The MESUR project: studying science from usage data

Implications for scholarly impact metrics

Johan Bollen - jbollen@indiana.edu

Indiana University
School of Informatics and Computing
Center for Complex Networks and Systems Research
Cognitive Science Program

April 9, 2010

Outline

1 Introduction
- Problem statement
- Usage data

2 MESUR
- MESUR overview
- Creating MESUR's reference data set

3 Mapping Science

4 Metrics Survey

5 Services

6 Discussion
- Overview
- Future Research
- Relevant papers

Why study science?

Tremendous importance

Enormous amount of resources and people involved:

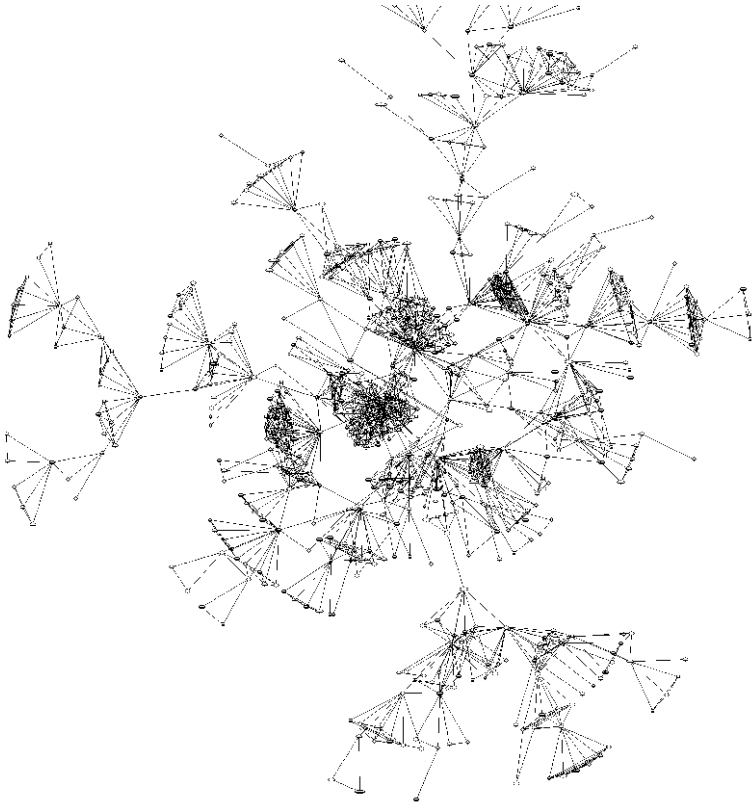
Allocation of resources: funding agencies, policy makers, the public, corporations, the scientists themselves

Social systems research: Ability to learn about general properties of similar social systems

Unrelated picture of my daughter

Science as a social system

- Actors: scientists, stakeholders
- Relations (often focused on exchange of knowledge)
  - Informal interactions: meetings, workshops, conversations, messages, etc
  - Formal: affiliations, collaborations, projects, publications
- Artifacts: articles, journals, reports, data, software



Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. Information Processing and Management, 41(6):14621480, December 2005

The measure of science in the print-era

Gold standard of measuring science: citation data

Extracted from "publication" data, so we know it's good (right?)

The measure of science in the print-era

Gold standard of measuring science: citation data

Extracted from "publication" data, so we know it's good (right?)

- Consequence of print economy:
  - Print-era: printed material (on "paper"! In "libraries")
  - Acknowledgement of influence: citations
  - More citations, more influence
  - Citation data: measurement of flow of influence and activity

The measure of science in the print-era

Gold standard of measuring science: citation data

Extracted from "publication" data, so we know it's good (right?)

- Consequence of print economy:
  - Print-era: printed material (on "paper"! In "libraries")
  - Acknowledgement of influence: citations
  - More citations, more influence
  - Citation data: measurement of flow of influence and activity
- Entire citation-based industry:
  - End-user services
  - Bibliometrics

Two problems with present citation-based approach

Data (1) and metrics (2)

Issues:

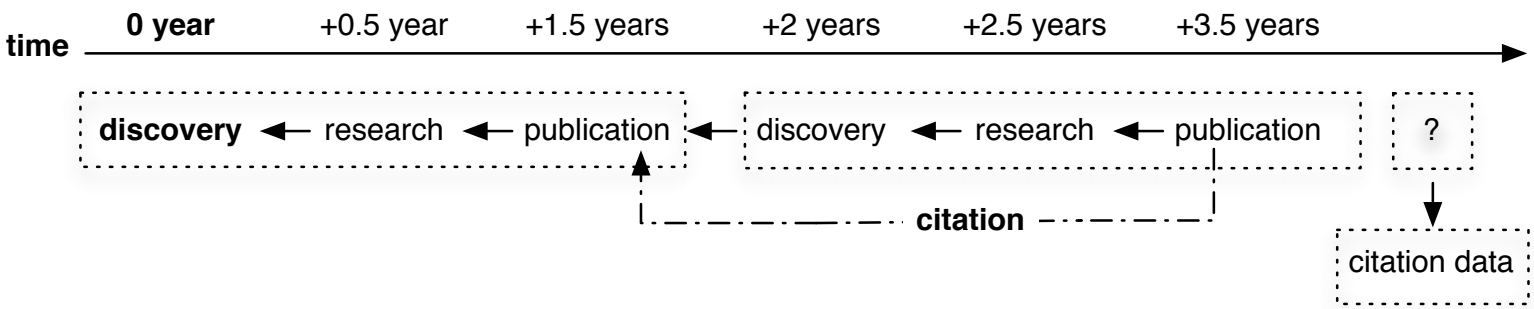
1. Inherent features of citation data due to its origins in published material
2. Existing metrics fail to acknowledge (1), and ignore network properties of science as a social system.

Citation data

Delays and partiality

Citation problem 1: Data

Citation data is late and partial indicator of scientific activity



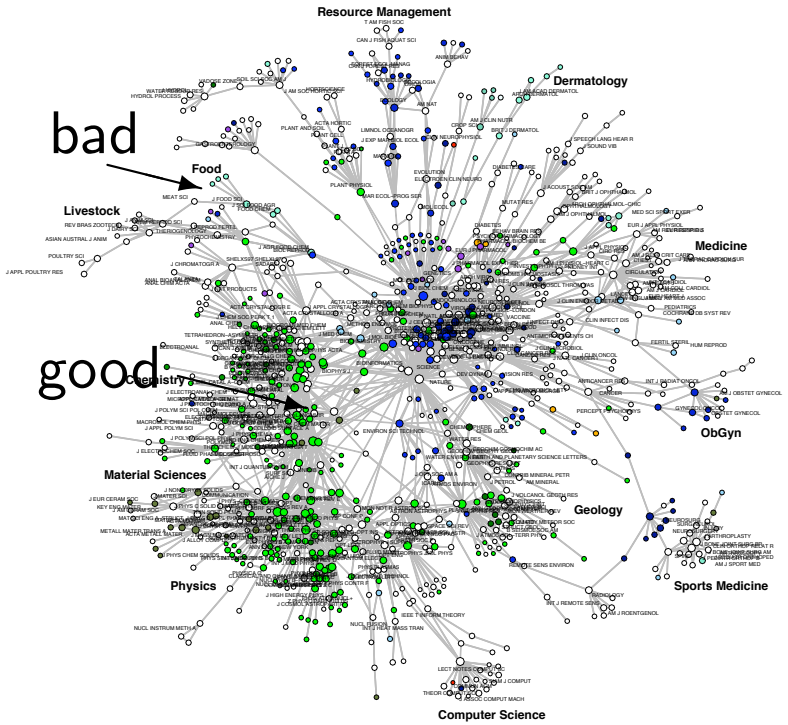
- Publication delays
- Publication$_t$ $\longrightarrow$ publication$_{t-1}$ $\longrightarrow$ citation DB
- Domain-dependent practices
- Community: Publishing authors only

Citation data

Metrics vs. networks

Citation problem 2: networks

Ignoring network properties of science



Measuring impact, influence, prestige from citation data:

- More citations is better than less citations
- In citation network, central position is better than outskirts.

Citation problem no. 2

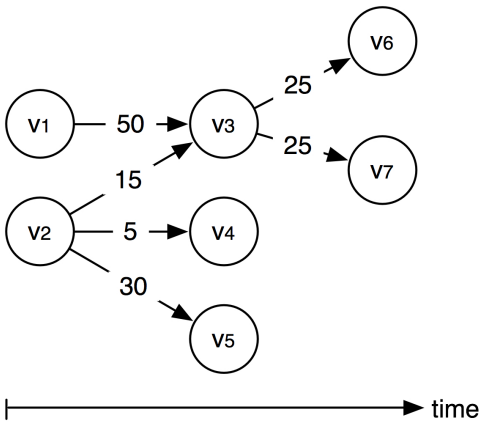
Metrics vs. networks

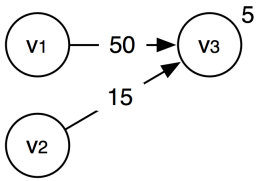
This approach is epitomized in Thomson-Reuters Impact Factor:

Citation problem no. 2

Metrics vs. networks

This approach is epitomized in Thomson-Reuters Impact Factor:

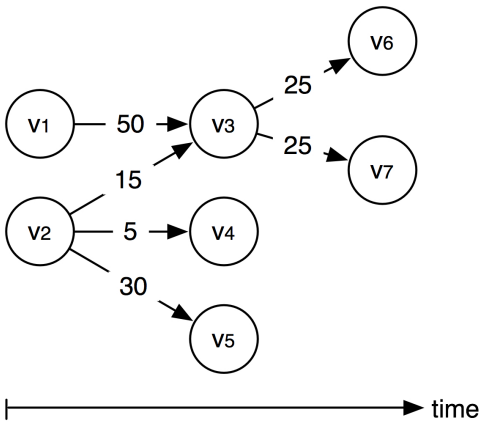




Citation problem no. 2

Metrics vs. networks

This approach is epitomized in Thomson-Reuters Impact Factor:



Standard citation graph representation

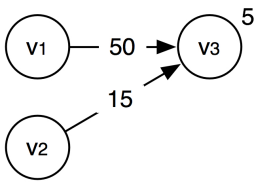
$G = (V, E, W)$

$E \subseteq V^2$

$W : E \rightarrow \mathbb{N}^+$

However:

$\forall (n_i, n_j) \in E : \text{pubtime}(n_i) > \text{pubtime}(n_j)$



Impact Factor: normalized in-degree

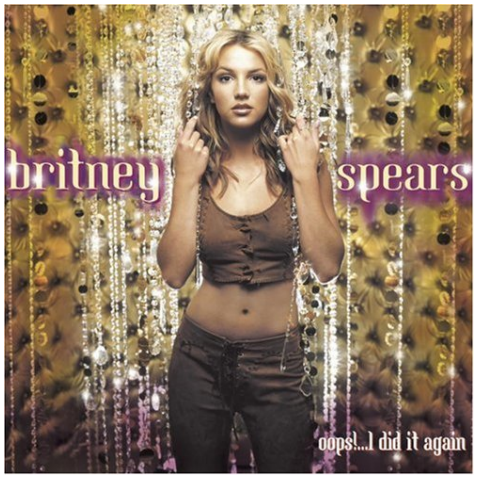
$$\text{IF}_j = \frac{\sum_i w_{ij}}{N_j}$$

Citation problem no. 2

Between Big Star and Britney Spears

Is more always better?

Why context matters

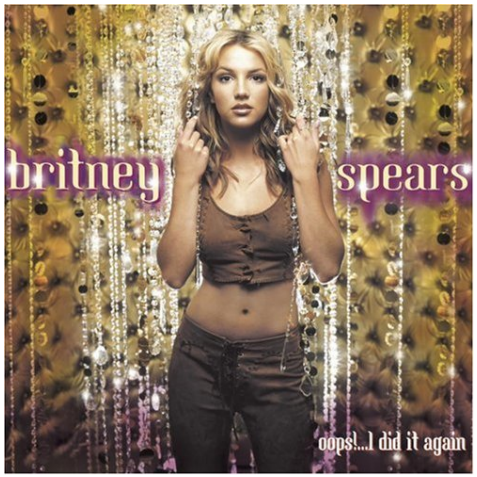


Citation problem no. 2

Between Big Star and Britney Spears

Is more always better?

Why context matters



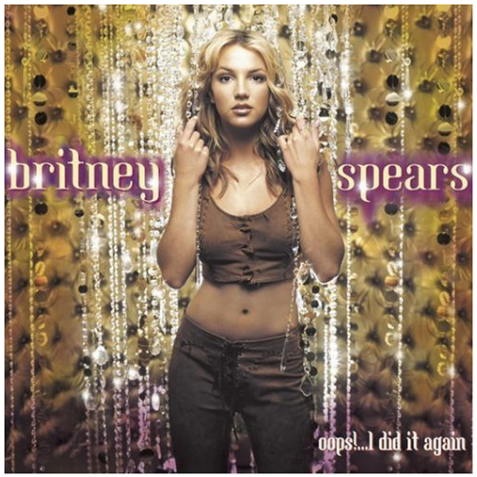
- Sold 50M records
- Influenced who?
- High popularity, low prestige.

Citation problem no. 2

Between Big Star and Britney Spears

Is more always better?

Why context matters



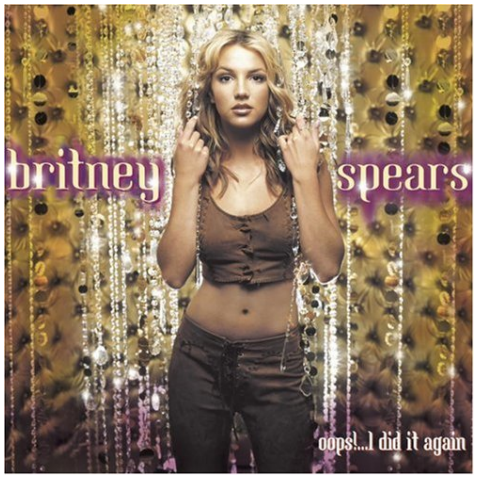
- Sold 50M records
- Influenced who?
- High popularity, low prestige.

Citation problem no. 2

Between Big Star and Britney Spears

Is more always better?

Why context matters



- Sold 50M records
- Influenced who?
- High popularity, low prestige.

- Sold 50k records
- Influenced REM, B52s, Teenage Fanclub, etc.
- Low popularity, high influence

Citation problem no. 2

Between Big Star and Britney Spears

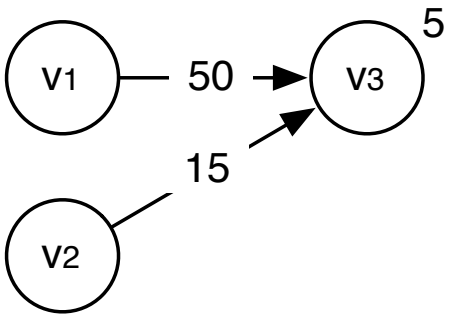
Silly? That's how we evaluate scientific impact right now!

The Impact Factor and lots of other citation-based metrics are presently being used to assess the quality of publications, journals, authors, institutions, and even entire countries by proxy.

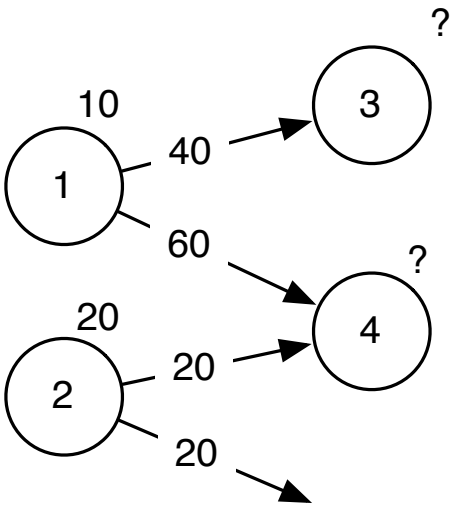
Note: Lots of developments on better metrics (cf. Eigenfactor: Bergstrom and Rosvall), but still reliance on citation data.

New developments

PageRank for citation graphs



Impact Factor: $\mathrm{IF}_j = \frac{\sum_i w_{ij}}{N_j}$:
normalized citation count
origin of citation disregarded



A different route:

- $IF(v_i,) \simeq \lambda \sum_j IF_j$
- $IF(v_i,) \simeq \lambda \sum_j IF_j \times \frac{1}{O(v_j)}$
- $PR(v_i) \simeq \lambda \sum_j PR(v_j) \times \frac{1}{O(v_j)}$
- $PR(v_i) \simeq \frac{(1-\lambda)}{N} + \lambda \sum_j PR(v_j) \times \frac{1}{O(v_j)}$
- $PR_w(v_i) = \frac{(1-\lambda)}{N} + \lambda \sum_j PR_w(v_j) \times w(v_j, v_i)$

New developments, contd

PageRank for citation graphs

	ISI IF		PR_w	
rank	value	Journal	value (x 10^3)	Journal
1	52.28	ANNU REV IMMUNOL	16.78	NATURE
2	37.65	ANNU REV BIOCHEM	16.39	J BIOL CHEM
3	36.83	PHYSIOL REV	16.38	SCIENCE
4	35.04	NAT REV MOL CELL BIO	14.49	P NATL ACAD SCI USA
5	34.83	NEW ENGL J MED	8.41	PHYS REV LETT
6	30.98	NATURE	5.76	CELL
7	30.55	NAT MED	5.70	NEW ENGL J MED
8	29.78	SCIENCE	4.67	J AM CHEM SOC
9	28.18	NAT IMMUNOL	4.46	J IMMUNOL
10	28.17	REV MOD PHYS	4.28	APPL PHYS LETT

Table: The highest ranking journals according to ISI IF and Weighted PageRank (JCR2003)

Two problems with present citation-based approach

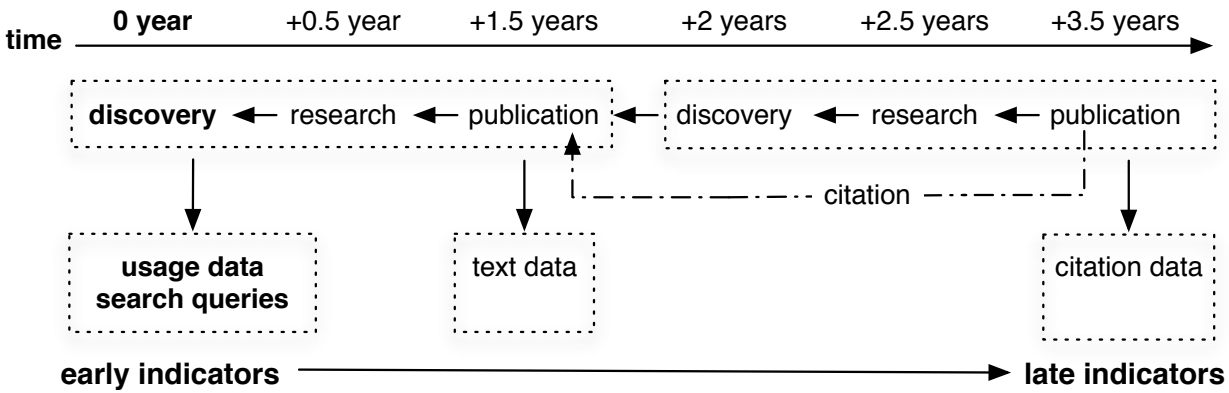
Data (1) and metrics (2)

Issues:

1. Inherent features of citation data due to its origins in published material
2. Existing metrics fail to acknowledge (1), and ignore network properties of science as a social system.

Post-print, online era, aka known as "today": usage data?

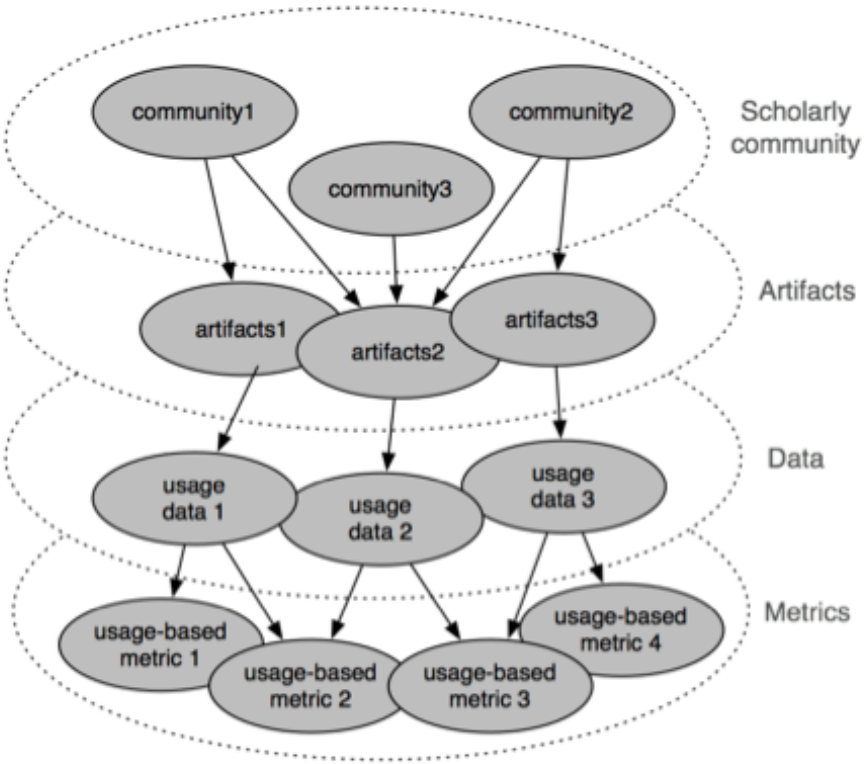
Most use of scholarly resources is now mediated by online services



- Recorded by nearly every scholarly service
- Captures early phases of scientific activity
- Includes a wide variety of resource types
- Activities of larger scientific communities, e.g. not limited only to authors
- Scale: cf. Elsevier announced 1B downloads in 2006 vs. 650M citation in WoS.

Issues with usage data

Lots of interest in leveraging usage data for models of science, tracking scientific activity, metrics, etc. Cf. COUNTER, Ex Libris bX, and many more.



However many different issues:

- Silo: recorded for *particular service* for a *particular user community*
- Lack of standards: recorded in different manner, for different systems
- Lack of research: how leverage usage data for scientific tracking, modeling, metrics, etc.

Outline

1 Introduction
- Problem statement
- Usage data

2 MESUR
- MESUR overview
- Creating MESUR's reference data set

3 Mapping Science

4 Metrics Survey

5 Services

6 Discussion
- Overview
- Future Research
- Relevant papers

MESUR project: survey the potential of usage data at very large-scale

Studying scientific activity

MESUR project: survey the potential of usage data at very large-scale

Studying scientific activity

- Can we model and study patterns of scientific activity from large-scale, representative usage data?

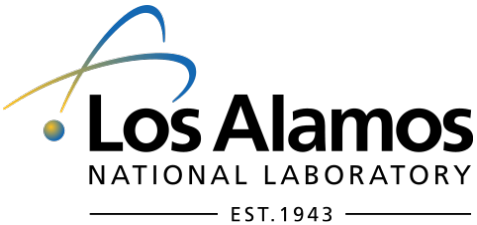
MESUR project: survey the potential of usage data at very large-scale

Studying scientific activity

- Can we model and study patterns of scientific activity from large-scale, representative usage data?
- What can we learn about impact and structure from patterns of scientific activity from usage data?

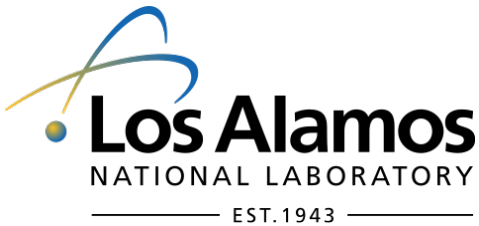
MESUR history

The Andrew W. Mellon Foundation



MESUR history

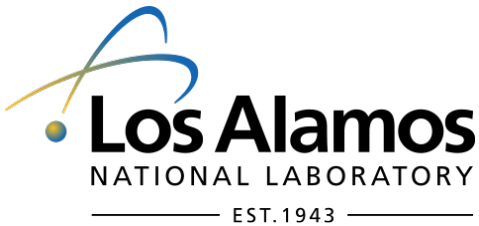
The Andrew W. Mellon Foundation



- 2006-2008: Andrew W. Mellon Foundation

MESUR history

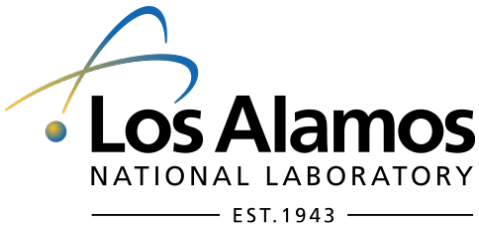
The Andrew W. Mellon Foundation



- 2006-2008: Andrew W. Mellon Foundation
- 2008-2009: Los Alamos National Laboratory

MESUR history

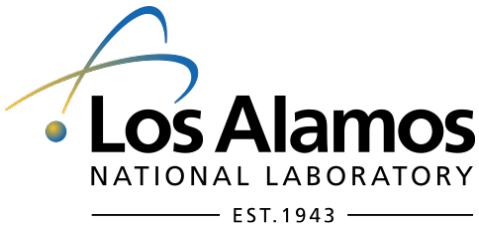
The Andrew W. Mellon Foundation



- 2006-2008: Andrew W. Mellon Foundation
- 2008-2009: Los Alamos National Laboratory
- 2009-: Indiana University, School of Informatics and Computing

MESUR history

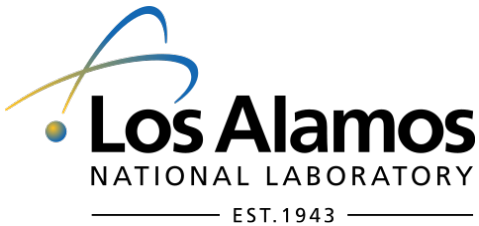
The Andrew W. Mellon Foundation



- 2006-2008: Andrew W. Mellon Foundation
- 2008-2009: Los Alamos National Laboratory
- 2009-: Indiana University, School of Informatics and Computing
- 2009-2013: National Science Foundation

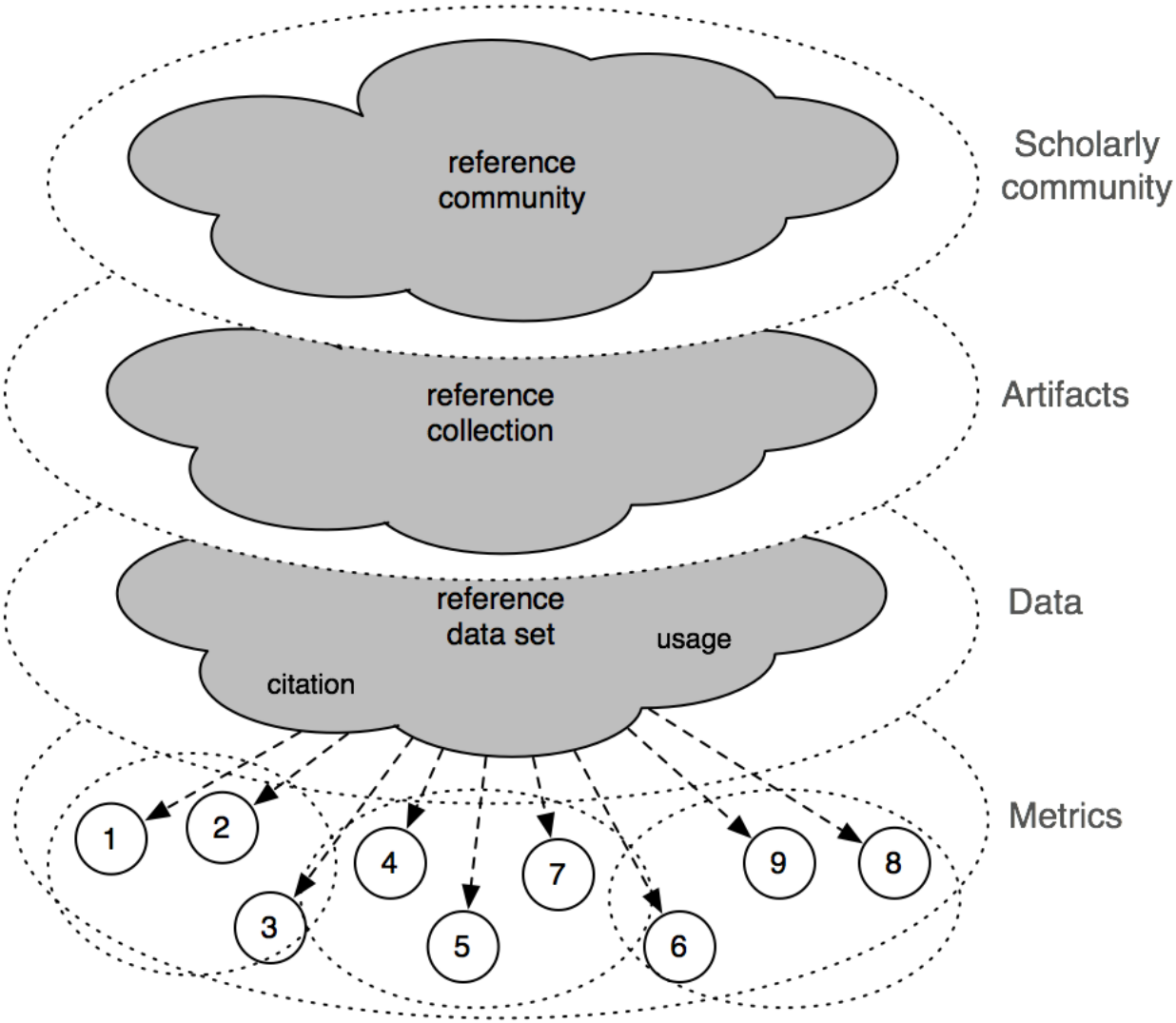
MESUR history

The Andrew W. Mellon Foundation

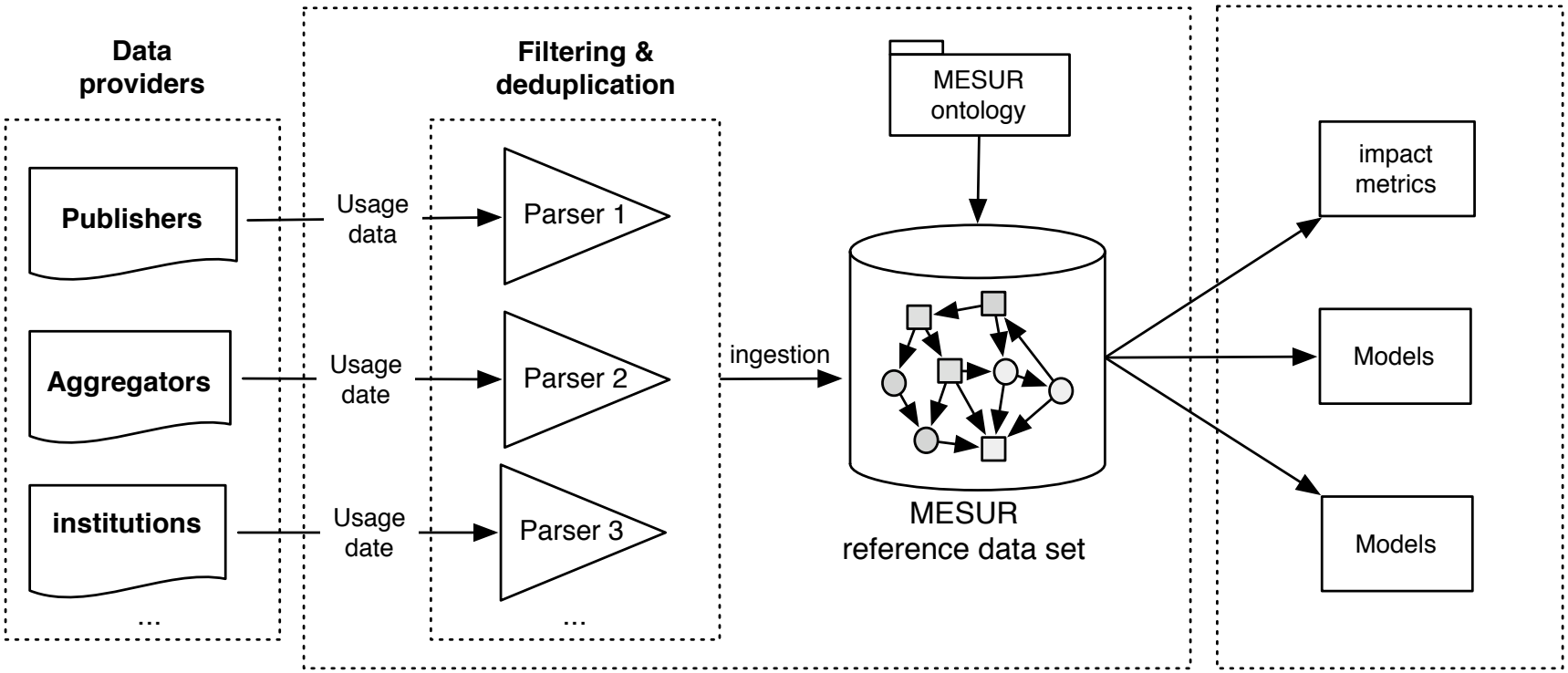


- 2006-2008: Andrew W. Mellon Foundation
- 2008-2009: Los Alamos National Laboratory
- 2009-: Indiana University, School of Informatics and Computing
- 2009-2013: National Science Foundation
- **Team: PI, co-PI, 2 scientists, 2 full-time developers, and 1 PhD student.**

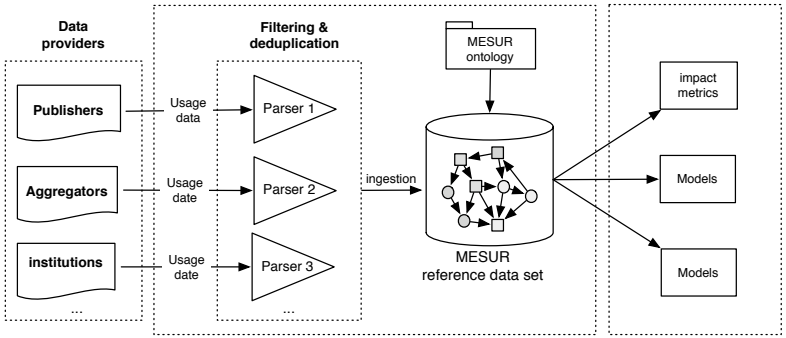
MESUR objective

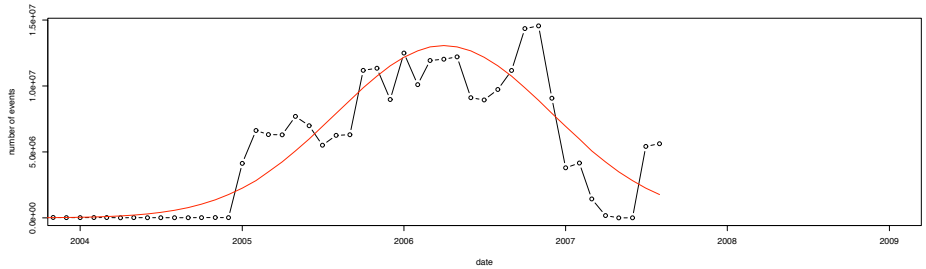


MESUR dataflow



Creating MESUR's data set





Providers 2006-20010: BMC, Blackwell, UC, CSU (23), EBSCO, ELSEVIER, EMERALD, INGENTA, JSTOR, LANL, MIMAS/ZETOC, THOMSON, UPENN (9), UTEXAS

events **1,000,000,000 usage events**

citations +500,000,000 citations,

articles and serials +50M articles, +-100,000 serials

Data requirements

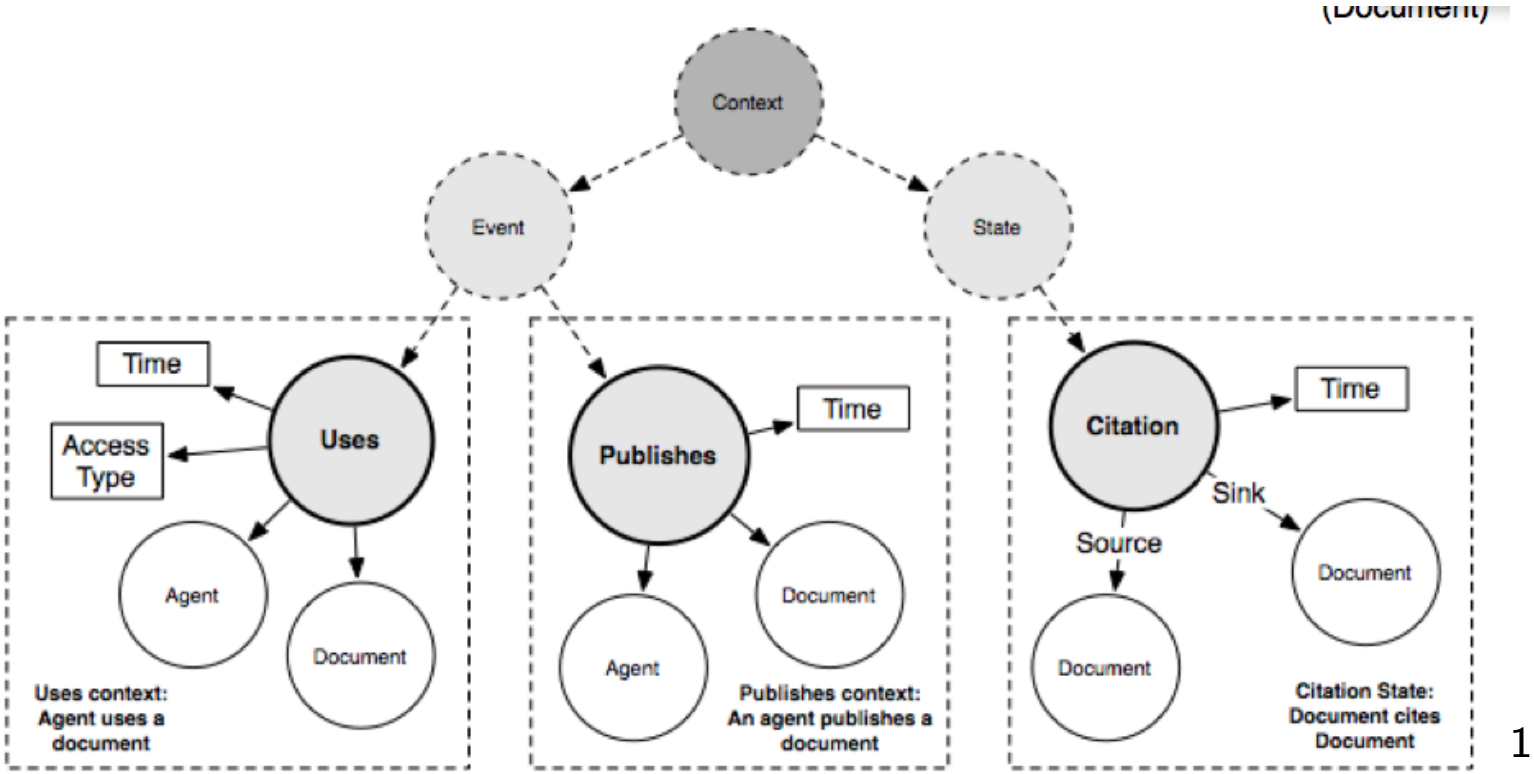
```
10.1186/1742-7622-1-4,Emerg Themes Epidemiol,PDF,01/05/2006 00:01:54,32F854DABC15C0FF70822AD1EBFD852C,,C3044832
10.1186/1477-7827-2-12,Reprod Biol Endocrinol,FULLTEXT,01/05/2006 00:01:56,CD3B42EA2299BD8241FB9A3C5AA4864D,,C3041834
10.1186/1741-7015-3-13,BMC Med,FULLTEXT,01/05/2006 00:01:58,2ED28AF4D204D1E75761A18D74246B3D,,C3041B35
10.1186/1476-7120-3-35,Cardiovasc Ultrasound,FULLTEXT,01/05/2006 00:01:58,AAE78CE04B89010156DC4A4702D6864F,,C3033446
10.1186/1743-0003-1-8,J NeuroEngineering Rehabil,FULLTEXT,01/05/2006 00:01:59,AD8DF686CEA39C65551297EACDA1D077,,C3040416
10.1186/1471-2164-5-13,BMC Genomics,FULLTEXT,01/05/2006 00:01:59,002D0E9F864278FC49DE0A57C069003D,,C3034954
10.1186/1477-5956-3-2,Proteome Sci,FULLTEXT,01/05/2006 00:02:01,FD8E014AFCEACD8EE033E80569419636,,C3031E15
10.1186/1471-2164-7-81,BMC Genomics,JUSTPUBLISHEDABSTRACT,01/05/2006 00:02:03,FD33B664F54D05B8B6C4A388650FC60F,,C304372C
10.1186/1475-2859-4-12,Microb Cell Fact,FULLTEXT,01/05/2006 00:02:03,DC884A02714D4B33A9E2ADA4FFE083E6,,C3043C26
10.1186/jbiol29,J Biol,FULLTEXT,01/05/2006 00:02:03,AC34C03556ADFB8E0A51D041FCA667D3,,C3034657
10.1186/gb-2000-2-1-reports4001,Genome Biol,FULLTEXT,01/05/2006 00:02:04,4284F454A9FAD032A1C20D90369D9884,,C3040142
10.1186/1471-2458-5-87,BMC Public Health,FULLTEXT,01/05/2006 00:02:04,917CE67A667693D29618899BCB14F697,,C306060D
10.1186/1471-2318-5-14,BMC Geriatr,PDF,01/05/2006 00:02:05,46F1D2204CB2CF88CAB621D416793C12,,
10.1186/ar1951,Arthritis Res Ther,FULLTEXT,01/05/2006 00:02:08,AC34C03556ADFB8E0A51D041FCA667D3,,C3045849
10.1186/1744-8069-2-4,Mol Pain,FULLTEXT,01/05/2006 00:02:09,354825729DD9B3BCE22E2832C6DD9ECB,612D0D60444356EB6C437EB8D73DA42A,C3041D5A
```

Minimum requirements

- Separate user requests
- Date-time stamp down to second
- Document identifier or sufficient metadata to de-duplicate
- Session ID (or anonymized user ID)
- Request type identifier

`http://tweety.lanl.gov/public/schemas/2007-01/mesur.owl`

MESUR's OWL/RDF ontology



[1]

[1] Rodriguez, Bollen & Van de Sompel. A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage. JCDL07 - Based on OntologyX work.

Outline

1 Introduction
- Problem statement
- Usage data

2 MESUR
- MESUR overview
- Creating MESUR's reference data set

3 Mapping Science

4 Metrics Survey

5 Services

6 Discussion
- Overview
- Future Research
- Relevant papers

Reference data set

Subsetting

Domain	Usage	UC Degrees	JCR
Natural Science	37%	39%	92.8%
Social Sciences	45%	46%	7.2%
Humanities	14%	15%	

Source: `http://www.ucop.edu/ucophome/uwnews/stat/statsum/fall2007/statsumm2007.pdf` (table 9)

Reference data set

Subsetting

Common time period March 1st 2006 - February 1st 2007

Domain	Usage	UC Degrees	JCR
Natural Science	37%	39%	92.8%
Social Sciences	45%	46%	7.2%
Humanities	14%	15%	

Source: `http://www.ucop.edu/ucophome/uwnews/stat/statsum/fall2007/statsumm2007.pdf` (table 9)

Reference data set

Subsetting

Common time period March 1st 2006 - February 1st 2007

Providers: Thomson Scientic (Web of Science), Elsevier (Scopus), JSTOR, Ingenta, University of Texas (9 campuses, 6 health institutions), and California State University (23 campuses)

Domain	Usage	UC Degrees	JCR
Natural Science	37%	39%	92.8%
Social Sciences	45%	46%	7.2%
Humanities	14%	15%	

Source: `http://www.ucop.edu/ucophome/uwnews/stat/statsum/fall2007/statsumm2007.pdf` (table 9)

Reference data set

Subsetting

Common time period March 1st 2006 - February 1st 2007

Providers: Thomson Scientic (Web of Science), Elsevier (Scopus), JSTOR, Ingenta, University of Texas (9 campuses, 6 health institutions), and California State University (23 campuses)

Scale: 346,312,045 usage events, 97,532 serials (many of which not journals)

Domain	Usage	UC Degrees	JCR
Natural Science	37%	39%	92.8%
Social Sciences	45%	46%	7.2%
Humanities	14%	15%	

Source: `http://www.ucop.edu/ucophome/uwnews/stat/statsum/fall2007/statsumm2007.pdf` (table 9)

Article clickstreams

usage data log: $U = \{u_1, u_2, \cdots, u_n\}$

$u = \{s, t, a\}$

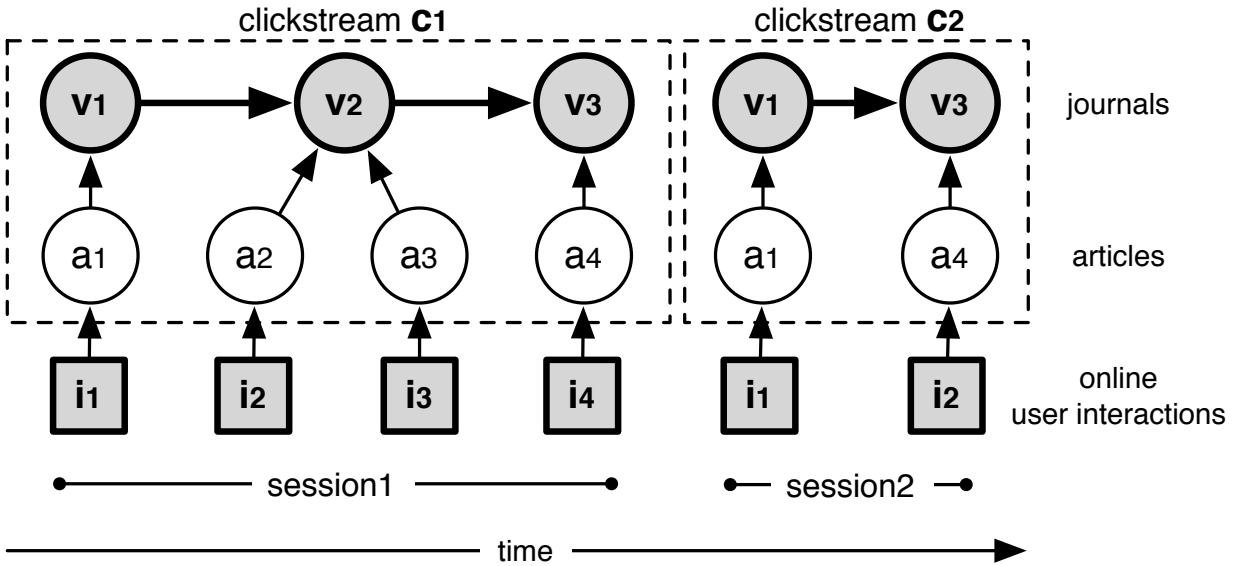
s = session identifier , t = a date-time and a = article

$f \in F$ = clickstreams extracted from U

$f \subset U$, where

$f = (\forall u \in U, \exists s : s(u) \wedge t(u_i) < t(u_{i+1}))$

$s(u)$ and $t(u)$: session identifier and date-time of interaction u.

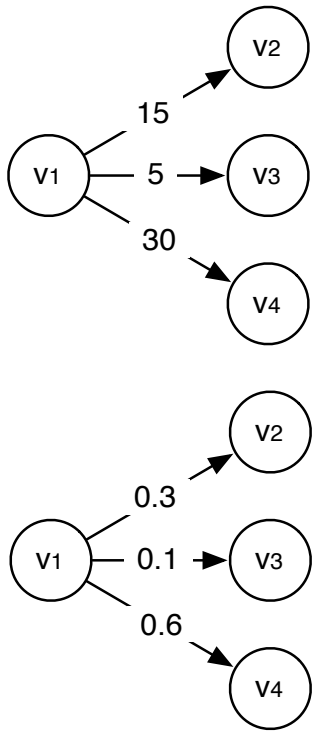


Journal Clickstreams

article clickstream: $f_a = (a_1, a_2, \cdots, a_k)$
journal clickstream: $f_v = (v_1, v_2, \cdots, v_k)$
We observe: $N(v_i, v_j)$
for all pairs (v_i, v_j) in which $j = i + 1$
From which follows:

$$P(v_i, v_j) = \frac{N(v_i, v_j)}{\sum_j N(v_i, v_j)}$$

and
M whose entries $m_{i,j} = P(v_i, v_j)$.



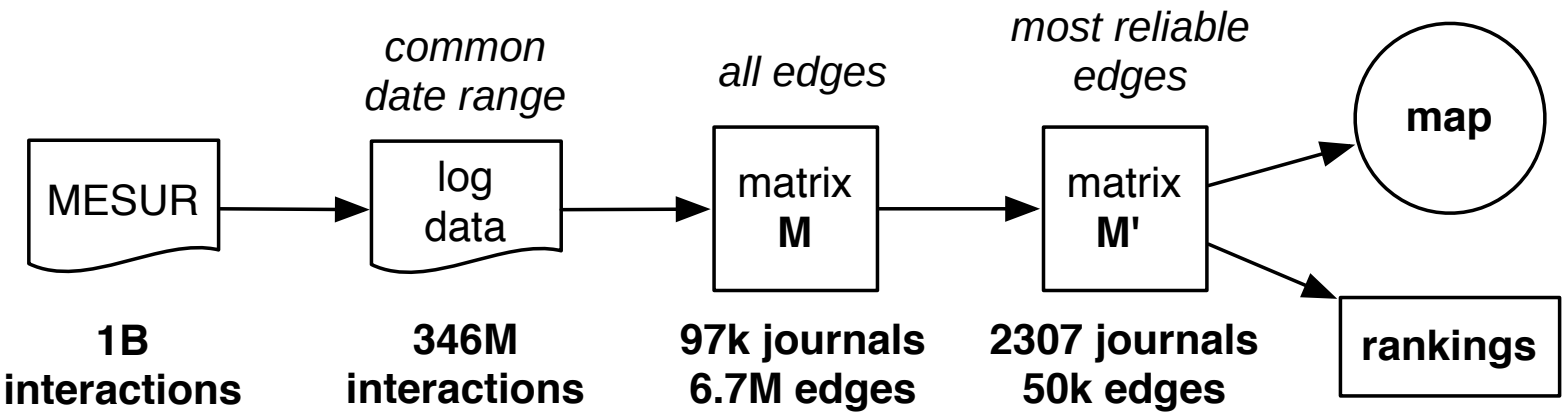
Examples of prominent connections

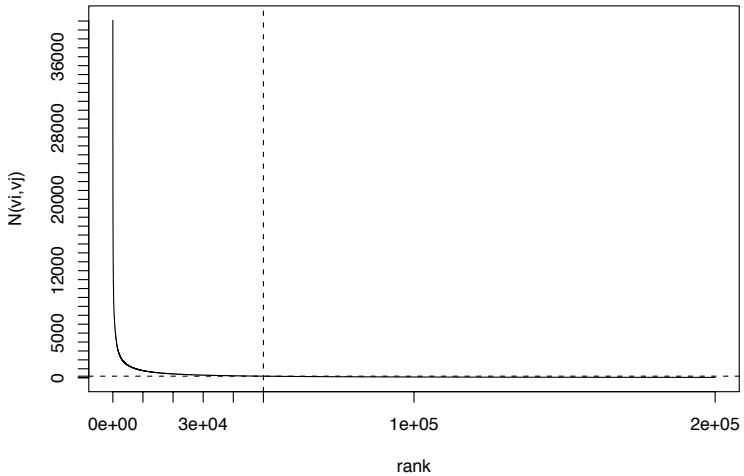
v_i	v_j	$p(v_i, v_j)$	$N(v_i, v_j)$	$N(v_i)$
American Journal of International Law	International Organization	0.0207	9,292	448,034
	International Affairs	0.0184	8,254	
	International and Comparative Law Quarterly	0.0171	7,654	
	Foreign Policy	0.0167	7,500	
	American Political Science Association	0.0140	6,291	
Journal of Educational Sociology	American Journal of Sociology	0.0334	2,790	83,419
	Journal of Higher Education	0.0303	2,529	
	Journal of Negro Education	0.0286	2,389	
	American Sociological Review	0.0276	2,303	
	Social Forces	0.0249	2,076	
Surface Science	Physical Review B	0.0704	2,555	36,282
	Applied Surface Science	0.0341	1,239	
	Physical Review Letters	0.0339	1,230	
	Journal of Chemical Physics	0.0333	1,207	
	Applied Physics Letters	0.0327	1,188	
Journal of Organic Chemistry	Journal of the American Chemical Society	0.0873	4,141	47,439
	Tetrahedron Letters	0.0865	4,105	
	Tetrahedron	0.0602	2,857	
	Organic Letters	0.0532	2,526	
	Angewandte Chemie	0.0305	1,448	
Ecological Applications	Ecology	0.0965	13,659	141,481
	Conservation Biology	0.0524	7,408	
	Bioscience	0.0215	3,043	
	Annual Review of Ecology and Systematics	0.0215	3,043	
	Clinical and Experimental Allergy	0.0191	2,699	
Annals of Mathematics	American Journal of Mathematics	0.0705	5,392	76,526
	American Mathematical Monthly	0.0579	4,432	
	PNAS	0.0156	1,195	
	Econometrica	0.0082	624	
	Mathematics Magazine	0.0077	587	

Network parameters

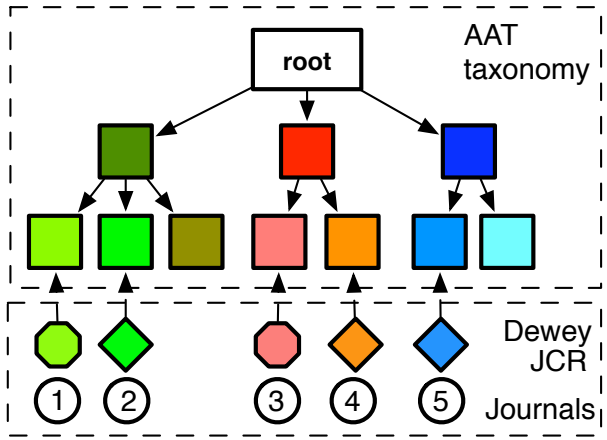
Network matrix

Parameter	M	M'
Journals	97,532	2,307
Edges	6,783,552	50,000
Matrix density	0.071%	0.939%
Strongly Connected Components (SCC)	16,474	236
Journals in SCC	80,934	1,944
Average journal clustering coefficient (SCC)	0.285	0.514
Diameter of largest SCC	37	14



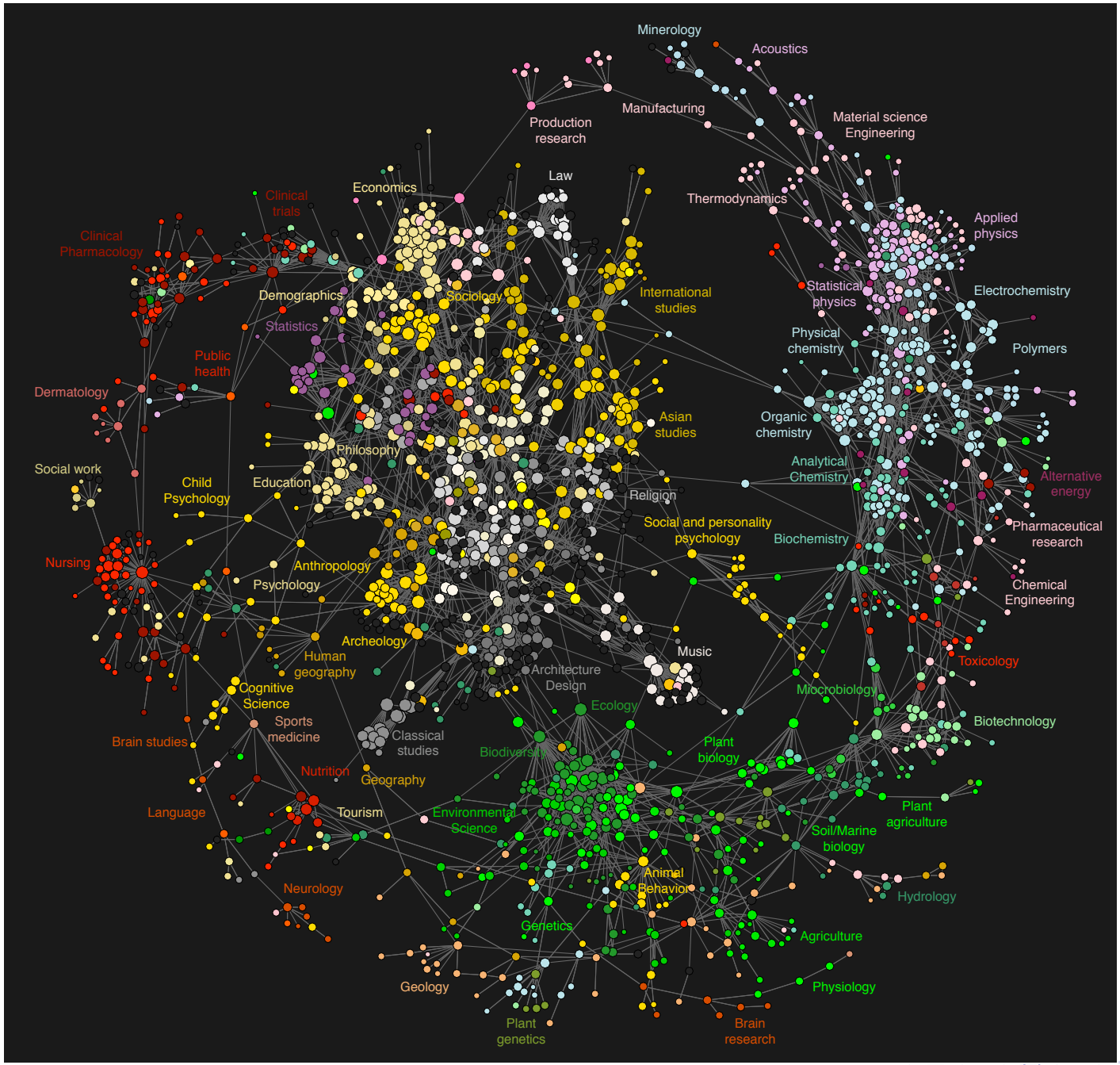


Classification code:



Fruchterman (1991) Graph Drawing by Force-directed

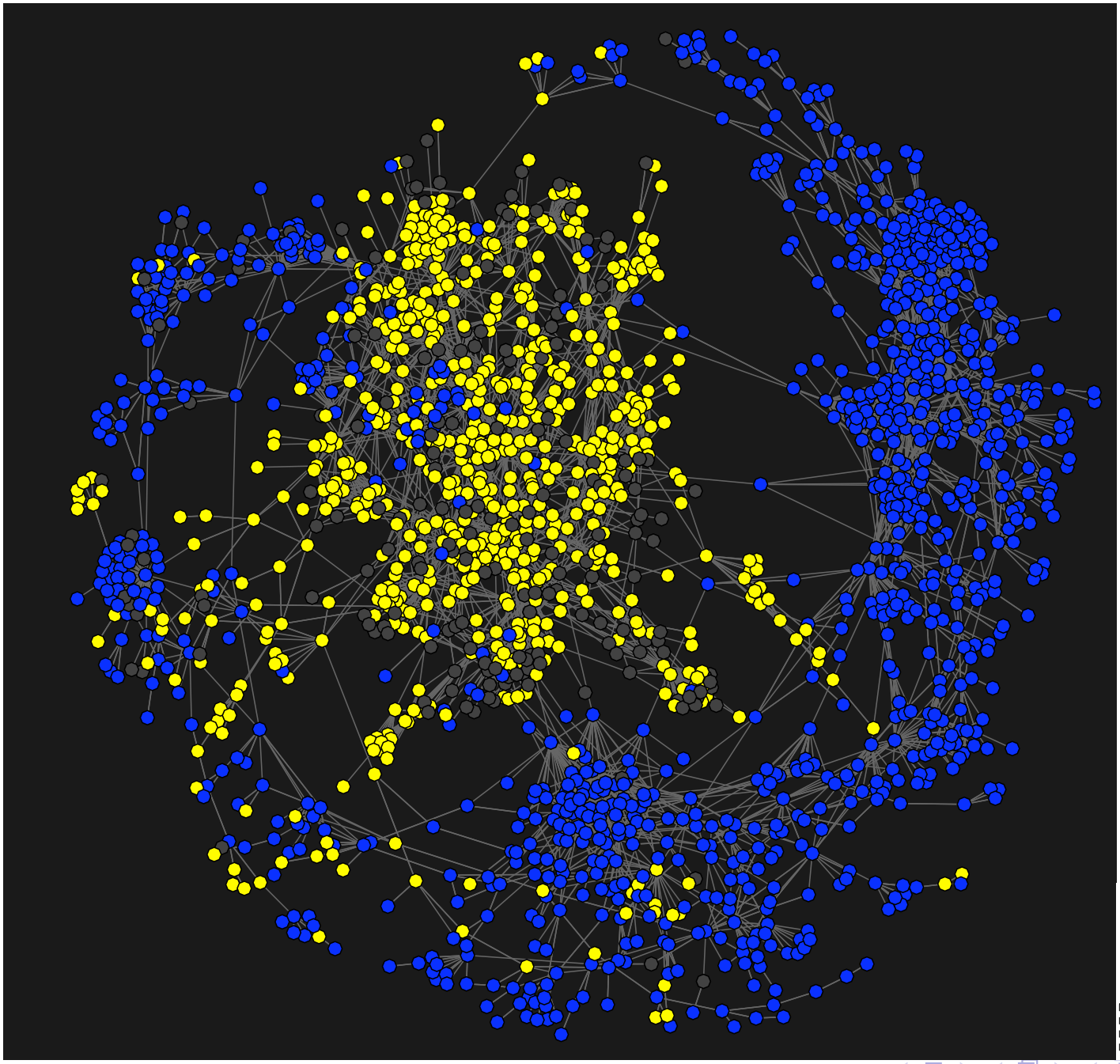
Placement. Software, 21(11):1129-1164

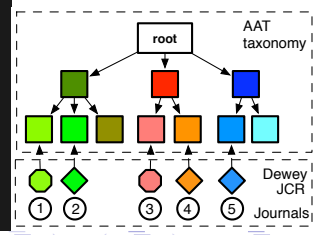

Minerology
Acoustics
Manufacturing
Production research
Material science Engineering
Law
Economics
Clinical trials
Thermodynamics
Applied physics
Clinical Pharmacology
Demographics
Sociology
International studies
Statistical physics
Electrochemistry
Statistics
Physical chemistry
Polymers
Public health
Dermatology
Asian studies
Organic chemistry
Philosophy
Analytical Chemistry
Social work
Child Psychology
Education
Religion
Alternative energy
Social and personality psychology
Biochemistry
Pharmaceutical research
Nursing
Anthropology
Psychology
Chemical Engineering
Archeology
Music
Human geography
Toxicology
Architecture
Design
Cognitive Science
Microbiology
Ecology
Biotechnology
Sports medicine
Brain studies
Classical studies
Plant biology
Biodiversity
Nutrition
Geography
Plant agriculture
Language
Tourism
Environmental Science
Soil/Marine biology
Animal Behavior
Neurology
Hydrology
Genetics
Agriculture
Geology
Physiology
Plant genetics
Brain research

The Arts and Architecture Thesaurus (Getty Research Center)

Table: **Distance from AAT root (α) and number of classifications N_c at that level. Each α produces a finer-grained separation of scientific disciplines.**

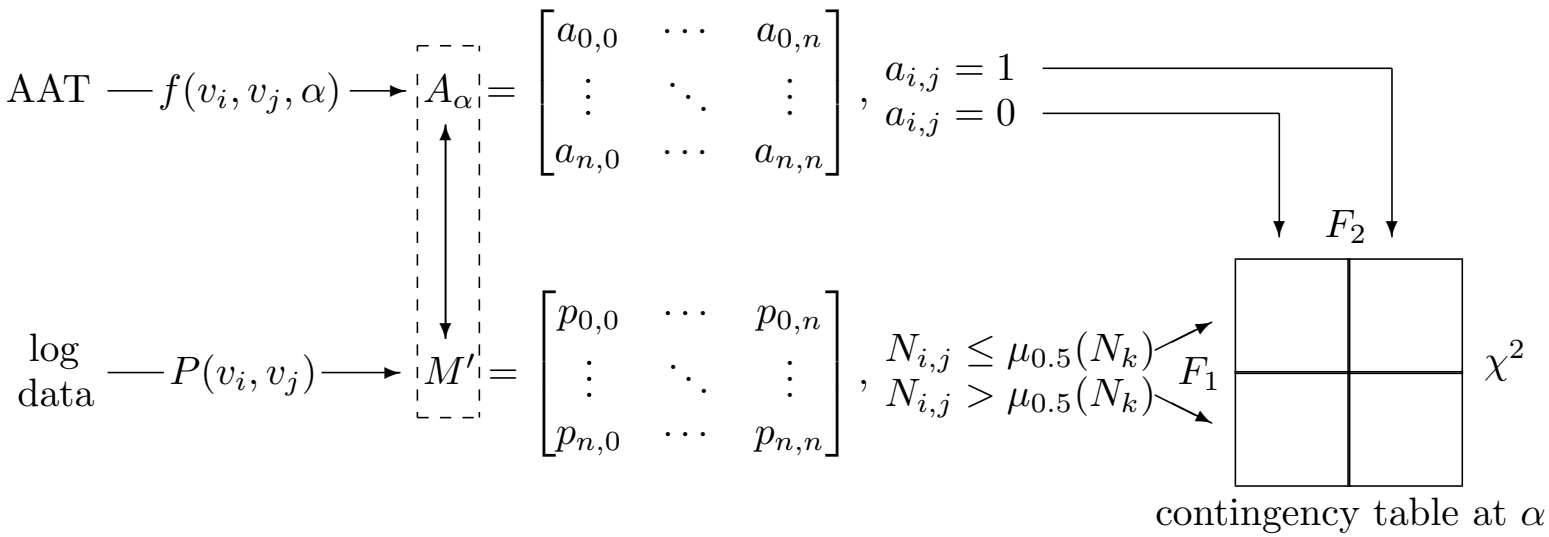
Distance (α)	N_c	Example classifications
1	4	Natural sciences, social sciences, humanities, $\cdots$
2	8	Biology, chemistry, physics, $\cdots$
3	31	Classics, communication, engineering, $\cdots$
4	195	Allergy, anesthesiology, applied linguistics, $\cdots$




root
AAT
taxonomy
Dewey
JCR
Journals
1
2
3
4
5

Cross-validating the clickstream map

$$f(v_i, v_j, \alpha) = \begin{cases} 1 & C_\alpha(v_i) = C_\alpha(v_j) \\ 0 & C_\alpha(v_i) \neq C_\alpha(v_j) \end{cases}$$



$\alpha = 1 : p < 0.0001$
$\alpha = 2 : p < 0.0001$
$\alpha = 3 : p < 0.0001$
$\alpha = 4 : p < 0.0001$

Betweenness centrality

$$C_b(v_k) = \sum_{i \neq j \neq k} \frac{\sigma_{i,j}(v_k)}{\sigma_{i,j}} \quad (1)$$

Table: **Ranking of journals from M' according to betweenness centrality.**

Rank	Journal	Top-level AAT classification
1	Science	Natural Sciences
2	Proceedings of the National Academy of Sciences	Natural Sciences
3	Environmental Health Perspectives	Natural Science
4	Chemosphere	Natural Sciences
5	Journal of Advanced Nursing	Natural Sciences
6	Nature	Natural Sciences
7	Ecology	Natural Sciences
8	Milbank Quarterly	Natural Sciences
9	Applied and Environmental Microbiology	Natural Sciences
10	Child Development	Social Sciences
11	Behavioral Ecology and Sociobiology	Social Sciences
12	Journal of Colloid and Information Science	Natural Sciences
13	American Anthropologist	Social Sciences
14	Journal of Biogeography	Natural Sciences
15	Materials Science and Technology	Natural Sciences

PageRank

$$PR(v_i) = \frac{1-\lambda}{N} + \lambda \sum_j \frac{PR(v_j)}{O(v_j)} \quad (2)$$

Table: **Ranking of journals from M' according to PageRank ($\lambda = 0.85$).**

Rank	Journal	Top-level AAT classification
1	Applied Physics Letters	Natural Sciences
2	Journal of Advanced Nursing	Natural Sciences
3	Journal of the American Chemical Society	Natural Sciences
4	Ecology	Natural Sciences
5	Nature	Natural Sciences
6	Physical Review B	Natural Sciences
7	Journal of Applied Physics	Natural Sciences
8	American Economic Review	Social Sciences
9	American Historical Review	Social Sciences
10	Physical Review Letters	Natural Sciences
11	Science	Natural Sciences
12	Langmuir	Natural Sciences
13	Journal of Chemical Physics	Natural Sciences
14	American Anthropologist	Social Sciences
15	Annals of the American Academy of Political and Social Science	Social Science

Outline

1 Introduction
- Problem statement
- Usage data

2 MESUR
- MESUR overview
- Creating MESUR's reference data set

3 Mapping Science

4 Metrics Survey

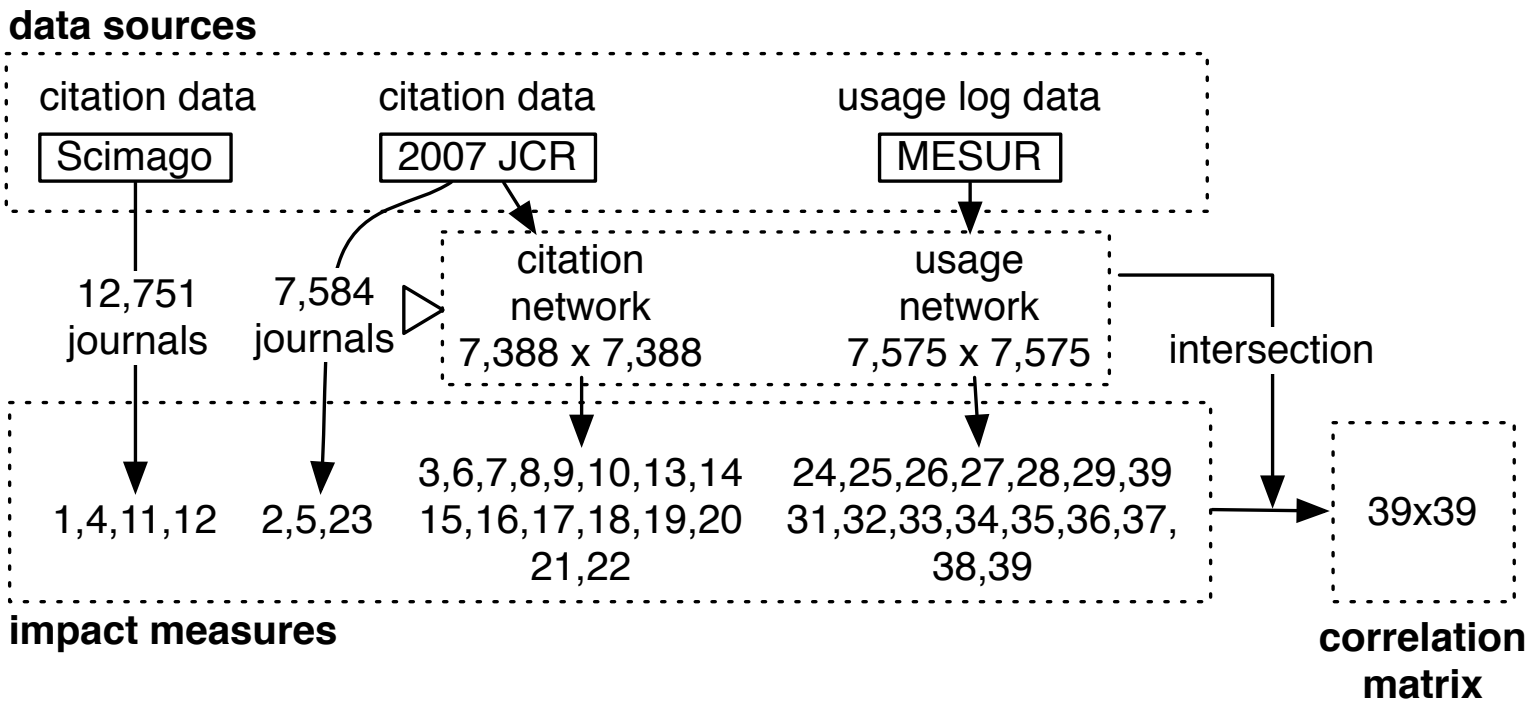
5 Services

6 Discussion
- Overview
- Future Research
- Relevant papers

Metrics

ID	Type	Measure	Source	Network parameters	PC1	PC2	$\bar{\rho}$
1	Citation	Scimago Journal Rank	Scimago/Scopus		-0.974	-8.296	0.556*
2	Citation	Immediacy Index	JCR 2007		1.659	-7.046	0.508*
3	Citation	Closeness Centrality	JCR 2007	Undirected, weighted	0.339	-6.284	0.565*
4	Citaton	Cites per doc	Scimago/Scopus		-1.311	-6.192	0.588*
5	Citation	Journal Impact Factor	JCR 2007		-1.854	-5.937	0.592*
6	Citation	Closeness centrality	JCR 2007	Undirected, unweighted	-1.388	-4.827	0.619
7	Citation	Out-degree centrality	JCR 2007	Directed, weighted	-3.191	-4.215	0.642
8	Citation	Out-degree centrality	JCR 2007	Directed, unweighted	-2.703	-4.015	0.640
9	Citation	Degree Centrality	JCR 2007	Undirected, weighted	-4.850	-2.834	0.690
10	Citation	Degree Centrality	JCR 2007	Undirected, unweighted	-4.398	-2.643	0.691
11	Citation	H-Index	Scimago/Scopus		-3.326	-2.003	0.681
12	Citation	Scimago Total cites	Scimago/Scopus		-4.926	-1.722	0.712
13	Citation	Journal Cite Probability	JCR 2007		-5.389	-1.647	0.710
14	Citation	In-degree centrality	JCR 2007	Directed, unweighted	-5.302	-1.429	0.717
15	Citation	In-degree centrality	JCR 2007	Directed, weighted	-5.380	-1.554	0.712
16	Citation	PageRank	JCR 2007	Directed, unweighted	-4.476	0.108	0.693
17	Citation	PageRank	JCR 2007	Undirected, unweighted	-4.929	0.731	0.726
18	Citation	PageRank	JCR 2007	Undirected, weighted	-4.160	0.864	0.696
19	Citation	PageRank	JCR 2007	Directed, weighted	-3.103	0.333	0.659
20	Citation	Y-factor	JCR 2007	Directed, weighted	-2.971	0.317	0.657
21	Citation	Betweenness centrality	JCR 2007	Undirected, weighted	-0.462	0.872	0.643
22	Citation	Betweenness centrality	JCR 2007	Undirected, unweighted	-0.474	1.609	0.642
23	*Citation*	*Citation Half-Life*	*JCR 2007*		/	/	*0.037*
24	Usage	Closeness centrality	MESUR 2007	Undirected, weighted	3.130	2.683	0.703
25	Usage	Closeness centrality	MESUR 2007	Undirected, unweighted	3.100	3.899	0.731
26	Usage	Degree centrality	MESUR 2007	Undirected, unweighted	3.271	3.873	0.729
27	Usage	PageRank	MESUR 2007	Undirected, unweighted	3.327	4.192	0.728
28	Usage	PageRank	MESUR 2007	Directed, unweighted	3.463	4.336	0.727
29	Usage	In-degree centrality	MESUR 2007	Directed, unweighted	3.463	4.015	0.728
30	Usage	Out-degree centrality	MESUR 2007	Directed, unweighted	3.484	3.994	0.727
31	Usage	PageRank	MESUR 2007	Directed, weighted	3.780	4.217	0.710
32	Usage	PageRank	MESUR 2007	Undirected, weighted	3.813	4.223	0.710
33	Usage	Betweenness centrality	MESUR 2007	Undirected, unweighted	3.988	4.271	0.699
34	Usage	Betweenness centrality	MESUR 2007	Undirected, weighted	3.957	3.698	0.693
35	Usage	Degree centrality	MESUR 2007	Undirected, weighted	5.293	3.528	0.683
36	Usage	Out-degree centrality	MESUR 2007	Directed, weighted	5.302	3.518	0.683
37	Usage	In-degree centrality	MESUR 2007	Directed, weighted	5.286	3.531	0.683
38	Usage	Journal Use Probability	MESUR 2007		8.914	1.833	0.593
39	*Usage*	*Usage Impact Factor*	*MESUR 2007*		/	/	*0.279*

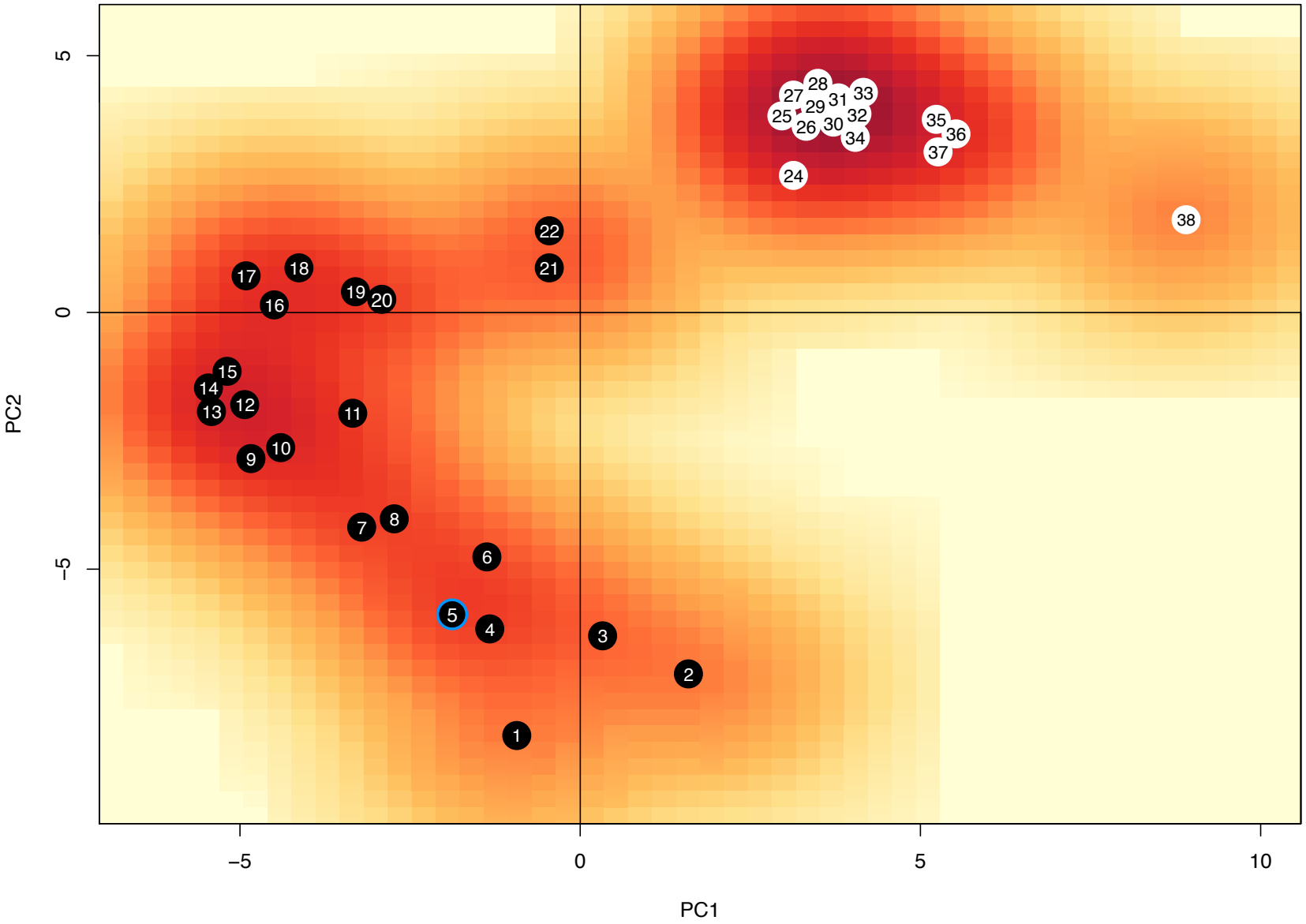
$$R_{10\times 10} = \begin{pmatrix} 1.00 & 0.71 & 0.77 & 0.52 & 0.79 & 0.55 & 0.69 & 0.63 & 0.60 & 0.18 \\ 0.71 & 0.99 & 0.52 & 0.69 & 0.79 & 0.85 & 0.49 & 0.44 & 0.49 & 0.22 \\ 0.77 & 0.52 & 1.00 & 0.62 & 0.63 & 0.39 & 0.70 & 0.73 & 0.68 & 0.20 \\ 0.52 & 0.69 & 0.62 & 1.00 & 0.68 & 0.78 & 0.49 & 0.56 & 0.65 & 0.06 \\ 0.79 & 0.79 & 0.63 & 0.68 & 1.00 & 0.82 & 0.66 & 0.62 & 0.66 & 0.15 \\ 0.55 & 0.85 & 0.39 & 0.78 & 0.82 & 1.00 & 0.40 & 0.40 & 0.50 & 0.13 \\ 0.69 & 0.49 & 0.70 & 0.49 & 0.66 & 0.40 & 1.00 & 0.89 & 0.85 & 0.53 \\ 0.63 & 0.44 & 0.73 & 0.56 & 0.62 & 0.40 & 0.89 & 1.00 & 0.97 & 0.45 \\ 0.60 & 0.49 & 0.68 & 0.65 & 0.66 & 0.50 & 0.85 & 0.97 & 1.00 & 0.42 \\ 0.18 & 0.22 & 0.20 & 0.06 & 0.15 & 0.13 & 0.53 & 0.45 & 0.42 & 1.00 \end{pmatrix} \begin{matrix} \text{19: Citation PageRank} \\ \text{5: Journal Impact Factor} \\ \text{22: Citation Betweenness} \\ \text{6: Citation Closeness} \\ \text{11: Citation H-index} \\ \text{1: Citation Scimago Journal Rank} \\ \text{31: Usage PageRank} \\ \text{34: Usage Betweenness} \\ \text{24: Usage Closeness} \\ \text{39: Usage Impact Factor} \end{matrix}$$

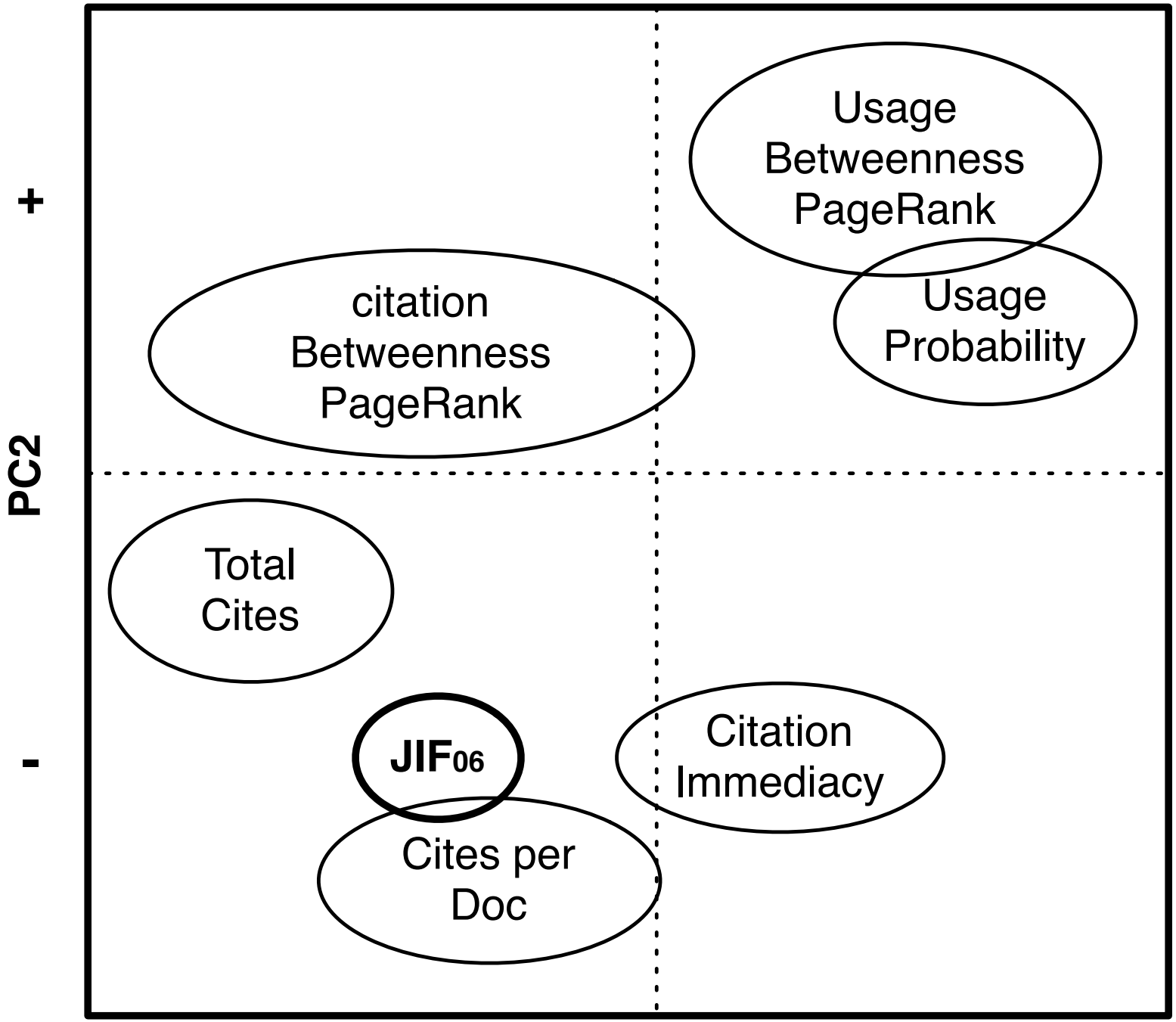


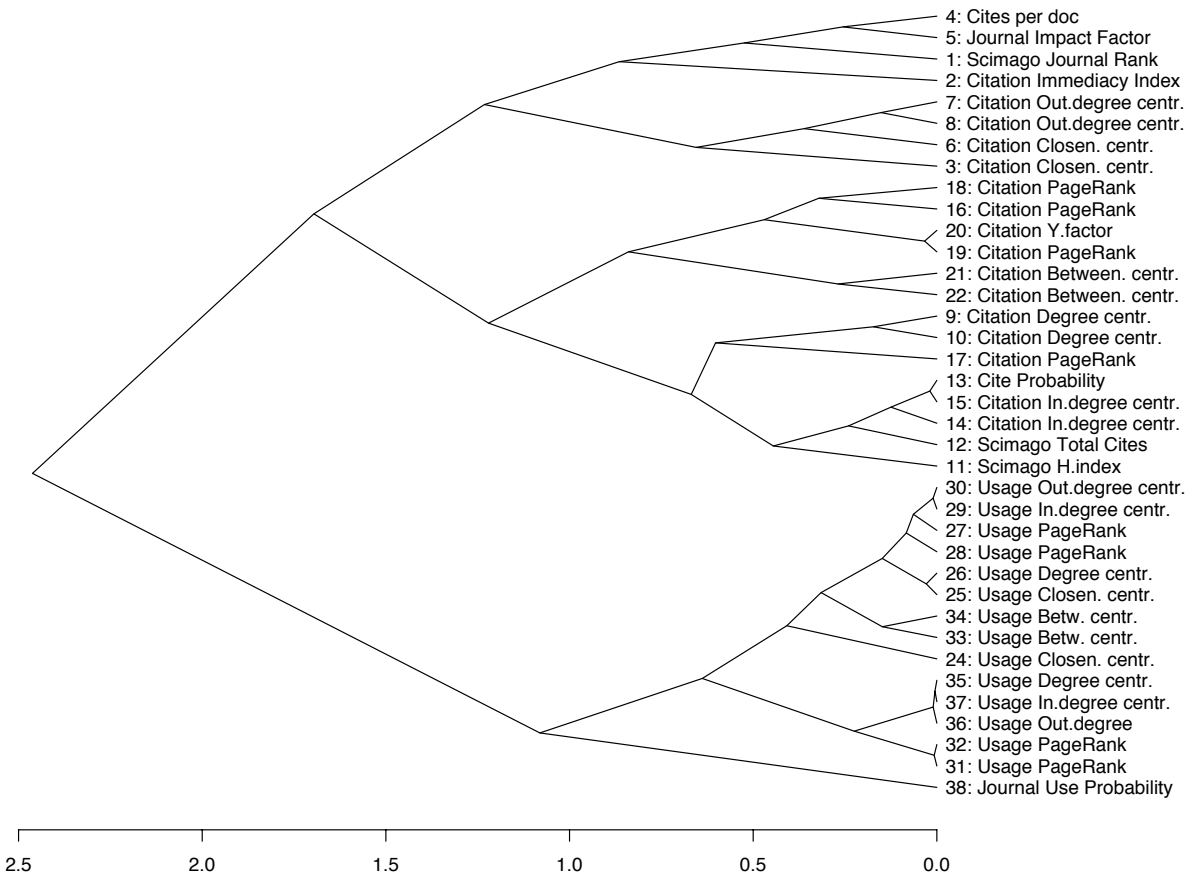
Schematic representation of data sources and processing. Impact measure identifiers.

Principal Component Analysis

	PC1	PC2	PC3	PC4	PC5
Proportion of Variance	66.1%	17.3%	9.2%	4.8%	0.9%
Cumulative Proportion	66.1%	83.4%	92.6%	97.4%	98.3%


PC2
PC1
5
0
−5
−5
0
5
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38


+
PC2
-
Usage
Betweenness
PageRank
Usage
Probability
citation
Betweenness
PageRank
Total
Cites
JIF06
Citation
Immediacy
Cites per
Doc



Cluster	Measures	Interpretation
1	38	Journal Use Probability
2	24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37	Usage measures
3	1, 2, 3, 4, 5	JIF, SJR, Cites per Document measures
4	6, 7, 8, 9, 10, 11, 12, 13, 14, 15	Total Citation rates and distributions
5	16, 17, 18, 19, 20, 21, 22	Citation Betweenness and PageRank

MESUR's end-user analytics services

Aimed at showcasing our results

Demos and prototypes:

`http://mesur.informatics.indiana.edu/demos/`

Outline

1 Introduction
- Problem statement
- Usage data

2 MESUR
- MESUR overview
- Creating MESUR's reference data set

3 Mapping Science

4 Metrics Survey

5 Services

6 Discussion
- Overview
- Future Research
- Relevant papers

MESUR: so far

Usage data: single largest reference data set of usage, citation and bibliographic data

- +1,000,000,000 usage events loaded
- multiple publishers, aggregators and institutions
- Infrastructure for research program

Usage graphs: track scientific flow of activity

- real-time studies of science
- inclusive of larger "scholarly community"

Metrics: valid, vetted indicators of scientific impact

- Different facets of scholarly impact
- Simple metrics, good results. Law of diminishing returns?
- Hybrid, consensus metrics?

MESUR: the bad

Sustainability: burden of data collection

- ad hoc, customized agreements with data providers
- restrictive agreements with regards to data sharing
- high costs of maintaining infrastructure: funding

Research program: too much to do

- 3rd party access: better science
- many eyes on data

Metrics and services: Community support and acceptance

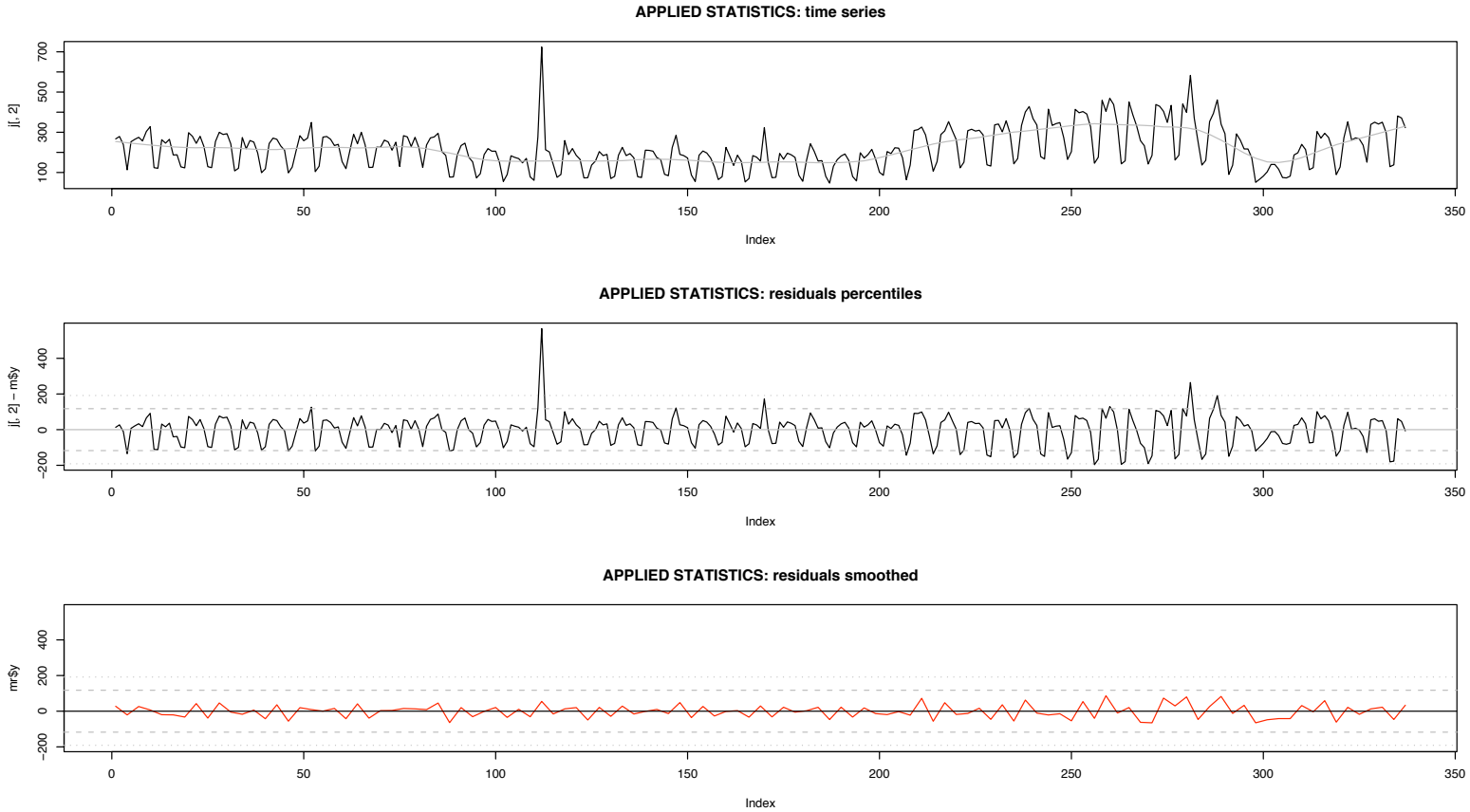
- Can't be limited to academic exercise
- Investigations need to be useful
- Accepted by community, become part of scholarly assessment system

Future Research: Two directions

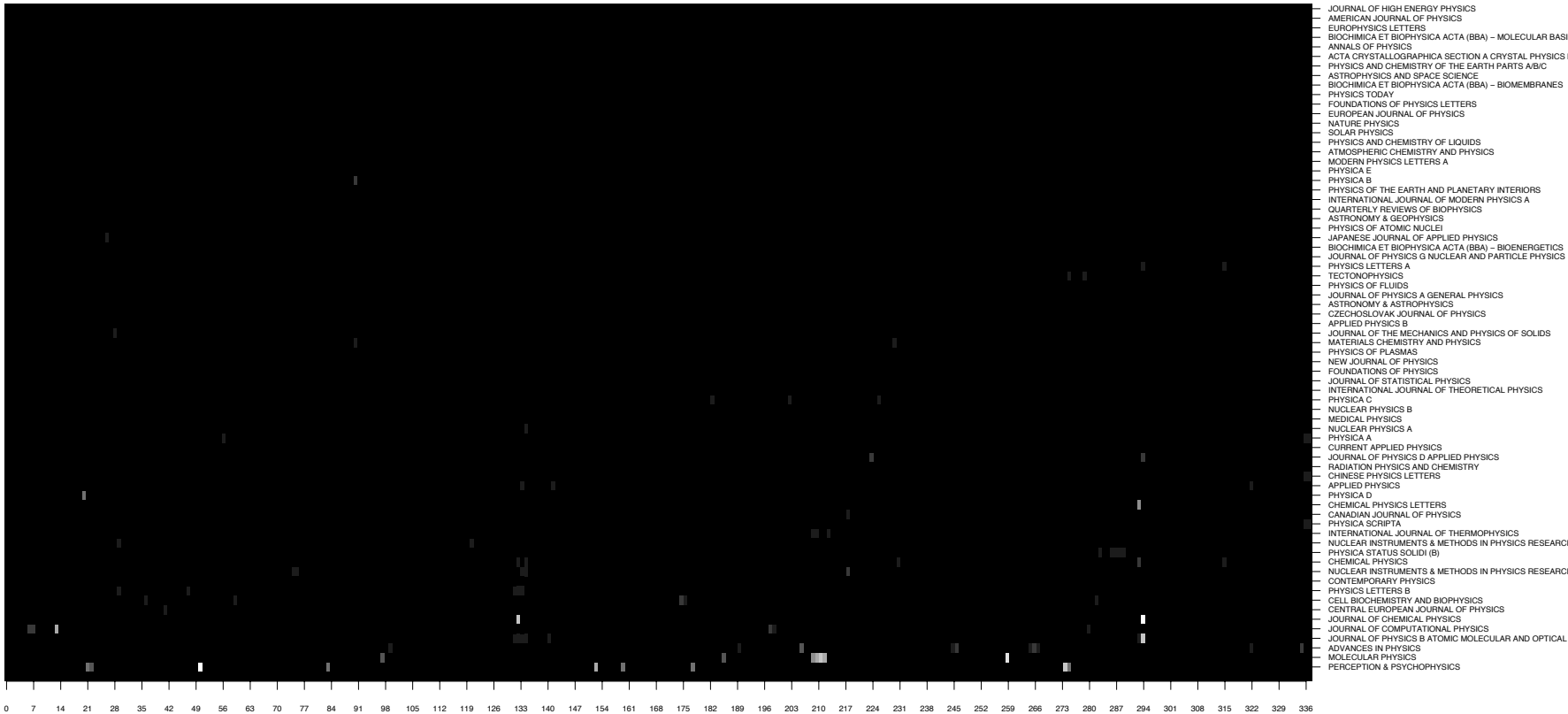
Longitudinal research and dynamic models of science

- Bursty behavior of science
  - Timeseries of reads over time are bursty
  - Coordinated: social networks at work?
- Modeling
  - Stochastic and agent-based models of scientific activity
  - Citation following? Group-think? Find parameters of human information searching behavior.

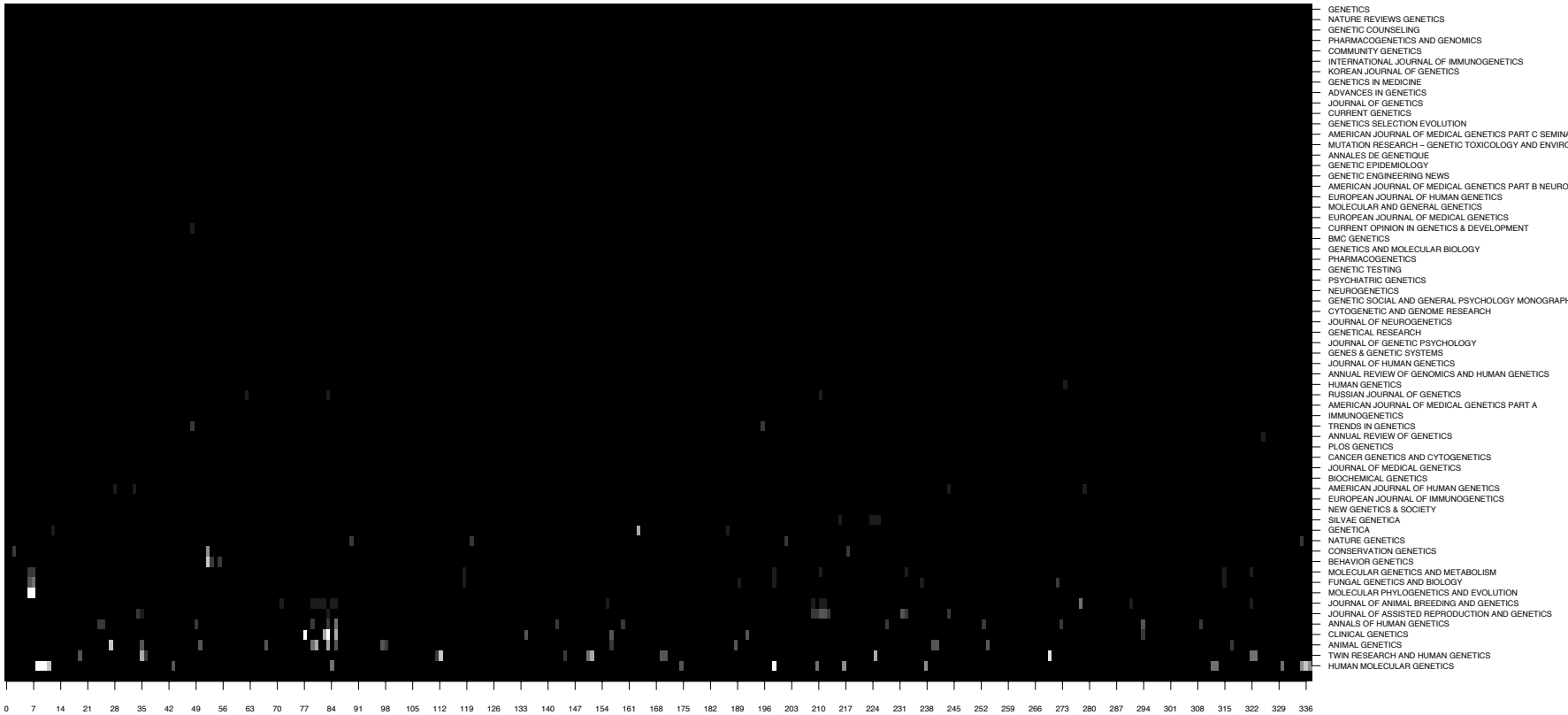
Bursty behavior



Contagion?


JOURNAL OF HIGH ENERGY PHYSICS
AMERICAN JOURNAL OF PHYSICS
EUROPHYSICS LETTERS
BIOCHIMICA ET BIOPHYSICA ACTA (BBA) – MOLECULAR BASI
ANNALS OF PHYSICS
ACTA CRYSTALLOGRAPHICA SECTION A CRYSTAL PHYSICS
PHYSICS AND CHEMISTRY OF THE EARTH PARTS A/B/C
ASTROPHYSICS AND SPACE SCIENCE
BIOCHIMICA ET BIOPHYSICA ACTA (BBA) – BIOMEMBRANES
PHYSICS TODAY
FOUNDATIONS OF PHYSICS LETTERS
EUROPEAN JOURNAL OF PHYSICS
NATURE PHYSICS
SOLAR PHYSICS
PHYSICS AND CHEMISTRY OF LIQUIDS
ATMOSPHERIC CHEMISTRY AND PHYSICS
MODERN PHYSICS LETTERS A
PHYSICA E
PHYSICA B
PHYSICS OF THE EARTH AND PLANETARY INTERIORS
INTERNATIONAL JOURNAL OF MODERN PHYSICS A
QUARTERLY REVIEWS OF BIOPHYSICS
ASTRONOMY & GEOPHYSICS
PHYSICS OF ATOMIC NUCLEI
JAPANESE JOURNAL OF APPLIED PHYSICS
BIOCHIMICA ET BIOPHYSICA ACTA (BBA) – BIOENERGETICS
JOURNAL OF PHYSICS G NUCLEAR AND PARTICLE PHYSICS
PHYSICS LETTERS A
TECTONOPHYSICS
PHYSICS OF FLUIDS
JOURNAL OF PHYSICS A GENERAL PHYSICS
ASTRONOMY & ASTROPHYSICS
CZECHOSLOVAK JOURNAL OF PHYSICS
APPLIED PHYSICS B
JOURNAL OF THE MECHANICS AND PHYSICS OF SOLIDS
MATERIALS CHEMISTRY AND PHYSICS
PHYSICS OF PLASMAS
NEW JOURNAL OF PHYSICS
FOUNDATIONS OF PHYSICS
JOURNAL OF STATISTICAL PHYSICS
INTERNATIONAL JOURNAL OF THEORETICAL PHYSICS
PHYSICA C
NUCLEAR PHYSICS B
MEDICAL PHYSICS
NUCLEAR PHYSICS A
PHYSICA A
CURRENT APPLIED PHYSICS
JOURNAL OF PHYSICS D APPLIED PHYSICS
RADIATION PHYSICS AND CHEMISTRY
CHINESE PHYSICS LETTERS
APPLIED PHYSICS
PHYSICA D
CHEMICAL PHYSICS LETTERS
CANADIAN JOURNAL OF PHYSICS
PHYSICA SCRIPTA
INTERNATIONAL JOURNAL OF THERMOPHYSICS
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH
PHYSICA STATUS SOLIDI (B)
CHEMICAL PHYSICS
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH
CONTEMPORARY PHYSICS
PHYSICS LETTERS B
CELL BIOCHEMISTRY AND BIOPHYSICS
CENTRAL EUROPEAN JOURNAL OF PHYSICS
JOURNAL OF CHEMICAL PHYSICS
JOURNAL OF COMPUTATIONAL PHYSICS
JOURNAL OF PHYSICS B ATOMIC MOLECULAR AND OPTICAL
ADVANCES IN PHYSICS
MOLECULAR PHYSICS
PERCEPTION & PSYCHOPHYSICS
0 7 14 21 28 35 42 49 56 63 70 77 84 91 98 105 112 119 126 133 140 147 154 161 168 175 182 189 196 203 210 217 224 231 238 245 252 259 266 273 280 287 294 301 308 315 322 329 336

Contagion?


GENETICS
NATURE REVIEWS GENETICS
GENETIC COUNSELING
PHARMACOGENETICS AND GENOMICS
COMMUNITY GENETICS
INTERNATIONAL JOURNAL OF IMMUNOGENETICS
KOREAN JOURNAL OF GENETICS
GENETICS IN MEDICINE
ADVANCES IN GENETICS
JOURNAL OF GENETICS
CURRENT GENETICS
GENETICS SELECTION EVOLUTION
AMERICAN JOURNAL OF MEDICAL GENETICS PART C SEMINA
MUTATION RESEARCH – GENETIC TOXICOLOGY AND ENVIRO
ANNALES DE GENETIQUE
GENETIC EPIDEMIOLOGY
GENETIC ENGINEERING NEWS
AMERICAN JOURNAL OF MEDICAL GENETICS PART B NEURO
EUROPEAN JOURNAL OF HUMAN GENETICS
MOLECULAR AND GENERAL GENETICS
EUROPEAN JOURNAL OF MEDICAL GENETICS
CURRENT OPINION IN GENETICS & DEVELOPMENT
BMC GENETICS
GENETICS AND MOLECULAR BIOLOGY
PHARMACOGENETICS
GENETIC TESTING
PSYCHIATRIC GENETICS
NEUROGENETICS
GENETIC SOCIAL AND GENERAL PSYCHOLOGY MONOGRAPH
CYTOGENETIC AND GENOME RESEARCH
JOURNAL OF NEUROGENETICS
GENETICAL RESEARCH
JOURNAL OF GENETIC PSYCHOLOGY
GENES & GENETIC SYSTEMS
JOURNAL OF HUMAN GENETICS
ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS
HUMAN GENETICS
RUSSIAN JOURNAL OF GENETICS
AMERICAN JOURNAL OF MEDICAL GENETICS PART A
IMMUNOGENETICS
TRENDS IN GENETICS
ANNUAL REVIEW OF GENETICS
PLOS GENETICS
CANCER GENETICS AND CYTOGENETICS
JOURNAL OF MEDICAL GENETICS
BIOCHEMICAL GENETICS
AMERICAN JOURNAL OF HUMAN GENETICS
EUROPEAN JOURNAL OF IMMUNOGENETICS
NEW GENETICS & SOCIETY
SILVAE GENETICA
GENETICA
NATURE GENETICS
CONSERVATION GENETICS
BEHAVIOR GENETICS
MOLECULAR GENETICS AND METABOLISM
FUNGAL GENETICS AND BIOLOGY
MOLECULAR PHYLOGENETICS AND EVOLUTION
JOURNAL OF ANIMAL BREEDING AND GENETICS
JOURNAL OF ASSISTED REPRODUCTION AND GENETICS
ANNALS OF HUMAN GENETICS
CLINICAL GENETICS
ANIMAL GENETICS
TWIN RESEARCH AND HUMAN GENETICS
HUMAN MOLECULAR GENETICS
0 7 14 21 28 35 42 49 56 63 70 77 84 91 98 105 112 119 126 133 140 147 154 161 168 175 182 189 196 203 210 217 224 231 238 245 252 259 266 273 280 287 294 301 308 315 322 329 336

Relevant papers

Johan Bollen, Herbert Van de Sompel, Aric Hagberg and Ryan Chute

A Principal Component Analysis of 39 Scientific Impact Measures.

PLoS ONE, June 2009. URL: http://dx.plos.org/10.1371/journal.pone.0006022.

Michael Kurtz and Johan Bollen.

Usage bibliometrics.

Annual Review of Information Science and Technology, 2010

Johan Bollen, Herbert Van de Sompel, Aric Hagberg,Luis Bettencourt, Ryan Chute, Marko A. Rodriguez, Lyudmila Balakireva.

Clickstream data yields high-resolution maps of science.

PLoS One, February 2009.