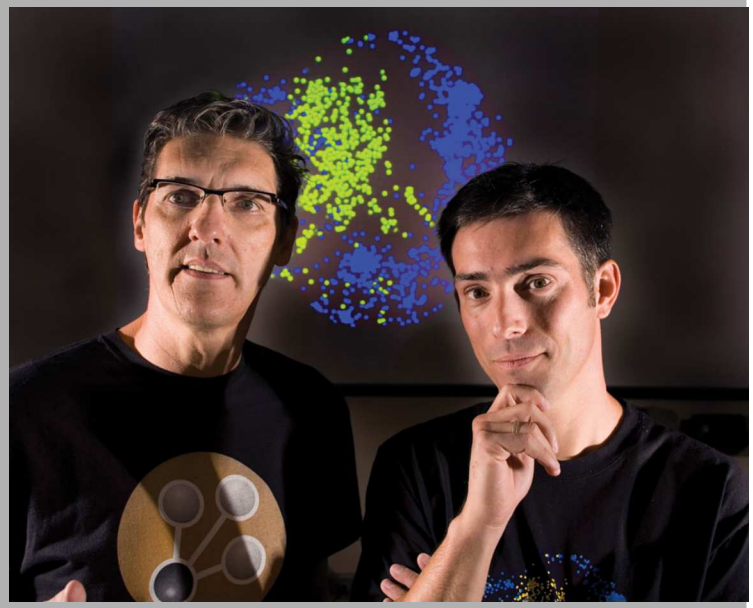
http://www.slideshare.net/hvdsomp/mesur-making-sense-and-use-of-usage-data

MESUR
Making Use and Sense of Scholarly Usage Data

<http://www.mesur.org>

Johan Bollen - jbollen@lanl.gov
Herbert Van de Sompel - herbertv@lanl.gov

Digital Library Research & Prototyping Team
Research Library
Los Alamos National Laboratory, USA



The MESUR research was funded by the Andrew W. Mellon Foundation

Acknowledgements: Marko A. Rodriguez (LANL), Ryan Chute (LANL), Lyudmila L. Balakireva (LANL), Aric Hagberg (LANL), Luis Bettencourt (LANL)

MESUR is Paradigm Shift Material

MESUR looks into new approaches to assess scholarly impact

- The Thomson Scientific IF was about the only metrics that could be computed in a paper-based era.
- But we don't live in the paper-based era anymore. So MESUR researches metrics for the digital era:
  - Usage-based metrics:
    - Access to scholarly materials happens via networked systems, not via paper stored in libraries.
    - Networked systems can record a great deal about access to materials; much more than libraries could about access to paper.
  - Network-based metrics:
    - Scholarly communication generates networks, e.g. citation networks, co-authorship networks, usage networks, …
    - A wide variety of metrics can be computed for such networks; much more than simple citation counts.

The Promise of Usage Data

Metrics based on usage data have significant potential

- Interactions be recorded for all digital scholarly content, i.e. papers, journals, preprints, blog postings, datasets, chemical structures, software, …
  - Not just for ~ 10,000 journals
- Interactions reflects the activities of all users of scholarly information, not only of scholarly authors
- Interactions are recorded starting immediately after *publication*
  - Not once read and cited (think publication delays)
  - Rapid indicator of scholarly trends

- So the interest in usage data from projects such as COUNTER, Citebase, IKS and MESUR should not come as a surprise!

And the Obvious Challenges of Usage Data

Usage data comes with significant challenges

- What exactly is usage?
  - E.g. various types of usage (download pdf, email abstract, …); impact of user interface on usage recordings, …
  - *Attention data* would be a better term.
- Privacy concerns
- Aggregating item-level usage data across networked systems:
  - Standardized recording
  - Standardized aggregating
  - Click-streams across networked systems
- How to deal with bots?

Network-Based Metrics

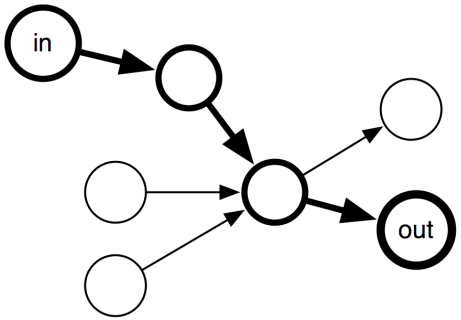
We have 50 years of network science available to us

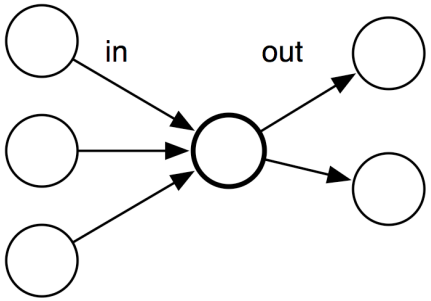
- A wide variety of metrics has been proposed to characterize networks, and to assess the importance of nodes in a network
  - E.g. social network analysis, small world graphs, graph theory, social modeling
- So when defining metrics for scholarly communication (clearly a network), we should probably leverage network science
  - Cf. Google's PageRank versus Alta Vista's statistical ranking
- A network (and hence a network-based metric) takes context into account; a statistical count does not.
- Readings:
  - Barabasi (2003) Linked.
  - Wasserman (1994). Social network analysis.

Network-Based Metrics

For an easy entry point, see http://en.wikipedia.org/wiki/Centrality

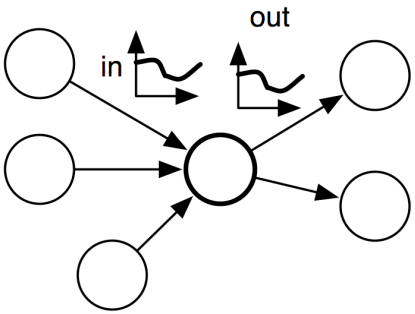
Shortest path
- Closeness
- Betweenness
- Newman

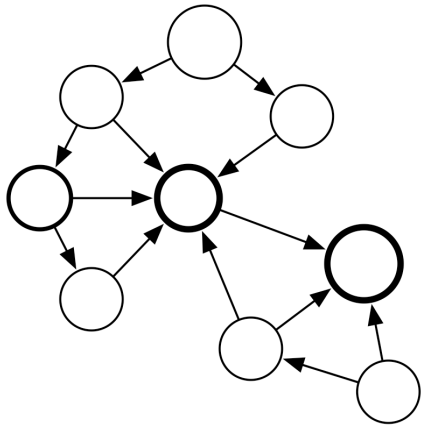




Degree
- In-degree
- Out-degree

Distribution
- In-degree entropy
- Out-degree entropy
- Bucket Entropy





Random walk
- PageRank
- Eigenvector

PageRank computed on Citation Network

	ISI IF		PR_w x 10^3		Y-factor x 10^2	
rank	value	Journal	value	Journal	value	Journal
1	52.28	ANNU REV IMMUNOL	17.46	J BIOL CHEM	51.15	NATURE
2	37.65	ANNU REV BIOCHEM	16.51	NATURE	47.72	SCIENCE
3	36.83	PHYSIOL REV	16.02	SCIENCE	19.92	NEW ENGL J MED
4	35.04	NAT REV MOL CELL BIO	13.77	PNAS	14.36	CELL
5	34.83	NEW ENGL J MED	8.90	PHYS REV LETT	14.14	PNAS
6	33.95	NAT REV CANCER	5.93	PHYS REV B	11.32	J BIOL CHEM
7	33.06	CANCER J CLIN	5.72	NEW ENGL J MED	8.73	JAMA
8	30.98	NATURE	5.40	ASTROPHYS J	7.83	LANCET
9	30.55	NAT MED	5.39	CELL	7.22	NAT GENET
10	30.17	ANNU REV NEUROSCI	4.90	J AM CHEM SOC	6.26	PHYS REV LETT

2003 JCR, Science Edition
5709 journals

Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel. Journal status. Scientometrics, 69(3), December 2006 (DOI:10.1007/s11192-006-0176-z)

Philip Ball. Prestige is factored into journal ratings. Nature **439**, 770-771, February 2006 (DOI:10.1038/439770a)

MESUR: A Thorough, Scientific Approach

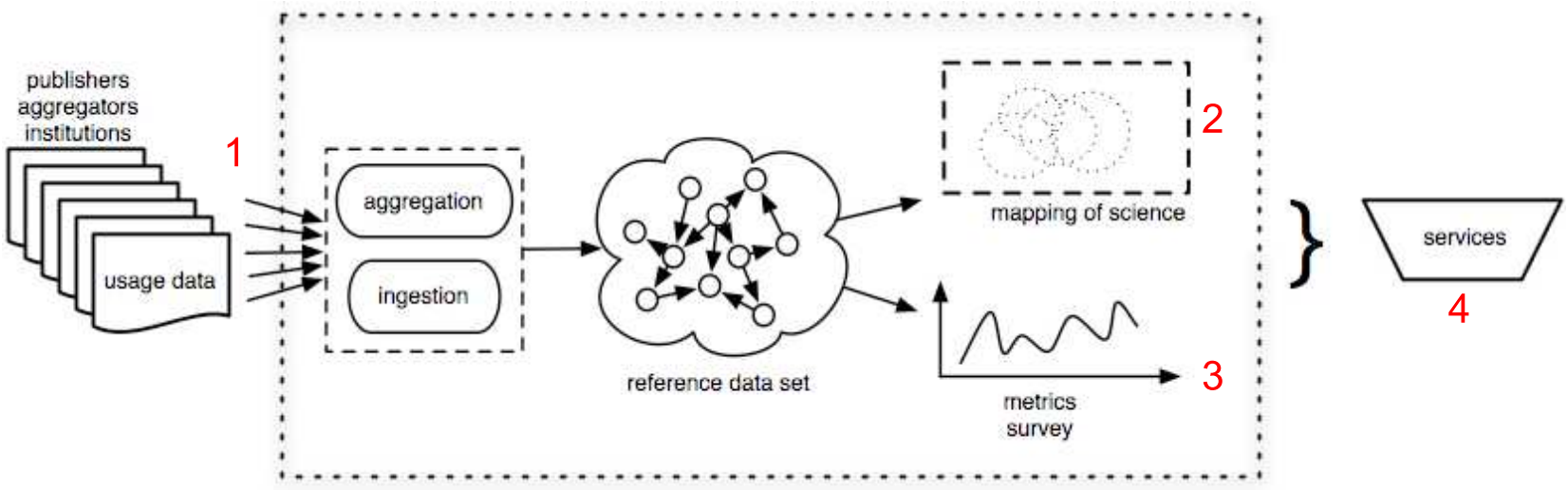
1. Create very large-scale reference data set
   a) Usage, citation and bibliographic data combined
   b) Various communities, various collections

1. Investigate validity of usage data and usage-based metrics – focus on journals:
   a) Is there any significant structure in usage data?
   b) Compute a variety of journal metrics for usage data & cross-validate with other journal metrics, e.g. citation-based IF

1. Deploy tools to explore usage-based journal metrics

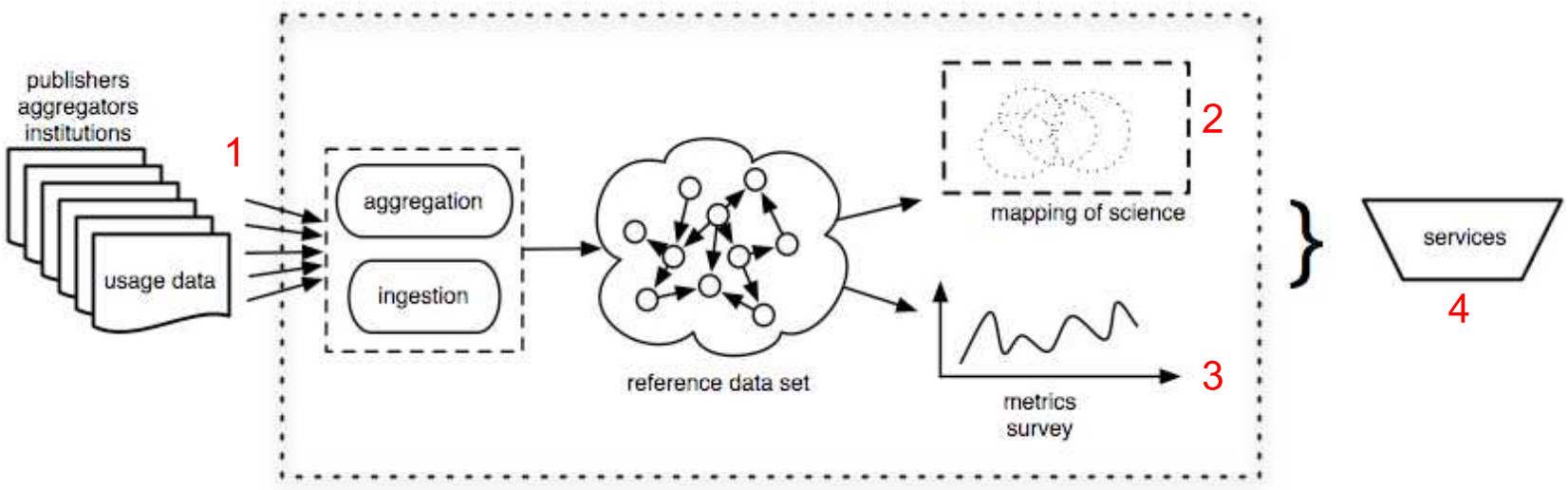
MESUR: Project Phases

1) Usage data acquisition
2) Structure in usage data - Map of Science
3) Metrics based on usage and citation - Compare
4) Services



MESUR: Project Phases

1) Usage data acquisition
1) Structure in usage data - Map of Science
2) Metrics based on usage and citation - Compare
3) Services



How to Obtain 1,000,000,000 Usage Events?

Politely ask publishers, aggregators, institutions

- Scale: > 1,000,000,000 usage events
- Period: 2002-2007, but mostly 2006
- Span:
  - > 50M articles ; > 100,000 journals (inc. newspapers, magazines,…)
- Publishers, Aggregators, Linking Servers, Proxy Servers:
  - BMC, Blackwell, UC, CSU (23), EBSCO, ELSEVIER, EMERALD, INGENTA, JSTOR, LANL, MIMAS/ZETOC, THOMSON, UPENN (9), UTEXAS
- Strict agreements regarding confidentiality of data

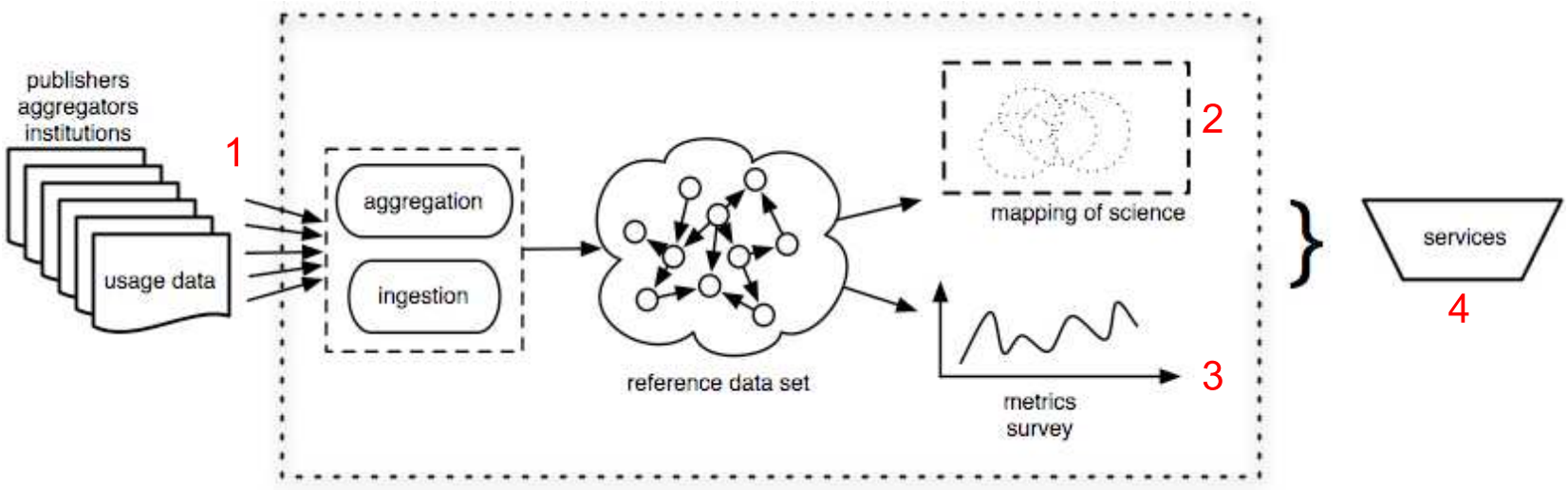
Some Minimal Requirements for Usage Data

In order to be able to construct usage-based networks

- Article level usage events
- Fields: unique session ID, date/time, unique document ID and/or metadata, request type

MESUR: Project Phases

1) Usage data acquisition
1) Structure in usage data - Map of Science
1) Metrics based on usage and citation - Compare
2) Services



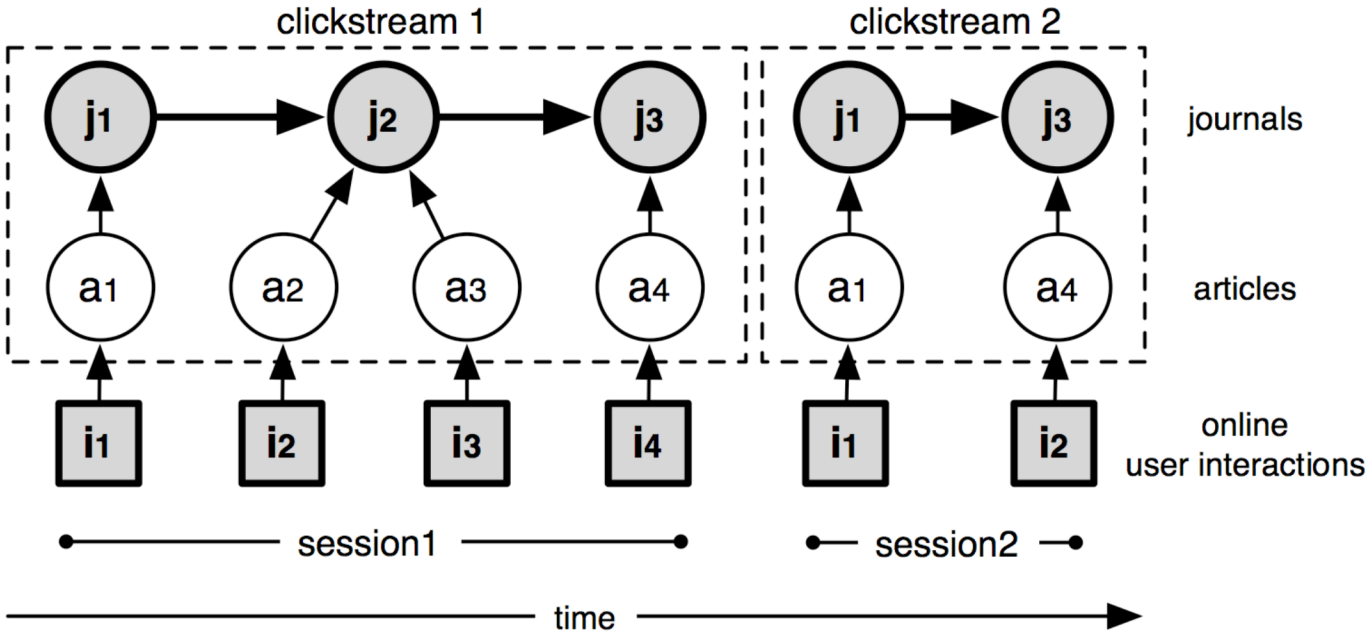
Data set: subset of MESUR

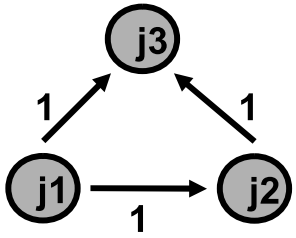
- Common time period:
  - March 1st 2006 - February 1st 2007
  - Thomson Scientific (Web of Science), Elsevier (Scopus), JSTOR, Ingenta, University of Texas (9 campuses, 6 health institutions), and California State University (23 campuses)

- 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%	

Generating a Network from Usage Data





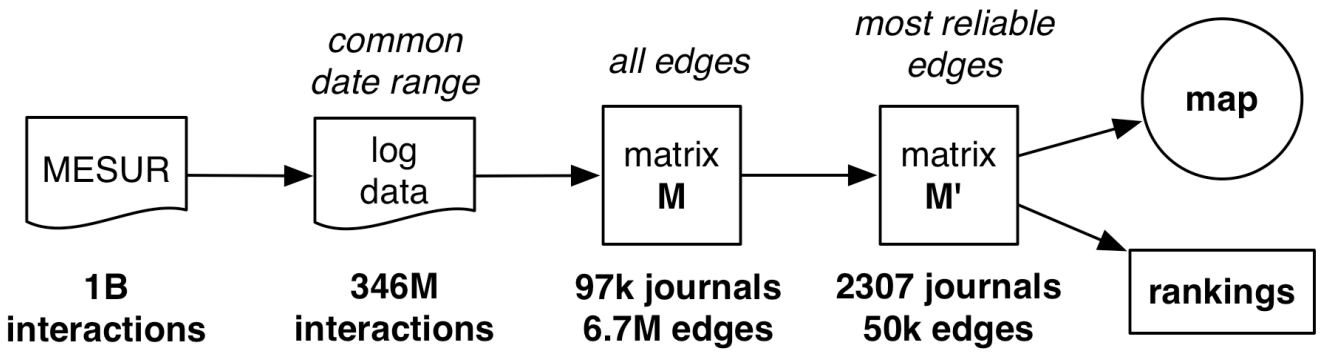
Same session ~ documents relatedness

- Same session, same user: common interest
- Frequency of co-occurrence = degree of relationship
- Normalized: conditional probability

Note: not something we invented

- Association rule learning in data mining
- Cf. Netflix, Amazon recommendations

Visualizing a Usage-Based Network

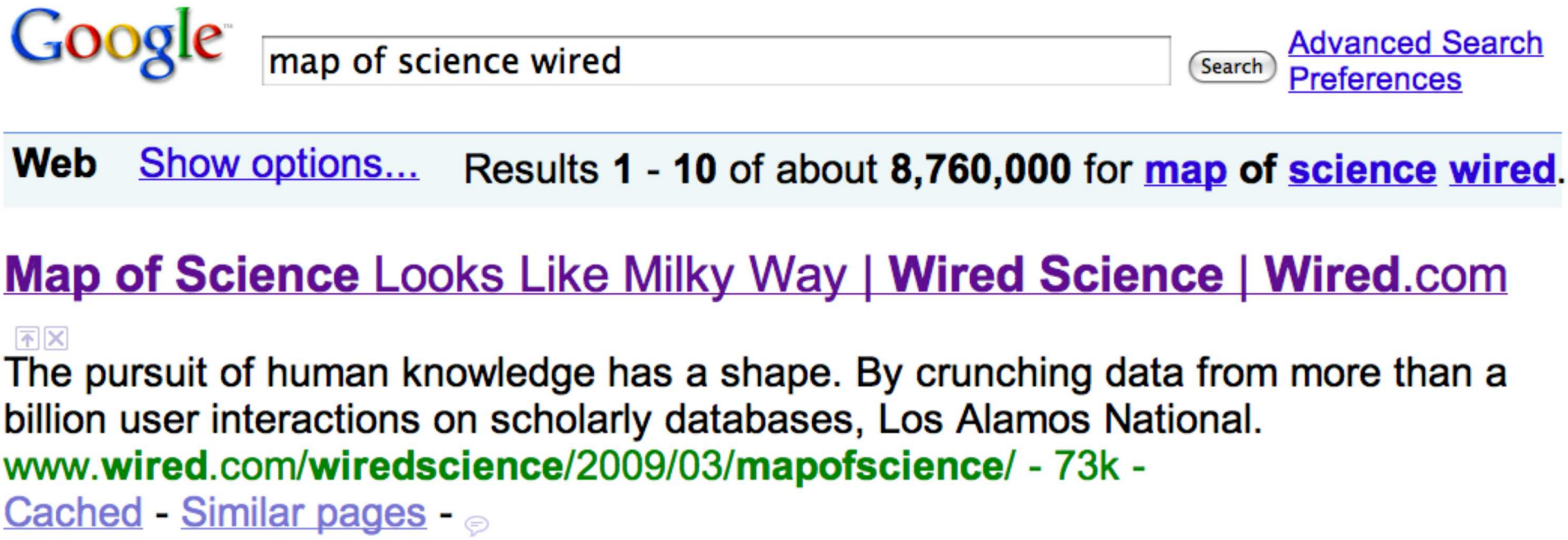


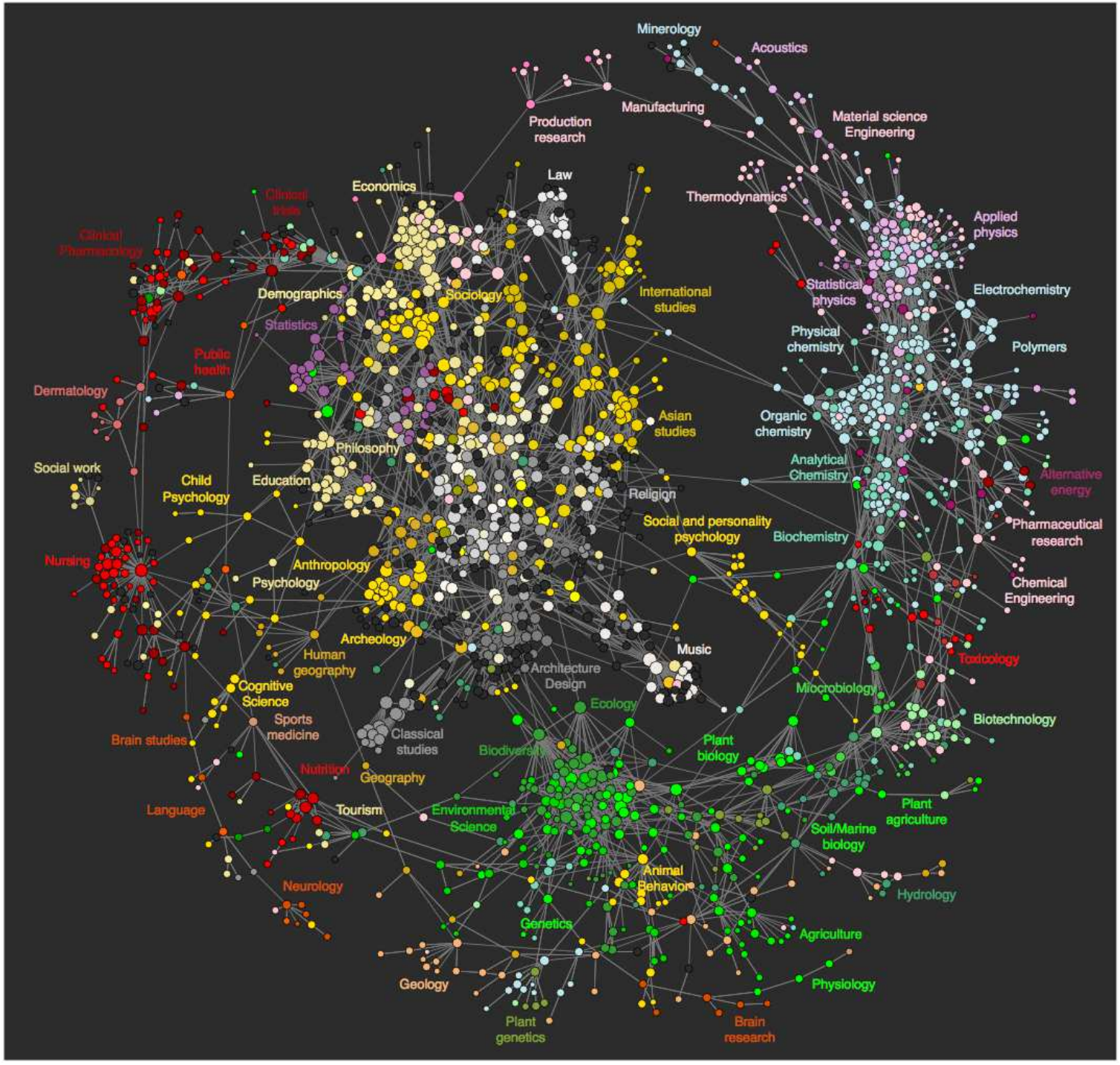
	Network matrix	
Parameter	M	M'
Journals	97,532	2,307
Edges	6,783,552	50,000
Matrix density	0.071%	0.939%
Strongly Connected Components (SCC)	16,474	236
Journals in SCC	80,934	1,944
Average journal clustering coefficient (SCC)	0.285	0.514
Diameter of largest SCC	37	14

Layout algorithm:
- "Fruchterman-Reingold" (1991)
- "Force-directed placement"
- Balancing node attraction (edges) with geometric repulsion (distance)

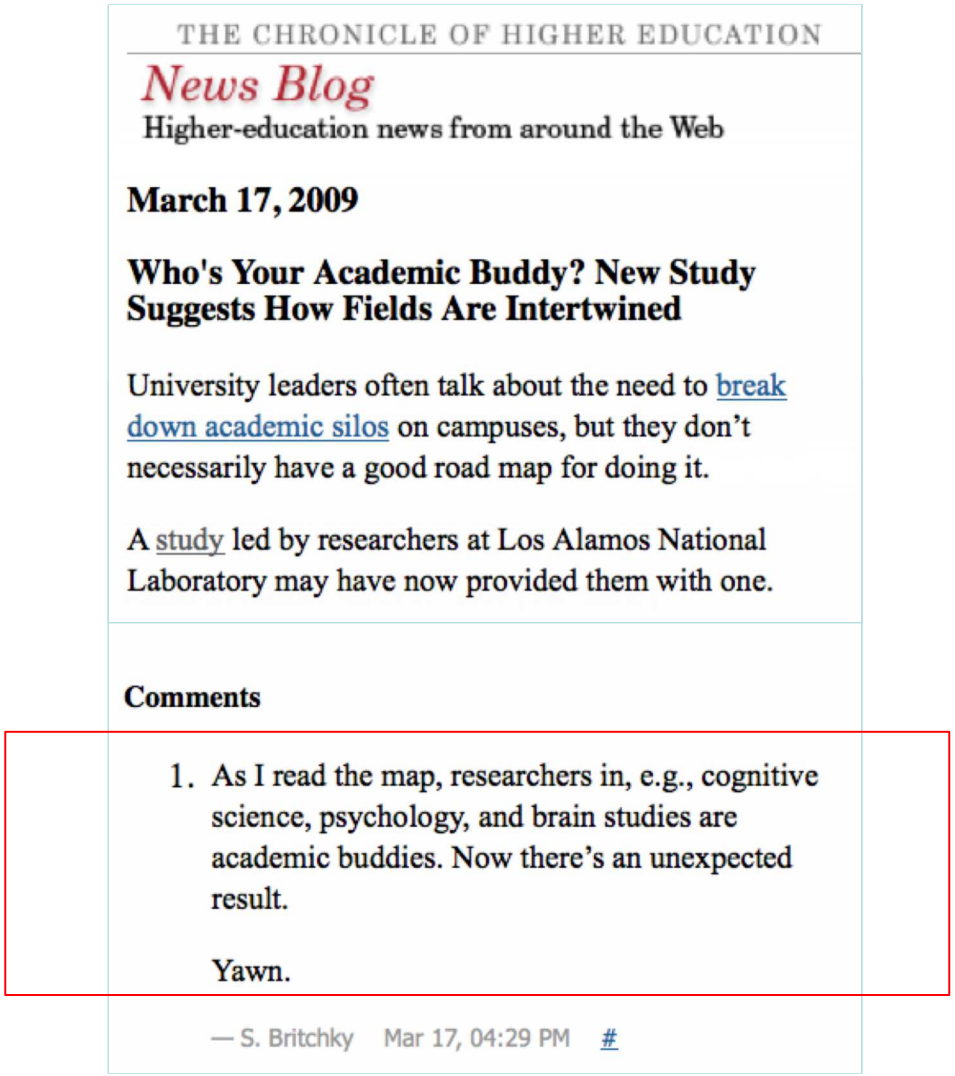
Bollen J, Van de Sompel H, Hagberg A, Bettencourt L, Chute R, et al. 2009 Clickstream Data Yields High-Resolution Maps of Science. PLoS ONE 4(3): e4803. DOI:10.1371/journal.pone.0004803

Google for: map of science wired




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

Yes, relevant structure in Usage-Based Map!



THE CHRONICLE OF HIGHER EDUCATION

News Blog

Higher-education news from around the Web

March 17, 2009

Who's Your Academic Buddy? New Study Suggests How Fields Are Intertwined

University leaders often talk about the need to break down academic silos on campuses, but they don't necessarily have a good road map for doing it.

A study led by researchers at Los Alamos National Laboratory may have now provided them with one.

Comments

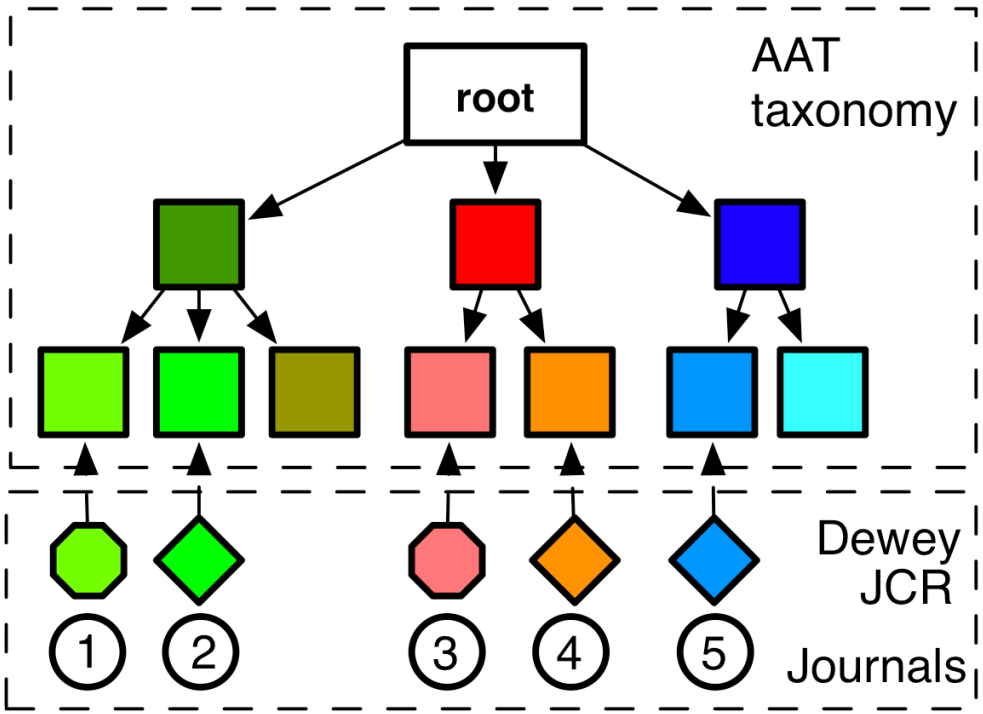
1. As I read the map, researchers in, e.g., cognitive science, psychology, and brain studies are academic buddies. Now there's an unexpected result.

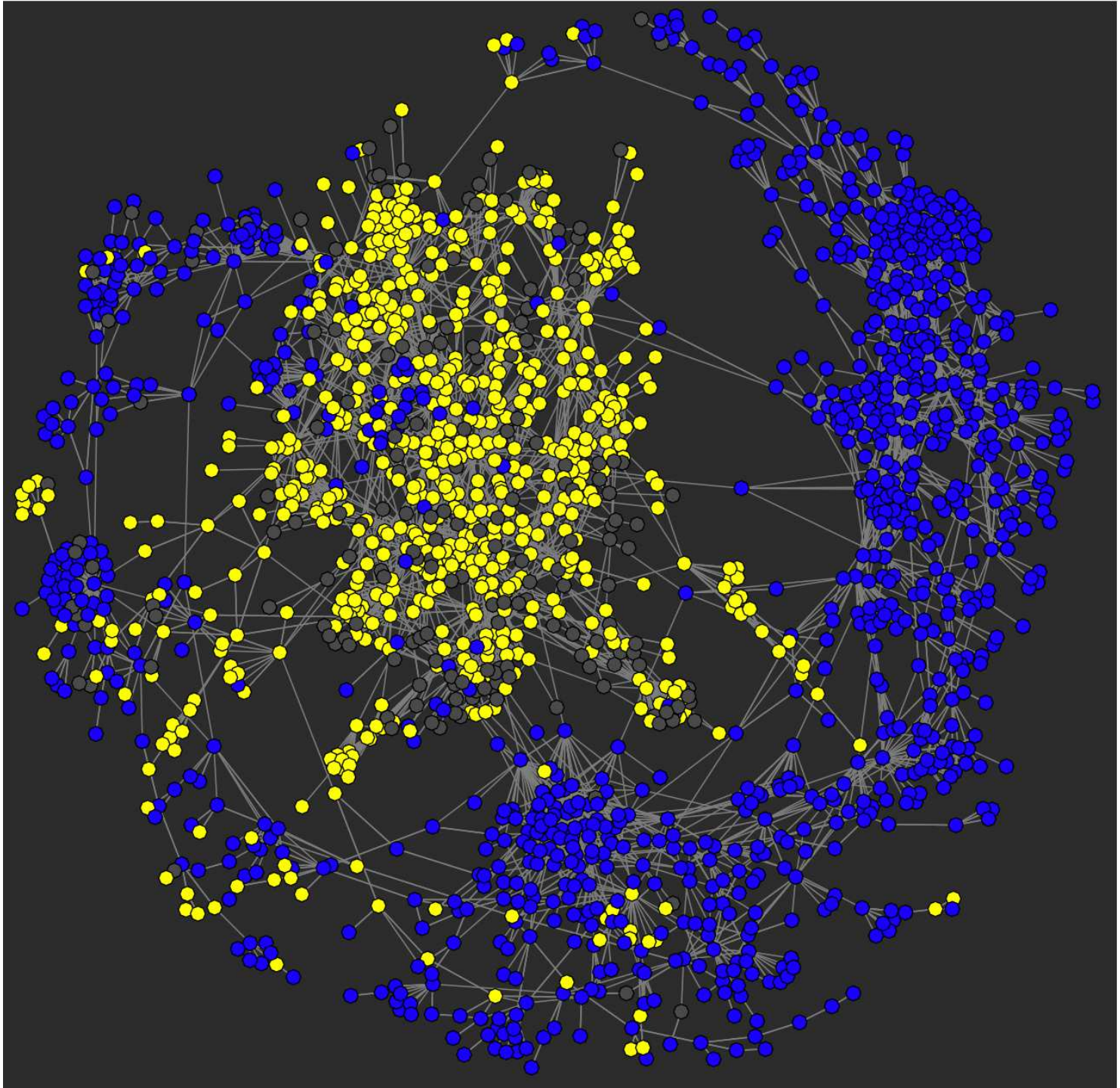
   Yawn.

   — S. Britchky Mar 17, 04:29 PM #

Validating the Usage-Based Map

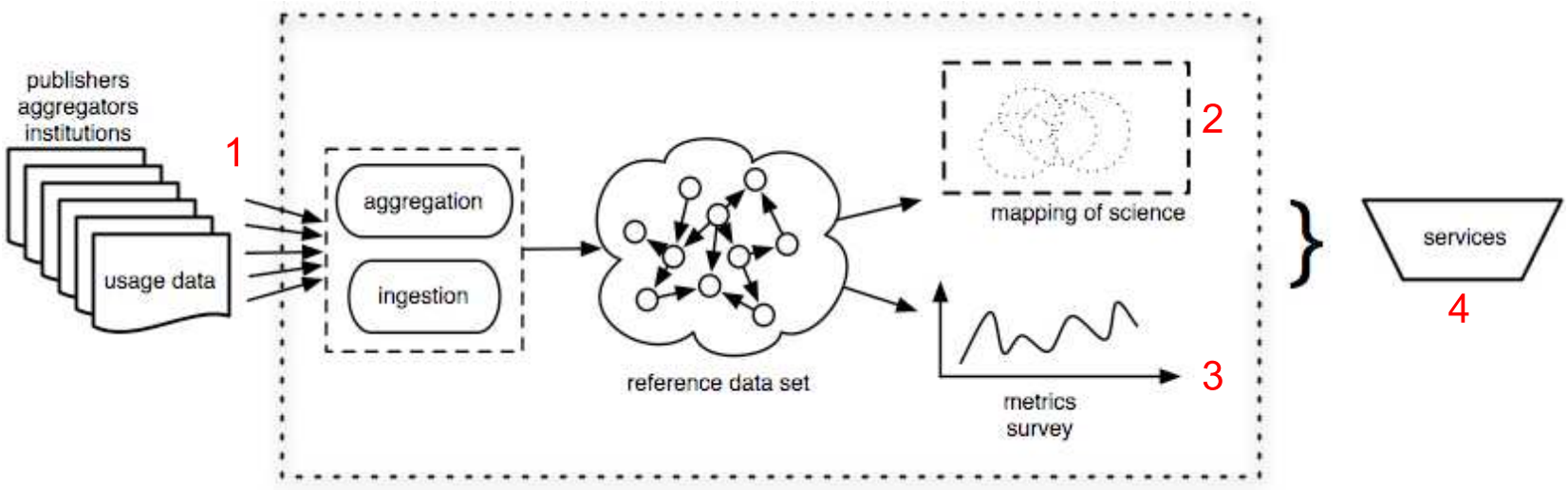
- Leverage Getty Research Art & Architecture thesaurus
- Cross-validation





MESUR: Project Phases

1) Usage data acquisition
2) Structure in usage data - Map of Science
1) Metrics based on usage and citation - Compare
1) Services



Metrics Computed for Usage and Citation Data

ID	Type	Measure	Source
1	Citation	Scimago Journal Rank	Scimago/Scopus
2	Citation	Immediacy Index	JCR 2007
3	Citation	Closeness	JCR 2007
4	Citaton	Cites per doc	Scimago/Scopus
5	Citation	Journal Impact Factor	JCR 2007
6	Citation	Closeness centrality	JCR 2007
7	Citation	Out-degree centrality	JCR 2007
8	Citation	Out-degree centrality	JCR 2007
9	Citation	Degree Centrality	JCR 2007
10	Citation	Degree Centrality	JCR 2007
11	Citation	H-Index	Scimago/Scopus
12	Citation	Scimago Total cites	Scimago/Scopus
13	Citation	Journal Cite Probability	JCR 2007
14	Citation	In-degree centrality	JCR 2007
15	Citation	In-degree centrality	JCR 2007
16	Citation	PageRank	JCR 2007
17	Citation	PageRank	JCR 2007
18	Citation	PageRank	JCR 2007
19	Citation	PageRank	JCR 2007
20	Citation	Y-factor	JCR 2007
21	Citation	Betweenness centrality	JCR 2007
22	Citation	Betweenness centrality	JCR 2007
23	*Citation*	*Citation Half-Life*	*JCR 2007*
24	Usage	Closeness centrality	MESUR 2007
25	Usage	Closeness centrality	MESUR 2007
26	Usage	Degree centrality	MESUR 2007
27	Usage	PageRank	MESUR 2007
28	Usage	PageRank	MESUR 2007
29	Usage	In-degree centrality	MESUR 2007
30	Usage	Out-degree centrality	MESUR 2007
31	Usage	PageRank	MESUR 2007
32	Usage	PageRank	MESUR 2007
33	Usage	Betweenness centrality	MESUR 2007
34	Usage	Betweenness centrality	MESUR 2007
35	Usage	Degree centrality	MESUR 2007
36	Usage	Out-degree centrality	MESUR 2007
37	Usage	In-degree centrality	MESUR 2007
38	Usage	Journal Use Probability	MESUR 2007
39	*Usage*	*Usage Impact Factor*	*MESUR 2007*

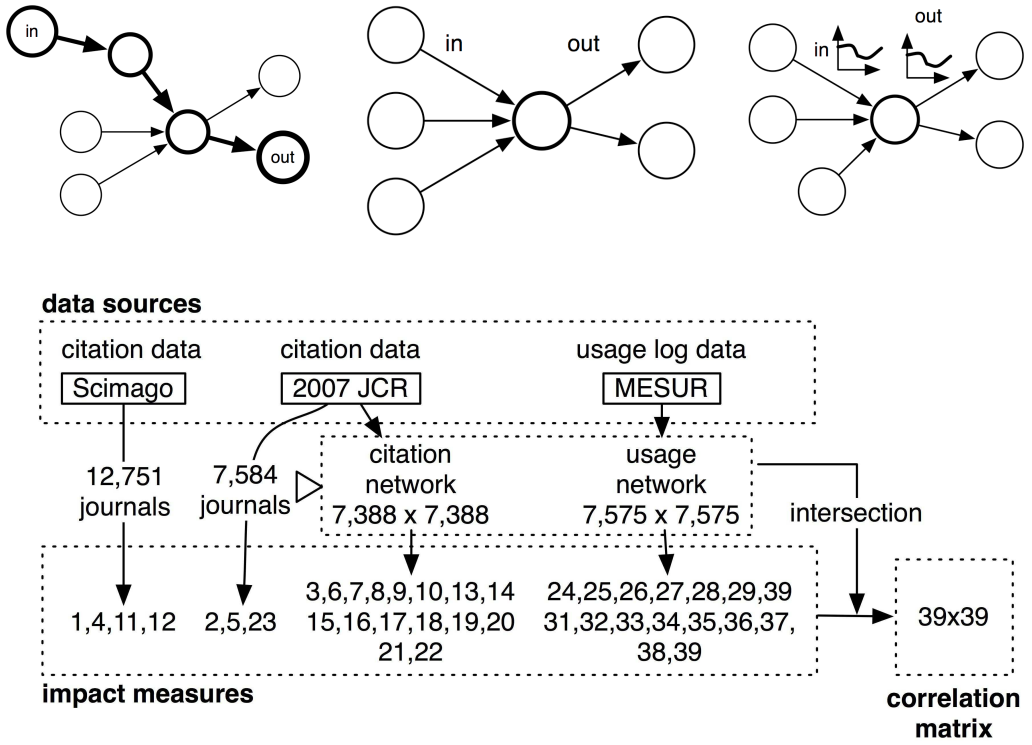


Fig. 4. Schematic representation of data sources and processing. Impact measure identifiers refer to Table 1.

Bollen J, Van de Sompel H, Hagberg A, Chute R. 2009 A principal component analysis of 39 scientific impact measures. http://arxiv.org/abs/0902.2183 Accepted by PLoS ONE

Citation Network Rankings

2004 Impact Factor

value	journal
1 49.794	CANCER
2 47.400	ANNU REV IMMUNOL
3 44.016	NEW ENGL J MED
4 33.456	ANNU REV BIOCHEM
5 31.694	NAT REV CANCER

Citation Pagerank

value	journal
1 0.0116	SCIENCE
2 0.0111	J BIOL CHEM
3 0.0108	NATURE
4 0.0101	PNAS
5 0.006	PHYS REV LETT

betweenness

value	journal
1 0.076	PNAS
2 0.072	SCIENCE
3 0.059	NATURE
4 0.039	LECT NOTES COMPUT SC
5 0.017	LANCET

Closeness

value	journal
1 7.02e-05	PNAS
2 6.72e-05	LECT NOTES COMPUT SC
3 6.43e-05	NATURE
4 6.37e-05	SCIENCE
5 6.37e-05	J BIOL CHEM

In-Degree

value	journal
1 3448	SCIENCE
2 3182	NATURE
3 2913	PNAS
4 2190	LANCET
5 2160	NEW ENGL J MED

In-degree entropy

Value	journal
1 9.849	LANCET
2 9.748	SCIENCE
3 9.701	NEW ENGL J MED
4 9.611	NATURE
5 9.526	JAMA

Usage Network Rankings

2004 Impact Factor

	value	journal
1	49.794	CANCER
2	47.400	ANNU REV IMMUNOL
3	44.016	NEW ENGL J MED
4	33.456	ANNU REV BIOCHEM
5	31.694	NAT REV CANCER

Pagerank

	value	journal
1	0.0016	SCIENCE
2	0.0015	NATURE
3	0.0013	PNAS
4	0.0010	LNCS
5	0.0008	J BIOL CHEM

betweenness

	value	journal
1	0.035	SCIENCE
2	0.032	NATURE
3	0.020	PNAS
4	0.017	LNCS
5	0.006	LANCET

Closeness

	value	journal
1	0.670	SCIENCE
2	0.665	NATURE
3	0.644	PNAS
4	0.591	LNCS
5	0.587	BIOCHEM BIOPH RES CO

In-Degree

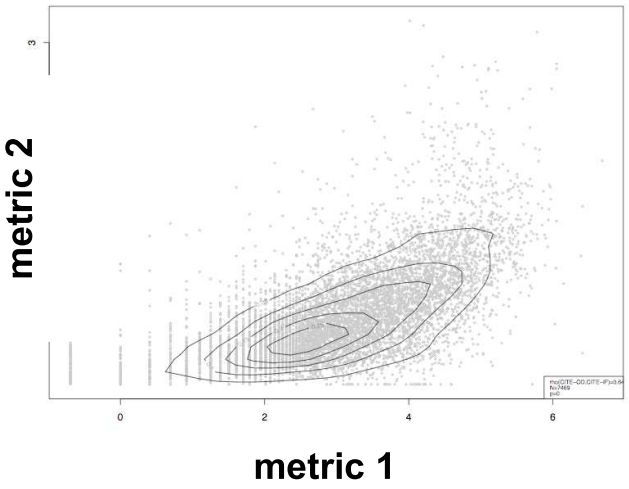
	value	journal
1	4195	SCIENCE
2	4019	NATURE
3	3562	PNAS
4	2438	J BIOL CHEM
5	2432	LNCS

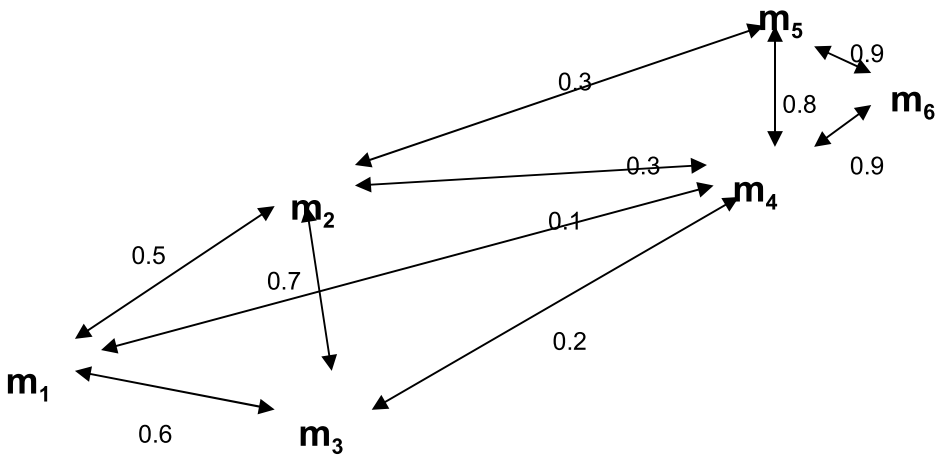
In-degree entropy

	Value	journal
1	9.364	MED HYPOTHESES
2	9.152	PNAS
3	9.027	LIFE SCI
4	8.939	LANCET
5	8.858	INT J BIOCHEM CELL B

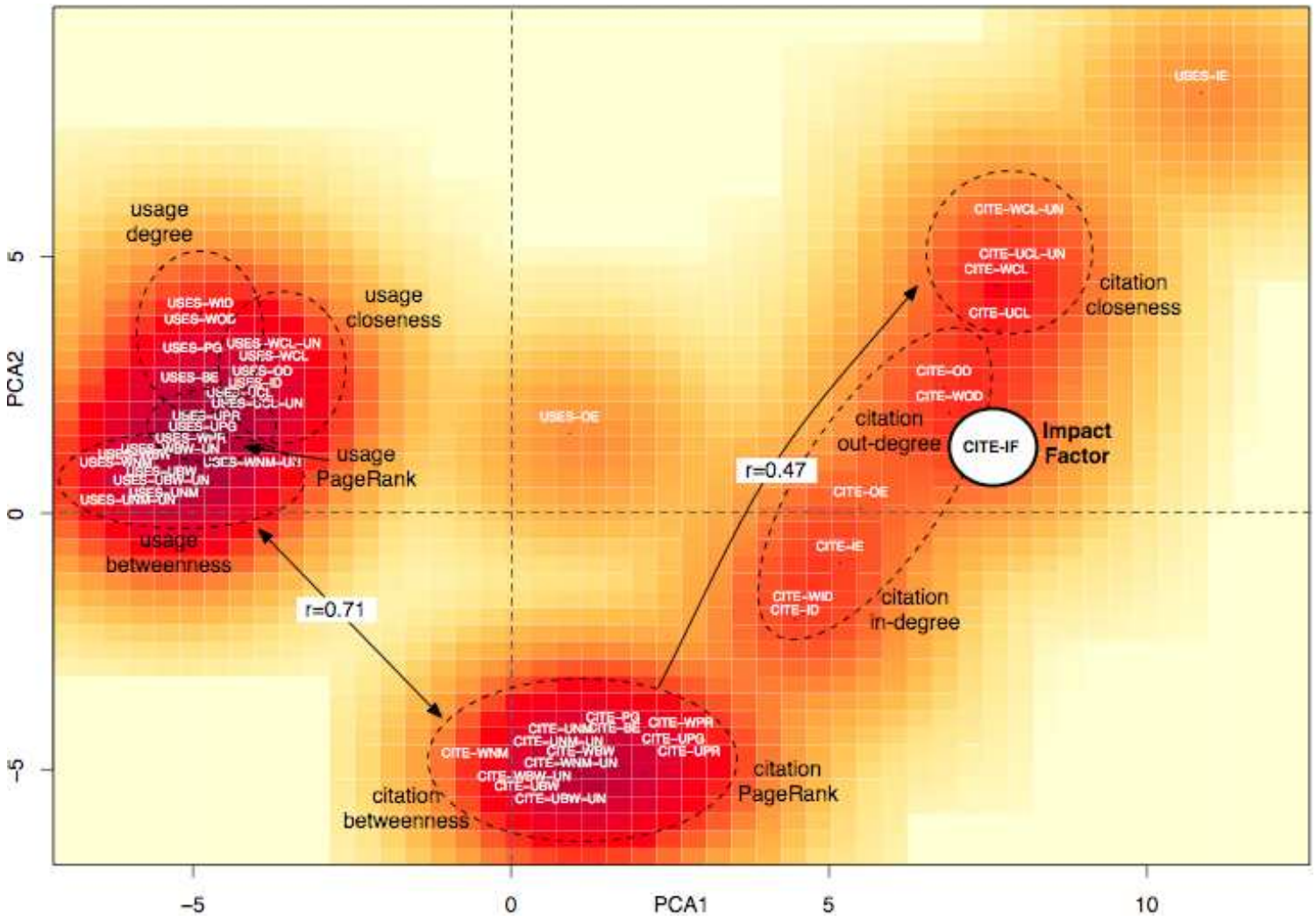
Metric Correlations: Metric Maps

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10
m1	1.00	0.75	0.67	0.61	0.46	0.57	0.99	0.79	0.79	0.40
m2	0.75	1.00	0.96	0.81	0.82	0.83	0.73	0.68	0.69	0.77
m3	0.67	0.96	1.00	0.77	0.77	0.81	0.65	0.62	0.63	0.72
m4	0.61	0.81	0.77	1.00	0.64	0.67	0.60	0.50	0.51	0.64
m5	0.46	0.82	0.77	0.64	1.00	0.92	0.44	0.57	0.58	0.89
m6	0.57	0.83	0.81	0.67	0.92	1.00	0.55	0.65	0.66	0.77
m7	0.99	0.73	0.65	0.60	0.44	0.55	1.00	0.78	0.79	0.39
m8	0.79	0.68	0.62	0.50	0.57	0.65	0.78	1.00	0.99	0.54
m9	0.79	0.69	0.63	0.51	0.58	0.66	0.79	0.99	1.00	0.55
m10	0.40	0.77	0.72	0.64	0.89	0.77	0.39	0.54	0.55	1.00



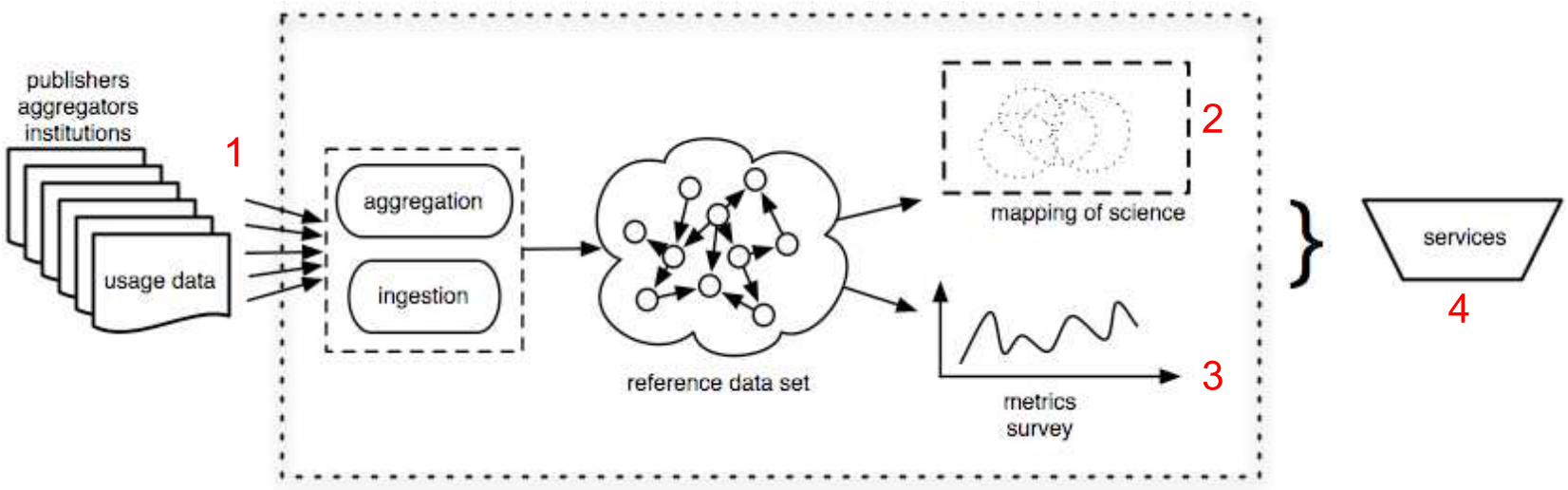


The MESUR Metrics Map

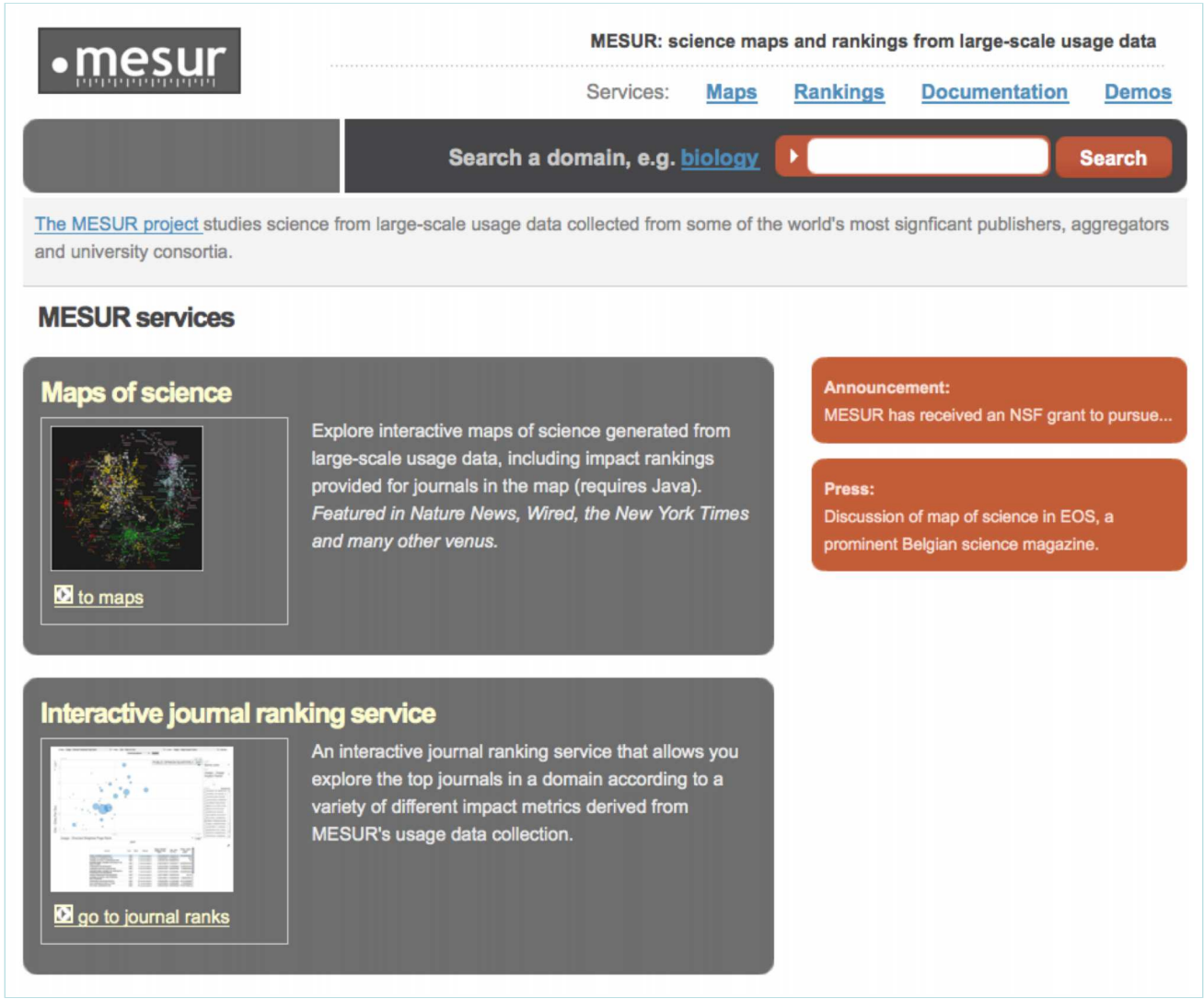


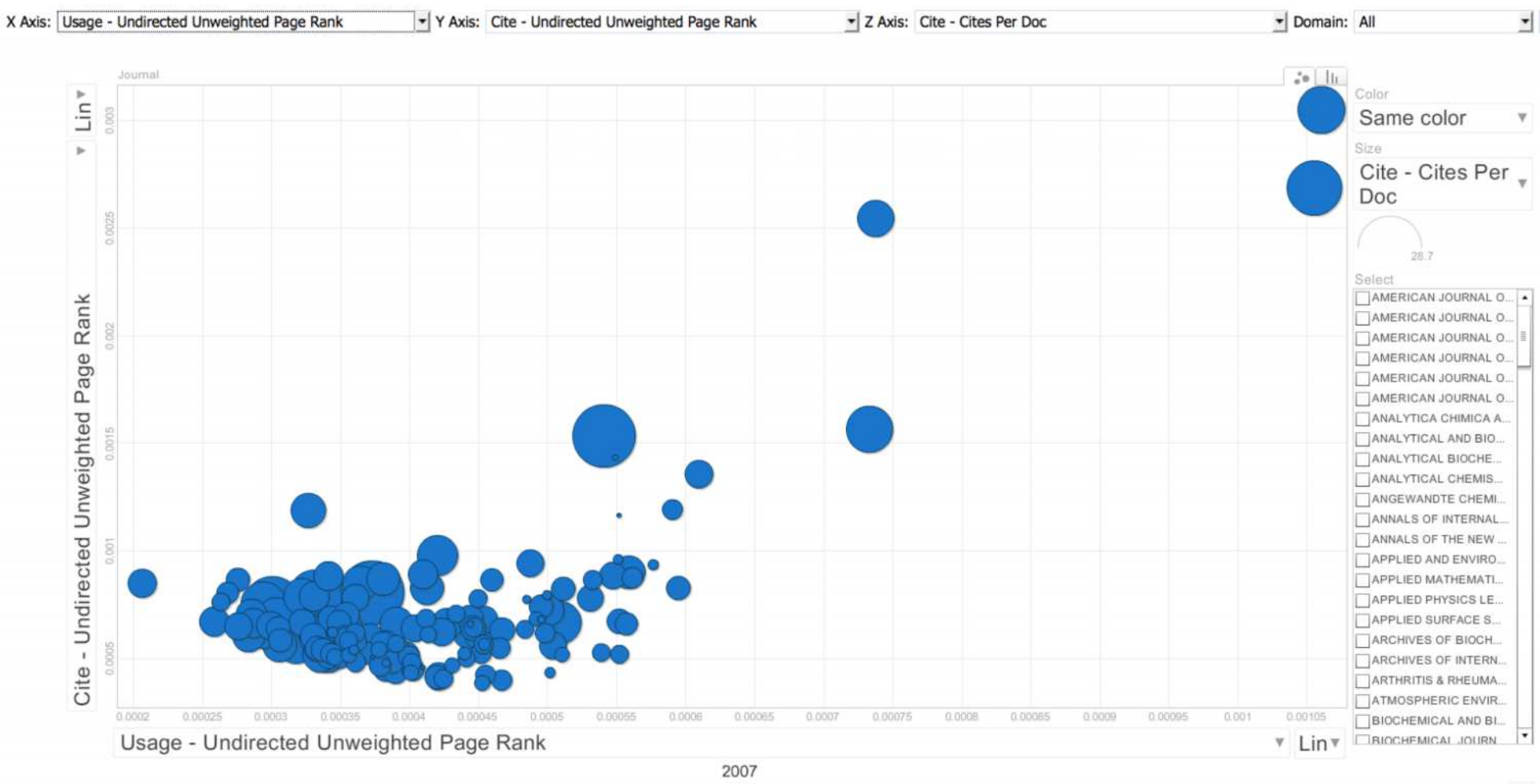
MESUR: Project Phases

1) Usage data acquisition
2) Structure in usage data - Map of Science
3) Metrics based on usage and citation - Compare
1) Services

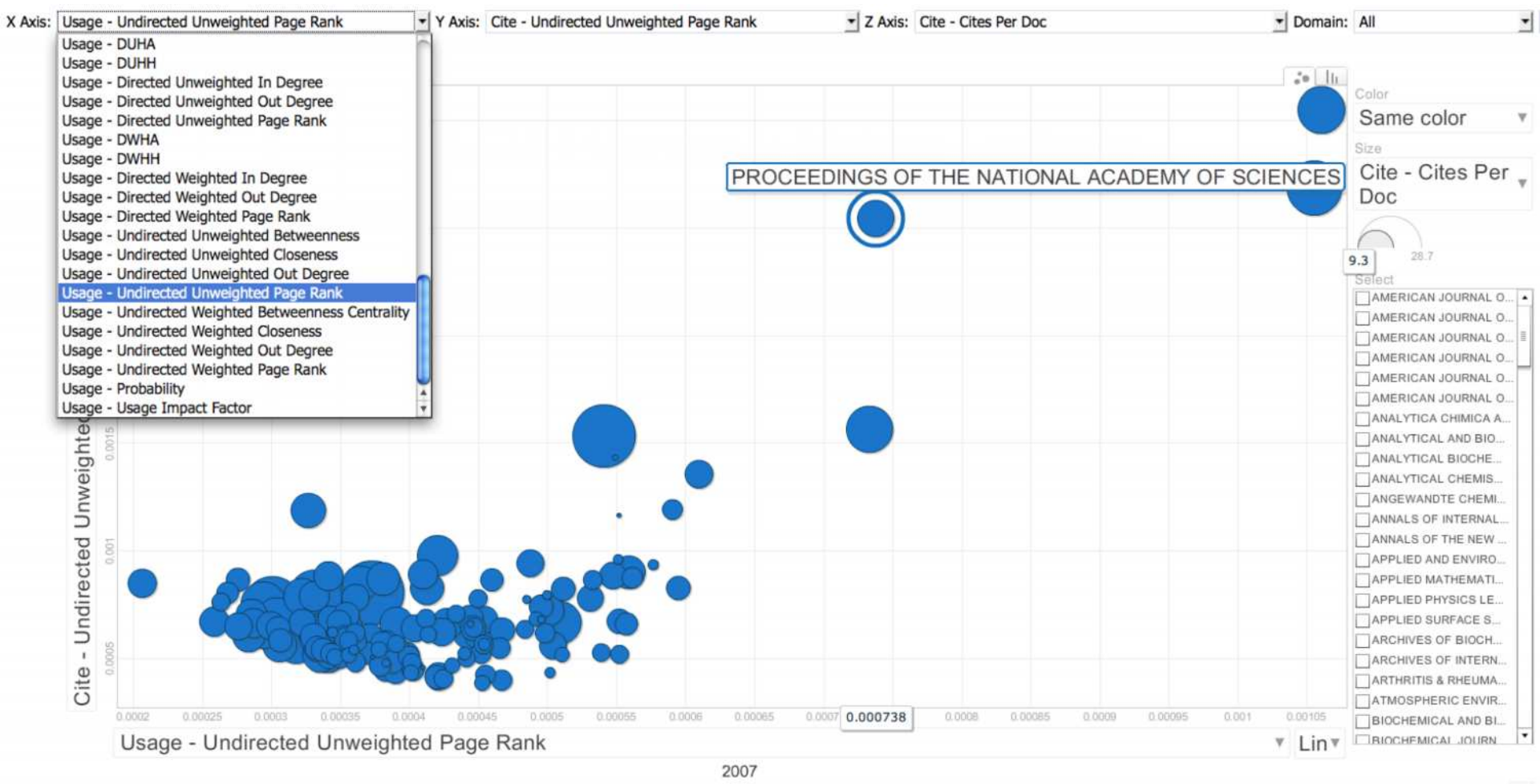


MESUR Services – http://www.mesur.org/services/

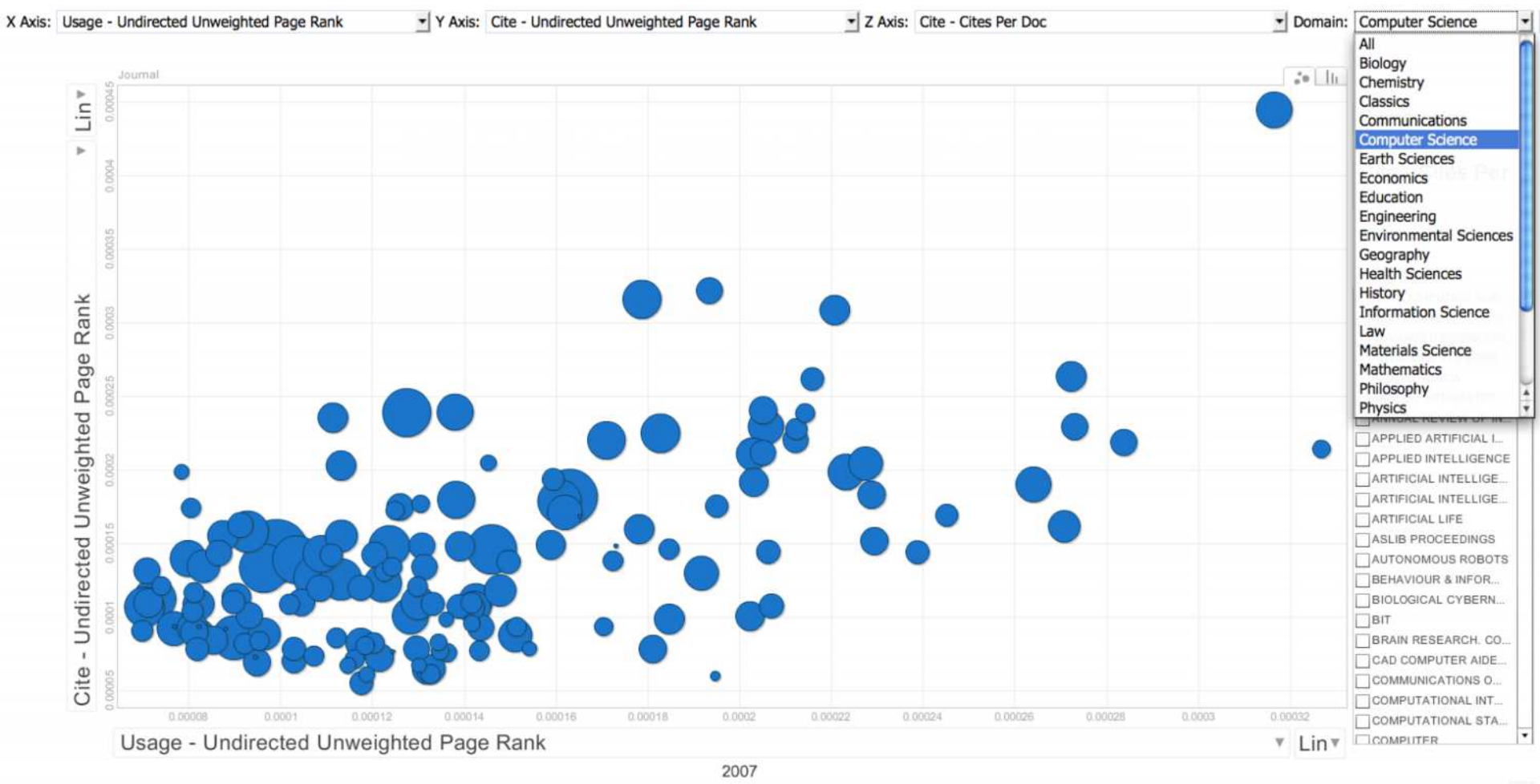




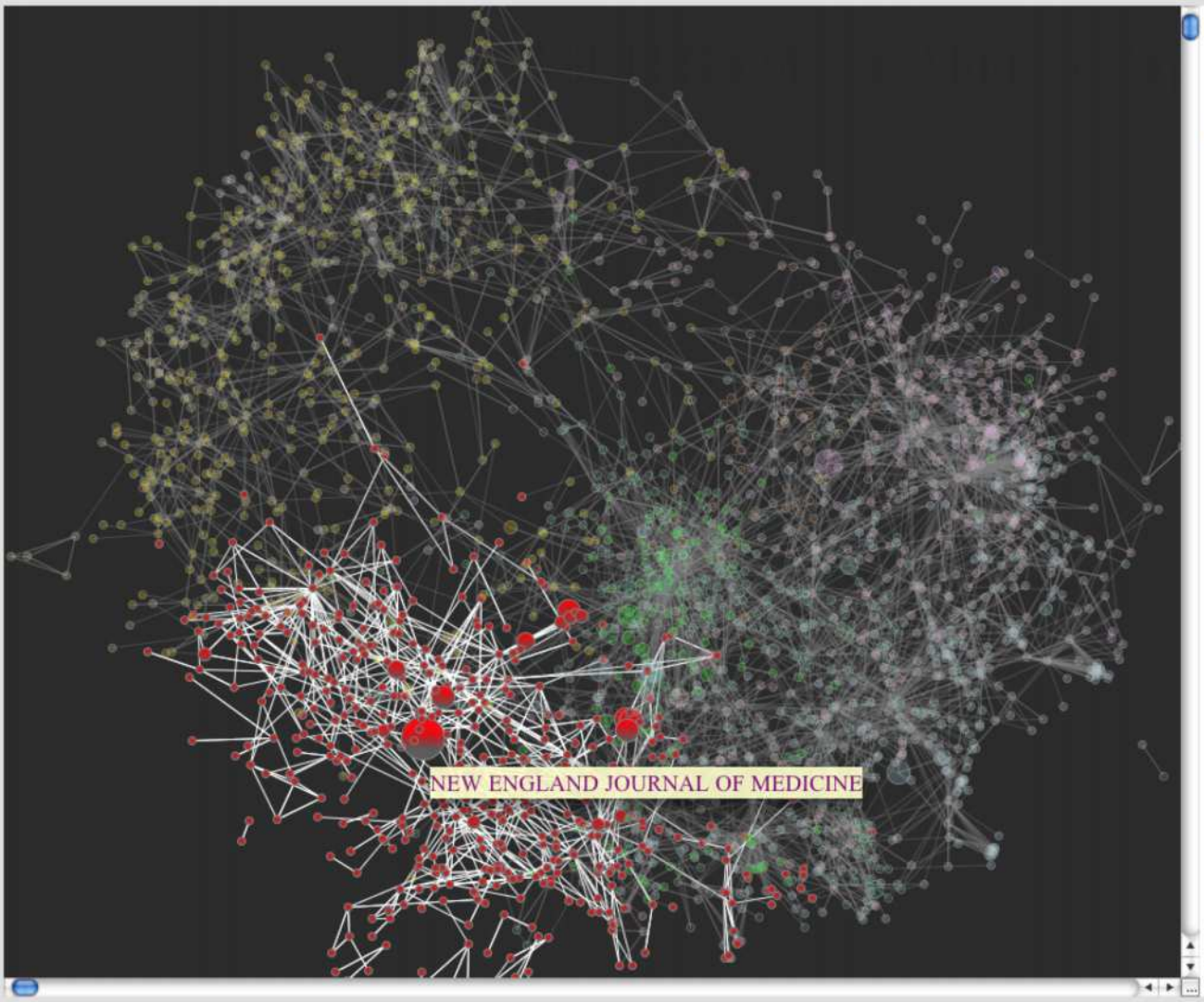
Journal	Year	Rank	Domain	Usage - Undirected Unweighted Page Rank	Cite - Undirected Unweighted Page Rank	Cite - Cites Per Doc
SCIENCE	2007	1	science	0.001060366	0.0030490425	15.4600000381
NATURE	2007	2	science	0.0010552662	0.0026860067	20.7700004578
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES	2007	3	science	0.0007375684	0.0025442564	9.2700004578
LANCET	2007	4	health sciences	0.0007331272	0.0015647924	15.0600004196
NEW ENGLAND JOURNAL OF MEDICINE	2007	5	health sciences	0.0005408772	0.0015339617	27.7099990845
JOURNAL OF BIOLOGICAL CHEMISTRY	2007	6	chemistry	0.0006096485	0.0013558392	5.4699997902
JAMA THE JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	2007	7	health sciences	0.0005490104	0.0014342124	0.200000003



Journal	Year	Rank	Domain	Usage - Undirected Unweighted Page Rank	Cite - Undirected Unweighted Page Rank	Cite - Cites Per Doc
SCIENCE	2007	1	science	0.001060366	0.0030490425	15.4600000381
NATURE	2007	2	science	0.0010552662	0.0026860067	20.7700004578
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES	2007	3	science	0.0007375684	0.0025442564	9.2700004578
LANCET	2007	4	health sciences	0.0007331272	0.0015647924	15.0600004196
NEW ENGLAND JOURNAL OF MEDICINE	2007	5	health sciences	0.0005408772	0.0015339617	27.7099990845
JOURNAL OF BIOLOGICAL CHEMISTRY	2007	6	chemistry	0.0006096485	0.0013558392	5.4699997902
JAMA THE JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	2007	7	health sciences	0.0005490104	0.0014342124	0.200000003



Journal	Year	Rank	Domain	Usage - Undirected Unweighted Page Rank	Cite - Undirected Unweighted Page Rank	Cite - Cites Per Doc
JOURNAL OF COMPUTATIONAL PHYSICS	2007	1	computer science	0.0003164522	0.0004445359	2.6700000763
PATTERN RECOGNITION	2007	2	computer science	0.0002722307	0.0002634107	1.8500000238
MATHEMATICAL AND COMPUTER MODELLING	2007	3	computer science	0.0003267292	0.0002140573	0.6399999857
COMMUNICATIONS OF THE ACM	2007	4	computer science	0.000220716	0.0003084775	1.8400000334
COMPUTERS & CHEMICAL ENGINEERING	2007	5	computer science	0.0002729736	0.000229205	1.4400000572
THEORETICAL COMPUTER SCIENCE	2007	6	computer science	0.0001934019	0.0003216609	1.3999999762
EXPERT SYSTEMS WITH APPLICATIONS	2007	7	computer science	0.0002836945	0.0002185048	1.4400000572



Find journal on the graph

ABDOMINAL IMAGING

Filter domain

health sciences

SubGraph

Collapse

Reset

+ Neighbors

Zoom

+

-

Magnifying Glass

Off

On

Edge threshold

0

10

Unbound journals

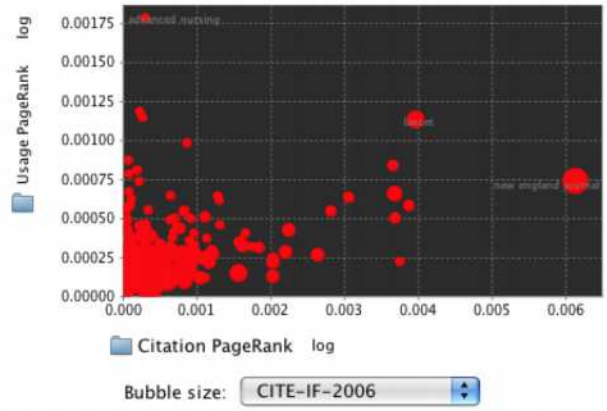
Off

On

Node size

CITE-IF-2006

Domain metrics



Bubble size: CITE-IF-2006

MESUR: the good ...

After 2 years of MESUR:

- Scientific exploration of metrics for scholarly evaluation
- Creation of large-scale reference data set
- Mapping science from the viewpoint of users: there **is** structure!
- Variety of Metrics that cover various aspects of scholarly impact and prestige
- MESUR dataset contains many more pearls for future research
- The (alternative / new) metrics issue is on many agendas. Usage data is on many agendas.

MESUR: the bad and the ugly …

Scalability of the approach:

- Lengthy negotiations to obtain log data
- No infrastructure standards: Recording, aggregating, normalization, ingestion, de-duplication,…
- No generally accepted policies: privacy, property, …
- No census data: when is a sample large and representative enough?

Quality control:

- Bots, Crawlers (detectable but never perfect)
- Cheating, manipulation (easier with usage statistics than network metrics)

Acceptance:

- Network-based usage metrics require session information. This is overlooked! As a result, will we end up with usage-based statistics only?

Publications related to MESUR

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, March 2009.

Johan Bollen, Herbert Van de Sompel, Aric HagBerg, Ryan Chute. **A principal component analysis of 39 scientific impact measures.** arXiv.org/abs/0902.2183

Johan Bollen, Herbert Van de Sompel, and Marko A. Rodriguez. **Towards usage-based impact metrics: first results from the MESUR project.** In Proceedings of the Joint Conference on Digital Libraries, Pittsburgh, June 2008

Marko A. Rodriguez, Johan Bollen and Herbert Van de Sompel. **A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage,** In Proceedings of the Joint Conference on Digital Libraries, Vancouver, June 2007

Johan Bollen and Herbert Van de Sompel. **Usage Impact Factor: the effects of sample characteristics on usage-based impact metrics.** (cs.DL/0610154)

Johan Bollen and Herbert Van de Sompel. **An architecture for the aggregation and analysis of scholarly usage data.** In Joint Conference on Digital Libraries (JCDL2006), pages 298-307, June 2006.

Johan Bollen and Herbert Van de Sompel. **Mapping the structure of science through usage.** Scientometrics, 69(2), 2006.

Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel. **Journal status.** Scientometrics, 69(3), December 2006 (arxiv.org:cs.DL/0601030)

Johan Bollen, Herbert Van de Sompel, Joan Smith, and Rick Luce. **Toward alternative metrics of journal impact: a comparison of download and citation data.** Information Processing and Management, 41(6):1419-1440, 2005.