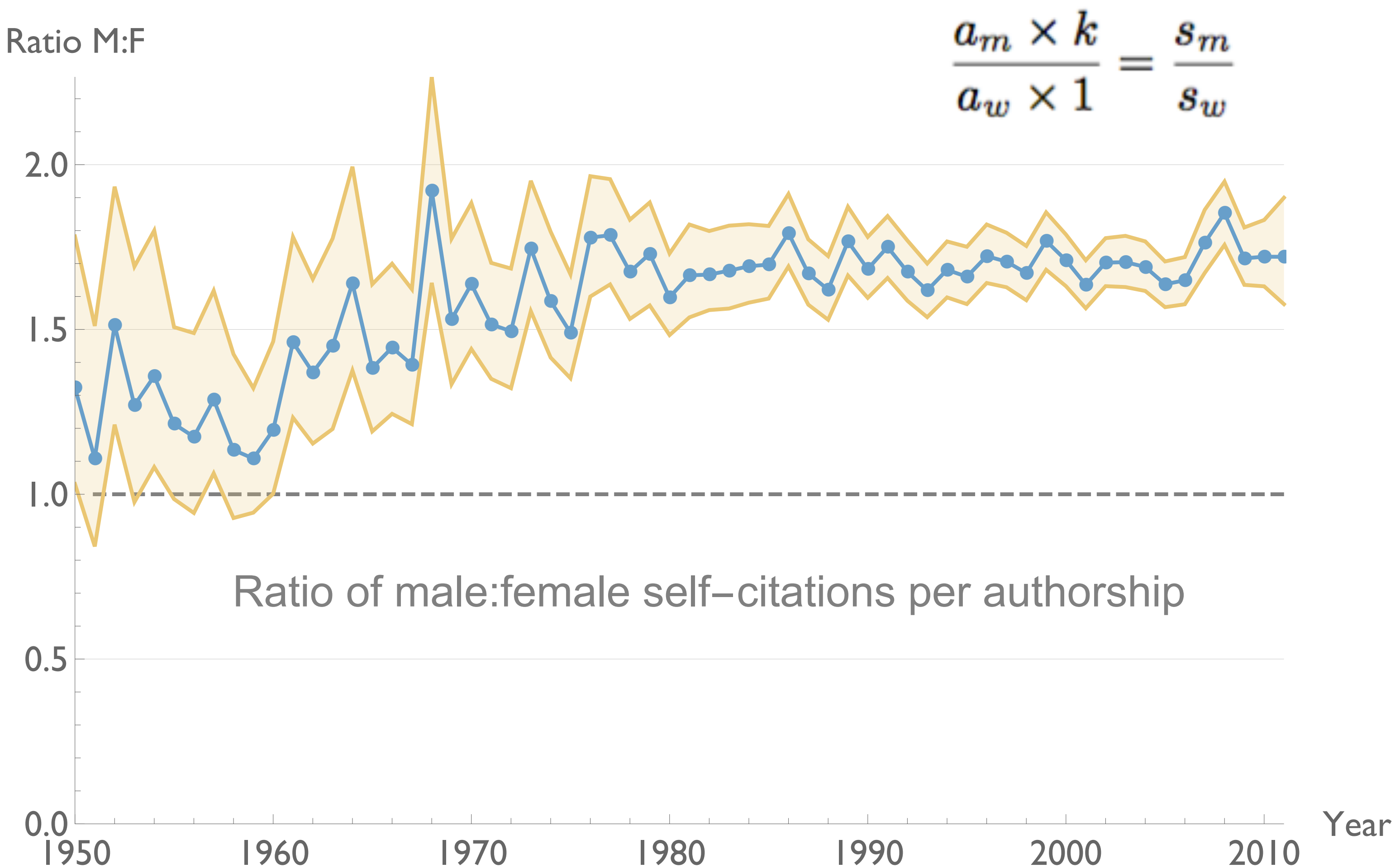


Science of Science

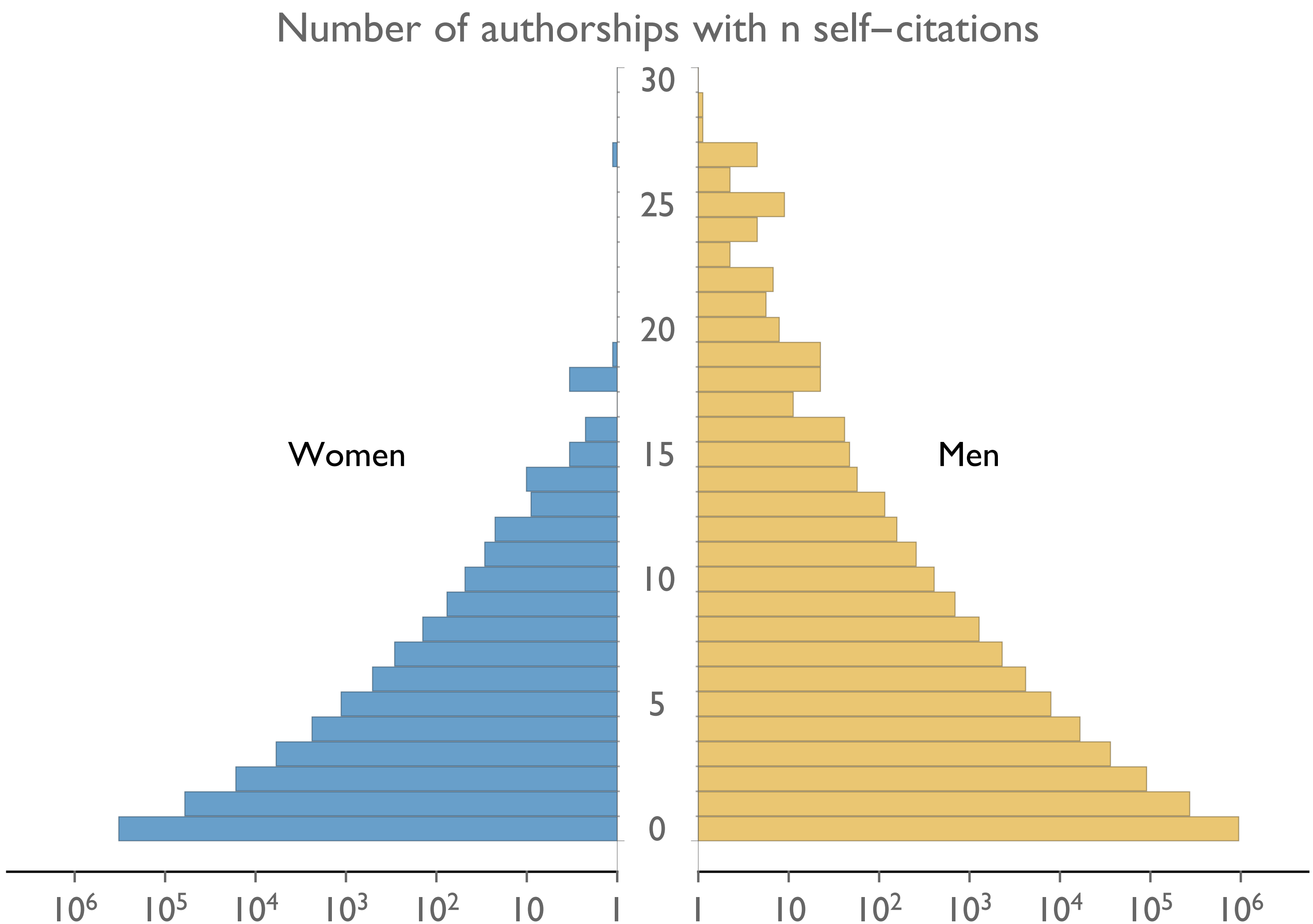


Incentives, reproducibility, biases, funding models, publishing economics…

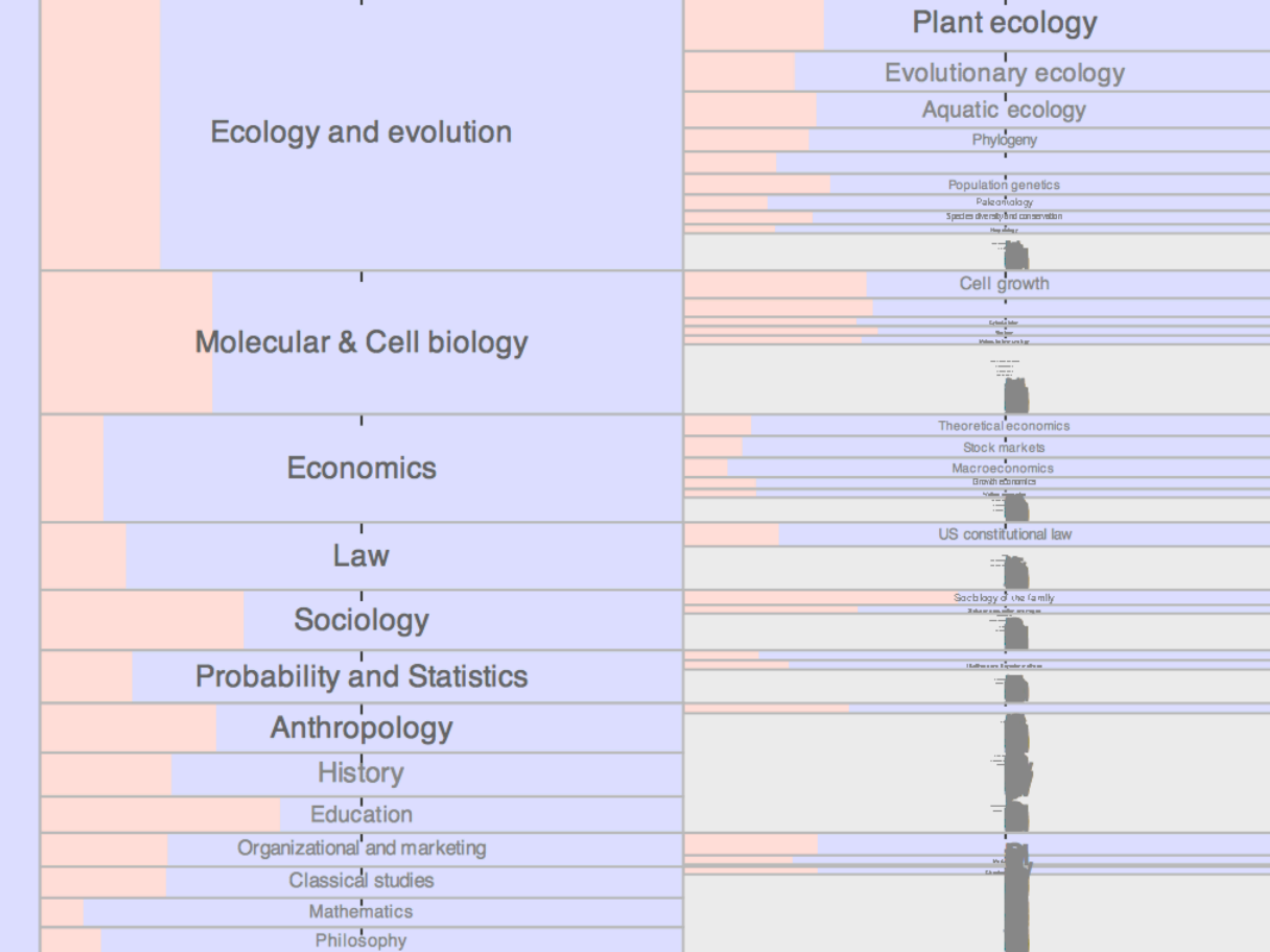
Self-citation over time



Excessive Self Citation

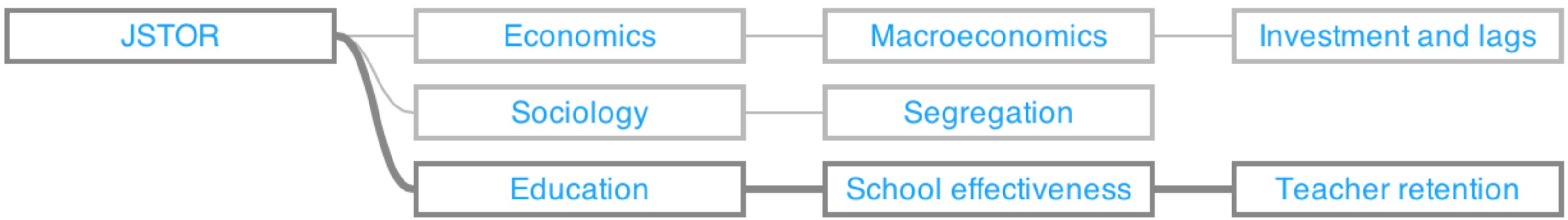


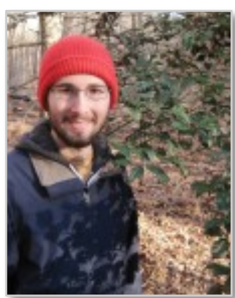
King et al (2016) in prep.



Plant ecology
Evolutionary ecology
Aquatic ecology
Phylogeny
Population genetics
Paleontology
Species diversity and conservation
Ecology and evolution
Cell growth
Molecular & Cell biology
Theoretical economics
Stock markets
Macroeconomics
Growth economics
Economics
US constitutional law
Law
Sociology of the family
Sociology
Probability and Statistics
Anthropology
History
Education
Organizational and marketing
Classical studies
Mathematics
Philosophy

Information Visualization



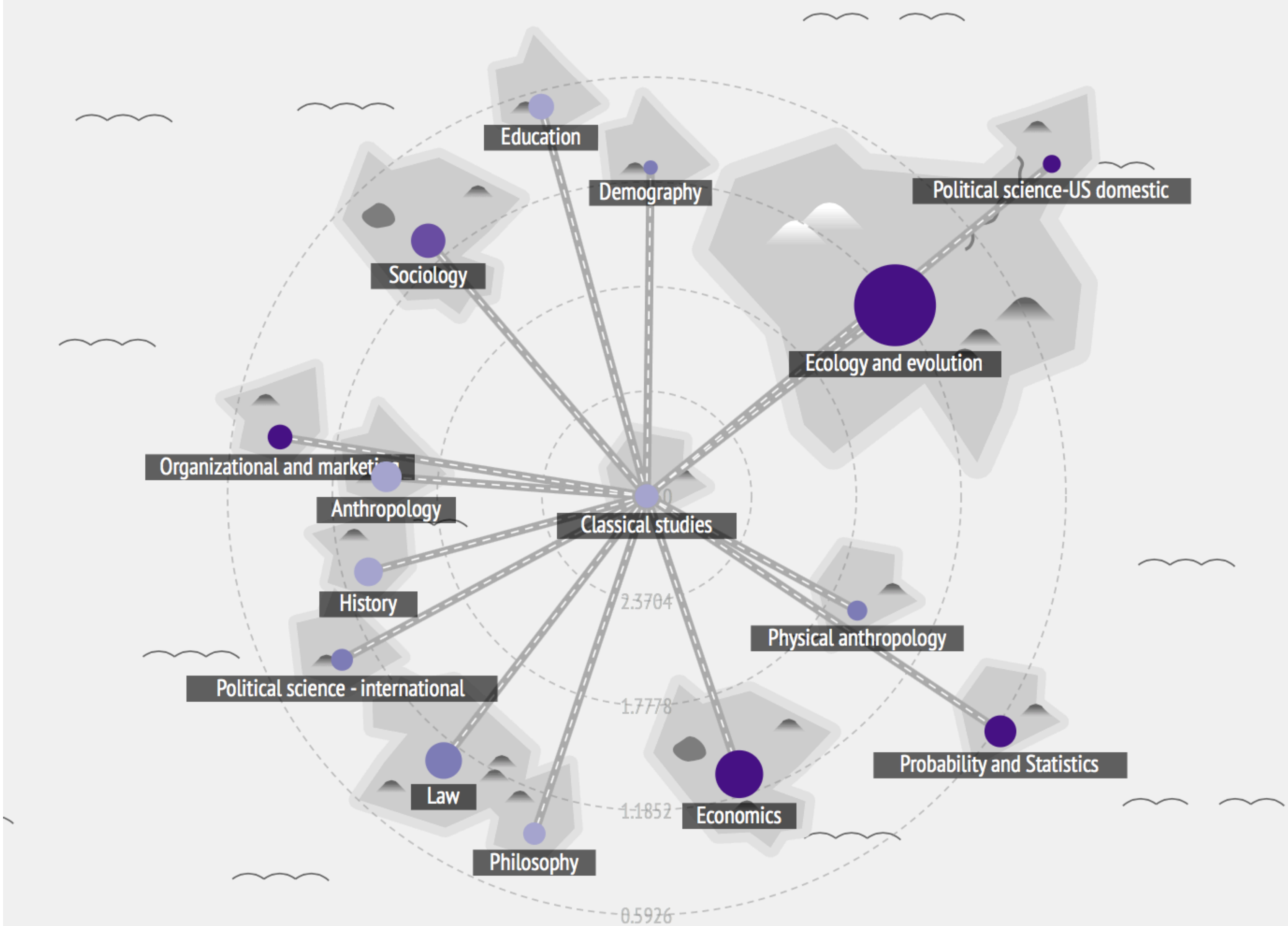


Cecilia Aragon
UW, HCDE

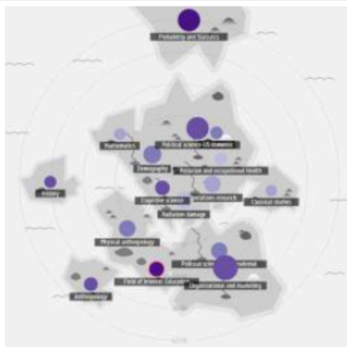
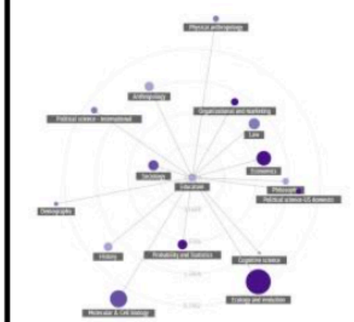
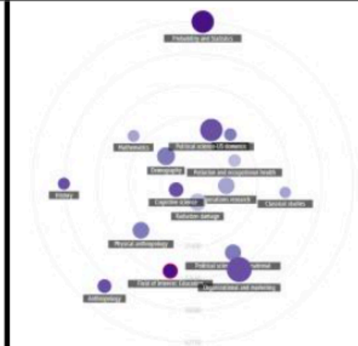
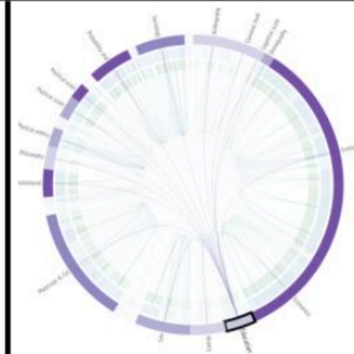
Michael Brooks
UW, HCDE

How do we augment human memory when zooming in/out of hierarchical trees?

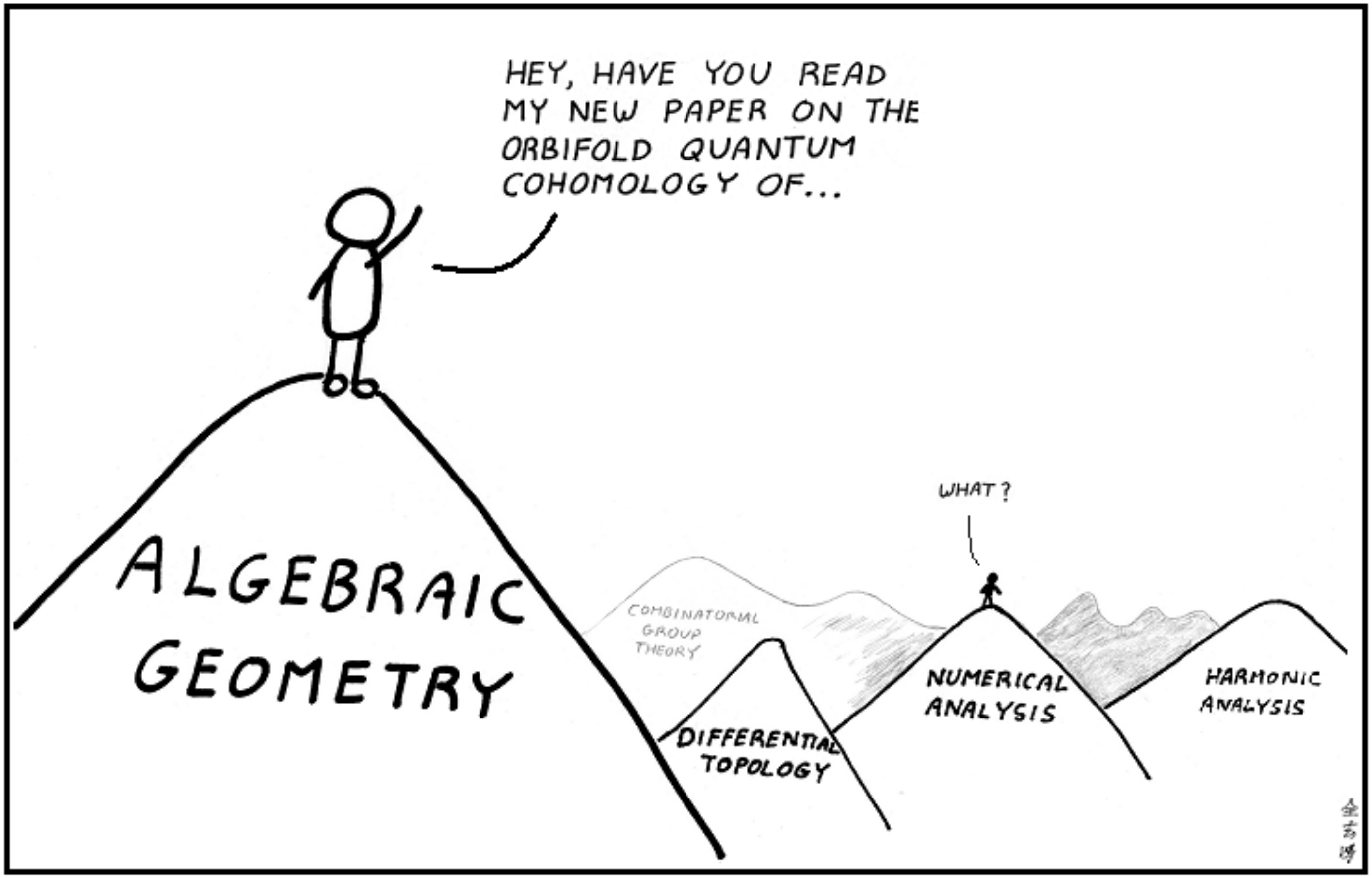
Brooks, M *et al.* (2013) *INTERACT*


Education
Demography
Political science-US domestic
Sociology
Ecology and evolution
Organizational and marketing
Anthropology
Classical studies
History
2.3704
Physical anthropology
Political science - international
1.7778
Probability and Statistics
Law
1.1852
Economics
Philosophy
0.5926

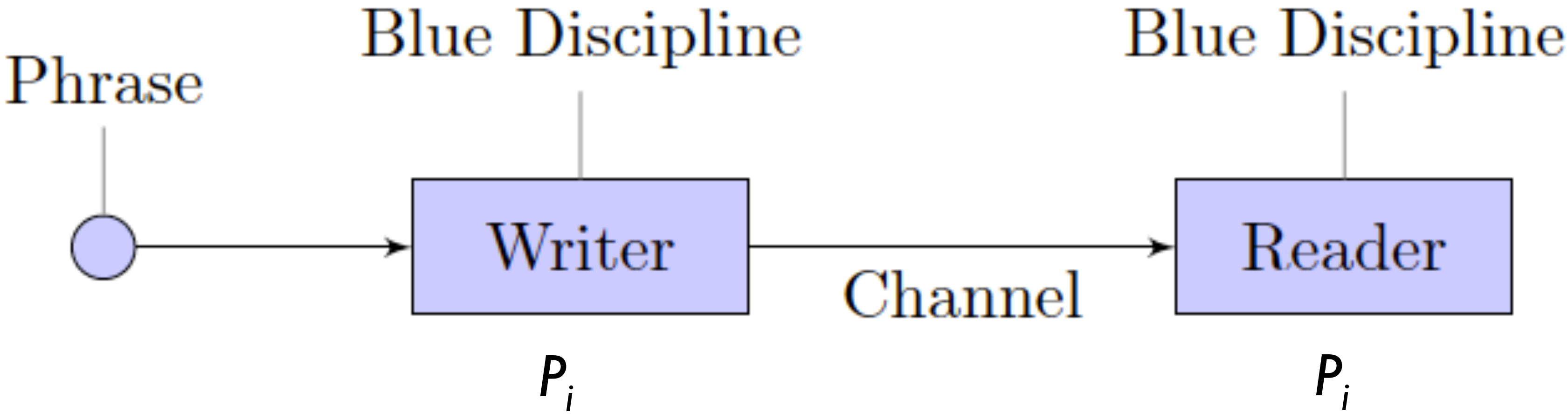
Navigating Hierarchical Knowledge Networks

1. Congruent Landscape	**2. Incongruent Landscape**	**3. Congruent Abstract**	**4. Incongruent Abstract**	**5. Designer Baseline**
				
Landscape visualization with data properties mapped to visual elements according to applicable image schemata	Landscape visualization with data properties mapped to visual elements deliberately breaking with image schemata	Identical to the Congruent Landscape tool but with all realistic details and overt "landscape" visuals removed	Identical to the Incongruent Landscape tool but with all realistic details and overt "landscape" visuals removed	Visualization designed by a hypothesis-blind designer attempting to make an effective visualization but without special emphasis on metaphor

The jargon barriers of science



The Landscape of Modern Mathematics



X ~ space of all phrases

P_i ~ probability distribution over χ_i with values $x \in \boldsymbol{X}$

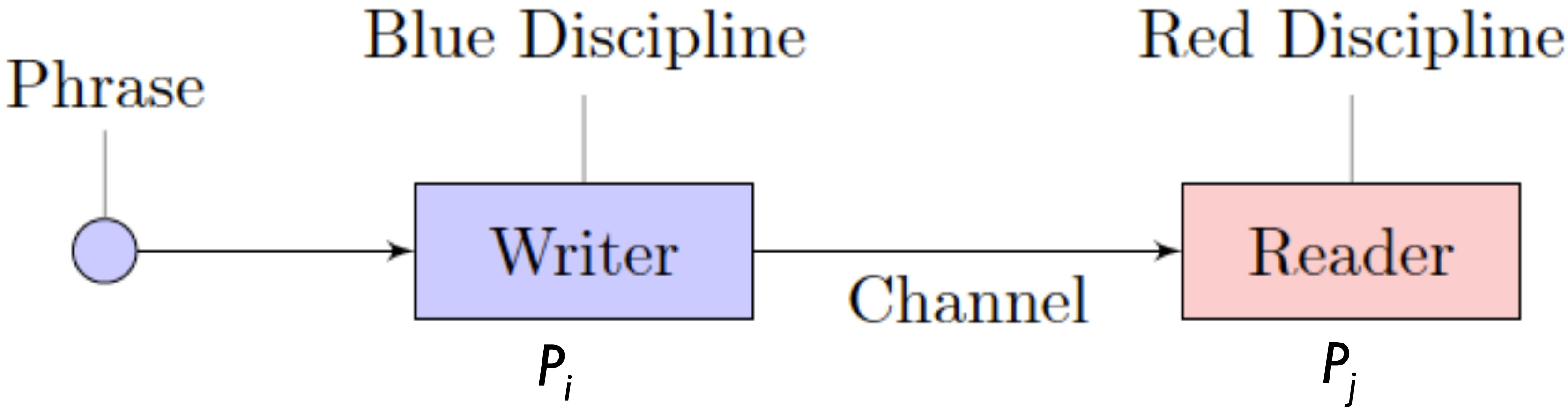
- writer chooses phrases with probability $p_i(x)$

- optimal codeword has length $-\log_2 p_i(x)$

expected message length $H(X_i) = -\sum_{x \in \mathcal{X}} p_i(x) \log_2 p_i(x)$

assumption: language of each scientific field is *optimized* based on frequency of phrases

Vilhena (2014) Sociological Science



cross entropy

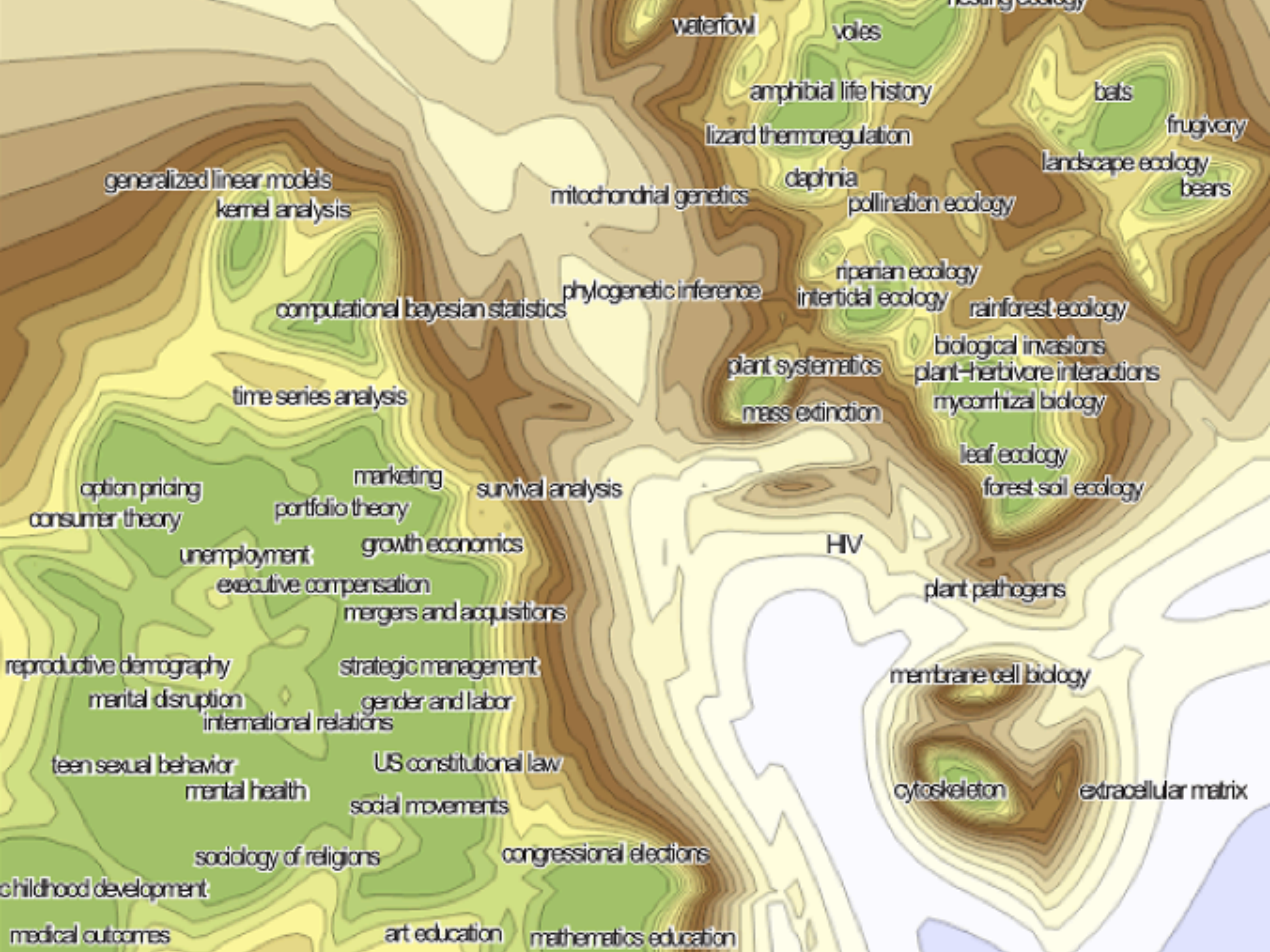
expected message length: $Q(p_i||p_j) = -\sum_{x \in \mathcal{X}} p_i(x) \log_2 p_j(x)$

efficiency of communication

$$E_{ij} = \frac{H(X_i)}{Q(p_i||p_j)} = \frac{-\sum_{x \in \mathcal{X}} p_i(x) \log_2 p_i(x)}{-\sum_{x \in \mathcal{X}} p_i(x) \log_2 p_j(x)}$$

$$C_{ij} = 1 - E_{ij}$$

cultural hole


waterfowl
voles
amphibial life history
bats
frugivory
lizard thermoregulation
landscape ecology
generalized linear models
daphnia
bears
mitochondrial genetics
pollination ecology
kernel analysis
riparian ecology
phylogenetic inference
intertidal ecology
rainforest ecology
computational bayesian statistics
biological invasions
plant systematics
plant–herbivore interactions
time series analysis
mycorrhizal biology
mass extinction
leaf ecology
marketing
option pricing
survival analysis
forest soil ecology
portfolio theory
consumer theory
HIV
growth economics
unemployment
executive compensation
plant pathogens
mergers and acquisitions
reproductive demography
strategic management
membrane cell biology
marital disruption
gender and labor
international relations
teen sexual behavior
US constitutional law
mental health
cytoskeleton
extracellular matrix
social movements
sociology of religions
congressional elections
childhood development
medical outcomes
art education
mathematics education

Challenges: Labeling

$$\text{label}(\mathcal{C}_i) = \mathcal{N}_j \text{ where } j = \underset{j}{\text{argmax}}\, \hat{I}\big(f(\varphi, i), g(\varphi, j)\big)$$

$$\hat{I}\Big(f_i(\varphi), g_j(\varphi)\Big) = \frac{I\big(f_i(\varphi), g_j(\varphi)\big)}{H\big(f_i(\varphi)\big)}$$

Normalized Mutual Information

Semantic Scholar

- Babel (data and recommendation)
- InfoMap (clustering citation networks)
- Scholar profiles (share data)
- Auto-detecting neuroscience
- Viziometrics (labeling tools)
- Visual interface design
- Auto-Labeling Fields
- Data cleaning (author disambiguation)

Acknowledgements

Carl Bergstrom, Department of Biology, University of Washington

Martin Rosvall, Department of Physics, Umea University

Ian Wesley-Smith, Information School, University of Washington

Jason Portenoy, Information School, University of Washington

Bill Howe, eScience, CSE, University of Washington

Poshen Lee, CSE, University of Washington