An update on the MESUR project:
a large-scale survey of usage-based metrics of scholarly impact.

Johan Bollen and Herbert van de Sompel

Digital Library Research & Prototyping Team

Los Alamos National Laboratory - Research Library

jbollen@lanl.gov

Acknowledgements:

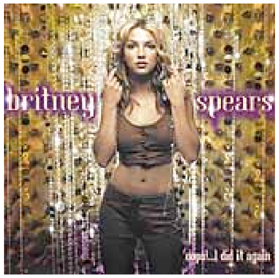
Marko A. Rodriguez (LANL), Ryan Chute (LANL),

Lyudmila L. Balakireva (LANL), Aric Hagberg (LANL), Luis Bettencourt (LANL)

Research supported by the Andrew W. Mellon Foundation.

Scholarly assessment: why more is not necessarily better…

So you want to know who's best?



83M **>** **50K**

↓ ?

↑ Who, Kinks, Byrds, Beatles → ↓ REM, Teenage Fanclub, Placebo, This Mortal Coil, Wilco

Silly?

We do the same in scholarly evaluation!

- Count citations, calculate IFs
- More citations > less citations

BUT that is **so** last century:

- relationships matter more than counts (cf Google)
- Web 2.0: social network thinking

Crucial distinction: data vs. statistics

- Data: who, what, when, how, …. Maintains sequence and context.
- Statistics: what, how much. Loss of most sequence and context.

How about usage data?

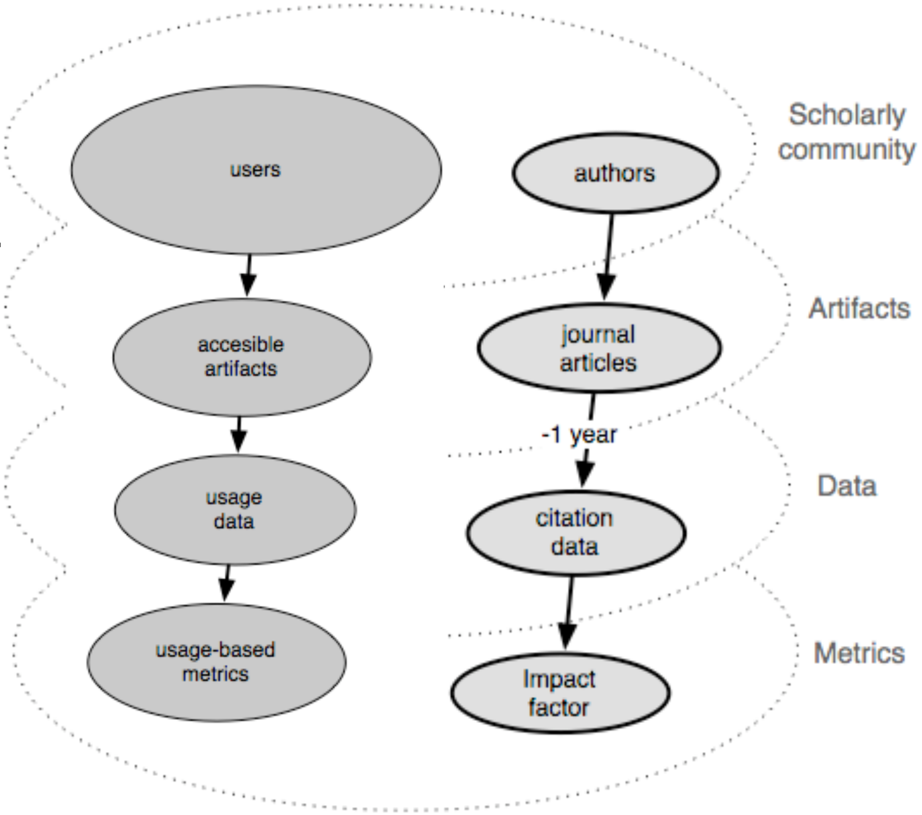
Citation data's limitations:

- Community: authors of journal articles.
- Artifacts: journal articles (8,000 journals?).
- Timing: +1 year publication delay.

Usage data's promise:

- Community: any user
- Artifacts: all that is accessible.
- Timing: recorded upon publication.

Hence, various initiatives focused on usage data: COUNTER, IRS, SUSHI, CiteBase



BEWARE: they're all statistics!
Counting usage is nearly as silly as counting citations (possibly even more so)!

We need to move to 21st century assessment paradigm suitable for e-science, open science:

1) usage data + 2) network metrics.

The promise of usage data and network metrics

LANL	Usage PR	IF (2003)	Title (abbv.)
1	60.196	7.035	PHYS REV LETT
2	37.568	2.950	J CHEM PHYS
3	34.618	1.179	J NUCL MATER
4	31.132	2.202	PHYS REV E
5	30.441	2.171	J APPL PHYS

→

CSU	Usage PR	IF (2003)	Title (abbv.)
1	78.565	21.455	JAMA-J AM MED ASSOC
2	71.414	29.781	SCIENCE
3	60.373	30.979	NATURE
4	40.828	3.779	J AM ACAD CHILD PSY
5	39.708	7.157	AM J PSYCHIAT

↓

MSR	Usage PR	IF (2005)	Title (abbv.)
1	15.830	30.927	SCIENCE
2	15.167	29.273	NATURE
3	12.798	10.231	PNAS
4	10.131	0.402	LECT NOTES COMP SCI
5	8.409	5.854	J BIOL CHEM

Counting citations doesn’t work very well.
Neither does counting usage…
But some network metrics do!

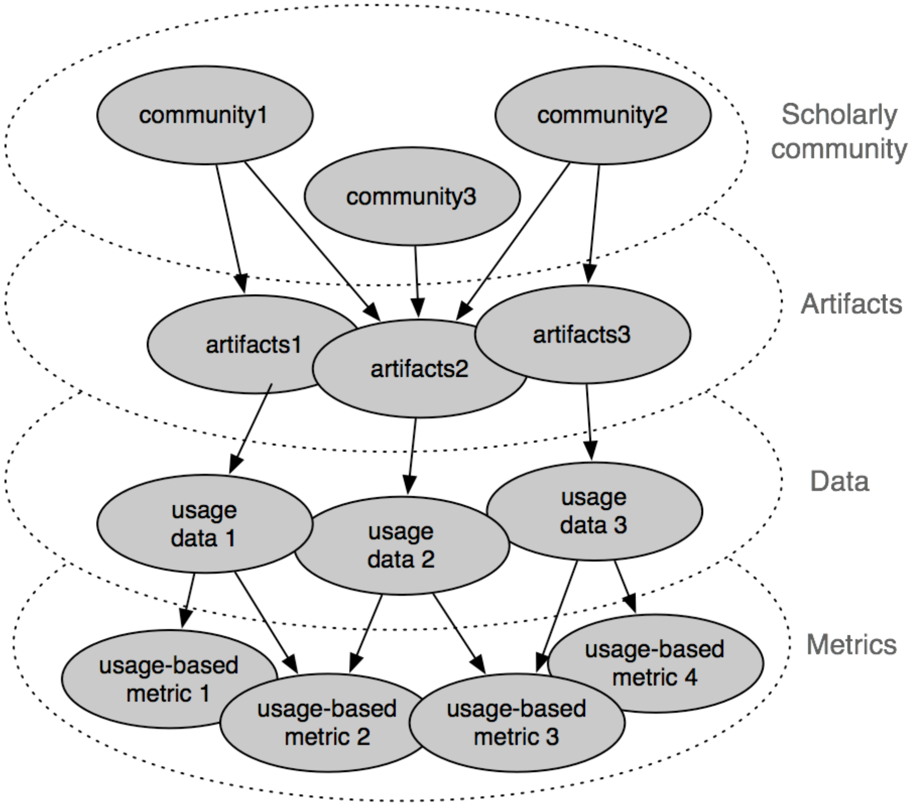
Also, note importance of:

- Community for which usage is recorded.
- Diversity for generalizable results.

Generalizable usage-based metrics?

Issues:

- Community: particular digital services.
- Artifacts: institutional policies and subscriptions.
- Data: particular sub-communities and collections of artifacts.
- Metrics: variety of possible metrics. What do they mean?



MESUR project

Previous/present studies of usage-data:

- Statistics: same mistake as citation analysis
- Single community usage data: not generalizable
- Small-scale investigations: reliability and validity?

MESUR:

- Beyond statistics: web 2.0/ social network metrics
- Multiple communities: generalizable results
- Large-scale investigation: 1,000,000,000 usage events, high reliability

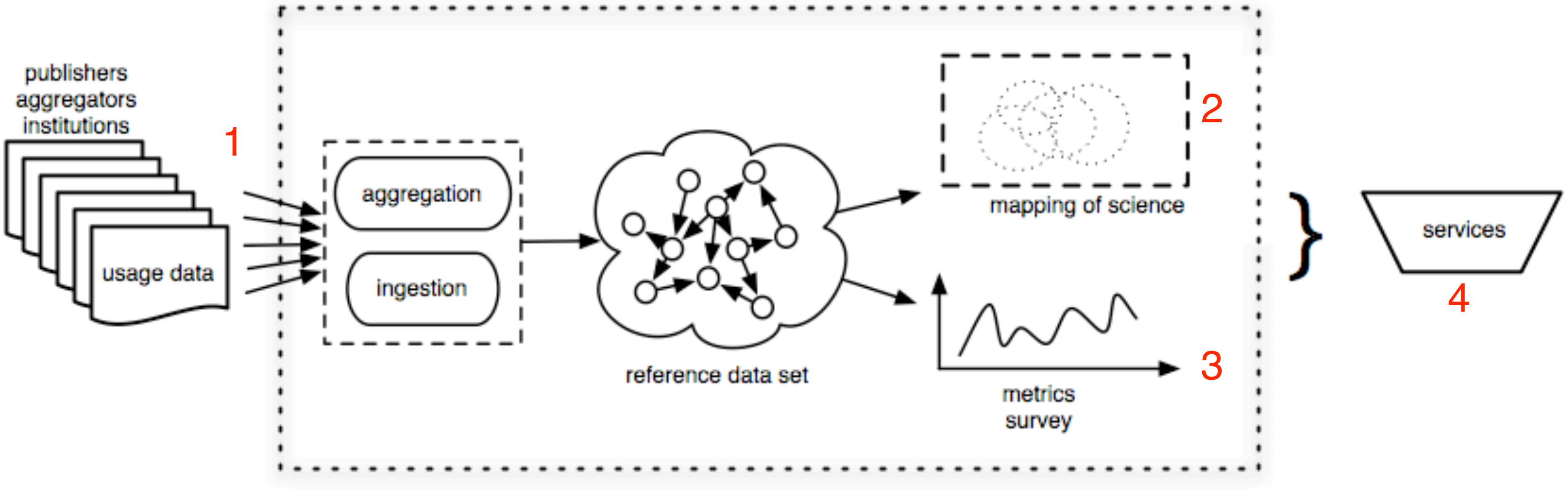
MESUR general approach

Generalizable, quantitative results

1. **Create very large-scale reference data set**
   1. Usage, citation and bibliographic data combined
   2. Various communities, various collections
2. **Investigate sampling issues:**
   1. Effects of sampling on usage-based assessment
   2. Mapping and characterization of scholarly community
   3. Uncertainty quantification: noise, bots, …
3. **Investigate validity of usage data and usage-based metrics**
   1. Cross-validation: compare to existing, accepted **journal**-focused metrics and data
   2. Not selling 1 metric: exploring many possibilities, many facets of impact
   3. Explorative approach: not top-down, bottom-up exploration
4. **Lay foundation for scientific, generalizable study** of usage data-based assessment

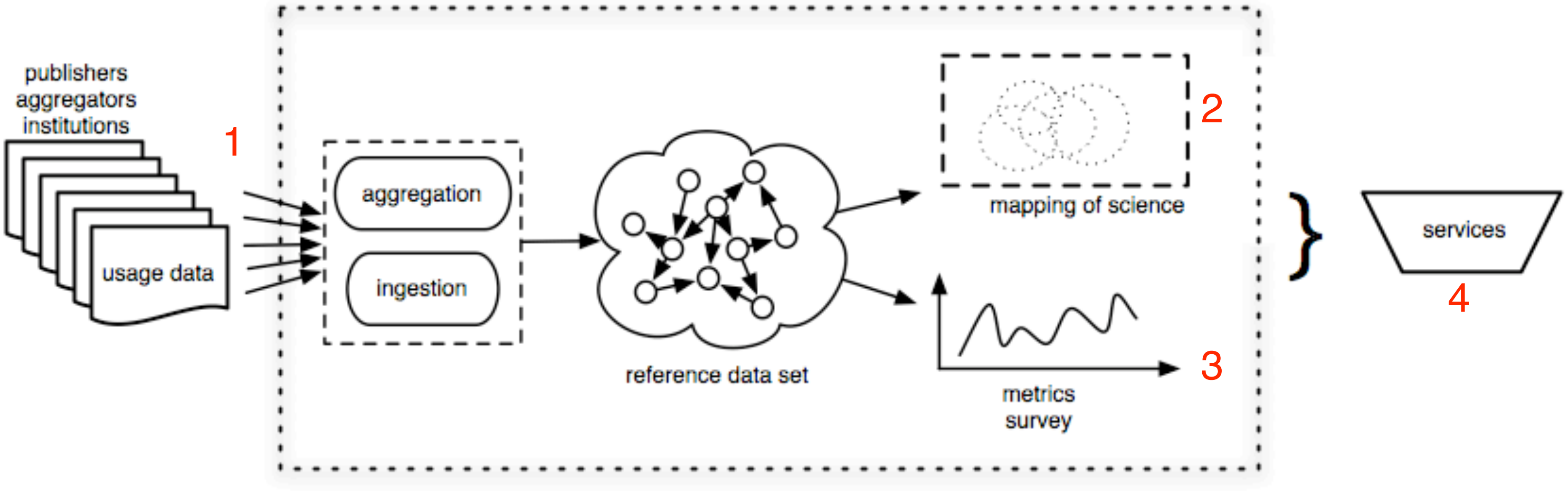
Presentation structure

1) **Usage data acquisition**
2) **Science mapping from usage graphs**
3) **Metrics survey**
4) **Services**
5) **Discussion**



Presentation structure

1) **Usage data acquisition**
2) Science mapping from usage graphs
3) Metrics survey
4) Services
5) Discussion

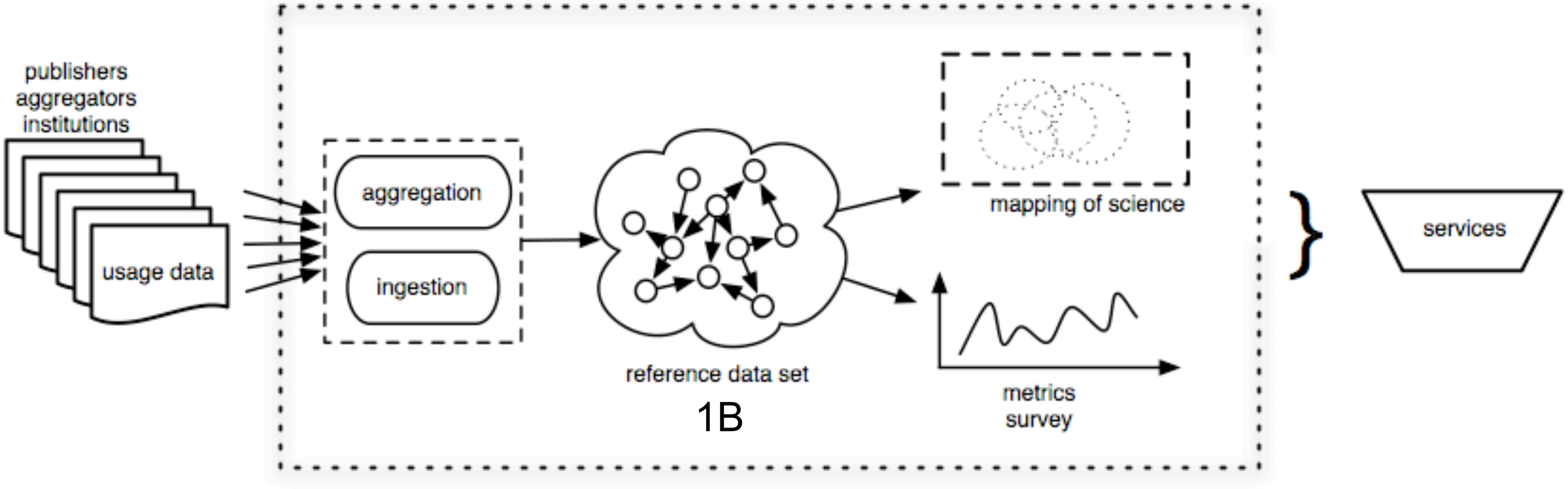


How to obtain 1,000,000,000 usage events?

Politely asked selected publishers, aggregators and institutional consortia for usage data:

- Scale: > 1,000,000,000 usage events and +500,000,000 citations
- Period: 2002-2007, but mostly 2006
- Span:
  - > 50M documents
  - > 100,000 journals (inc. newspapers, magazines,…)

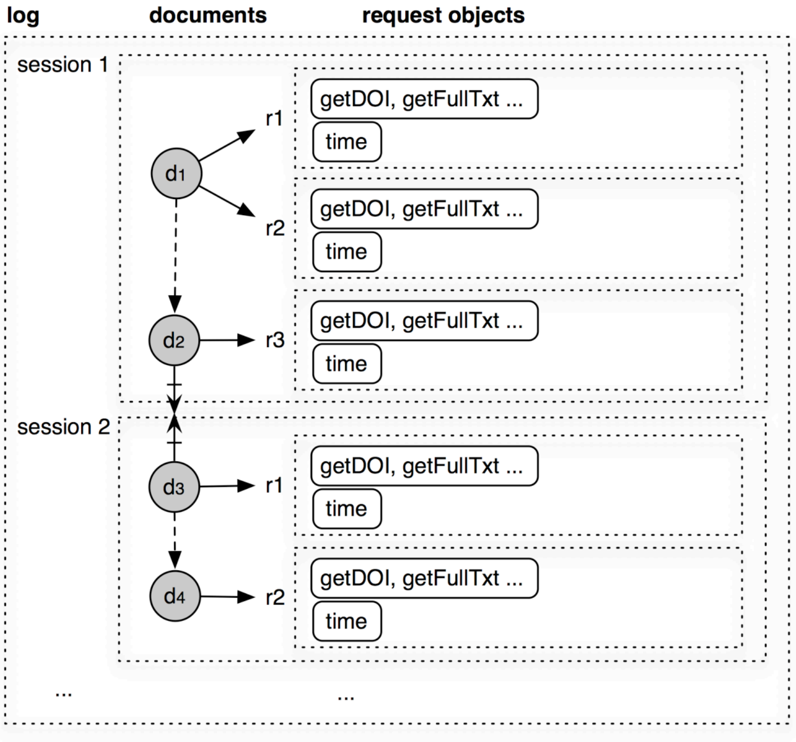
Use reference data set for mapping, metrics and services



Data normalization and ingestion

Minimal requirements for all usage data

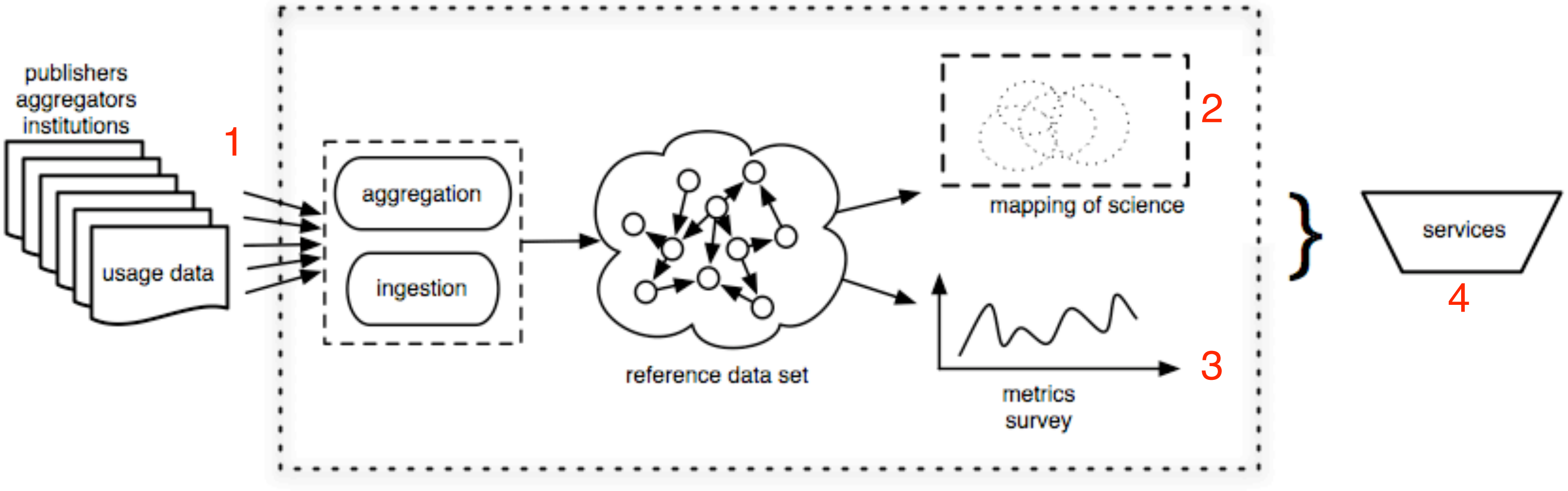
- Unique usage events (article level)
- Fields: unique session ID, date/time, unique document ID and/or metadata, request type
- Note difference with usage statistics



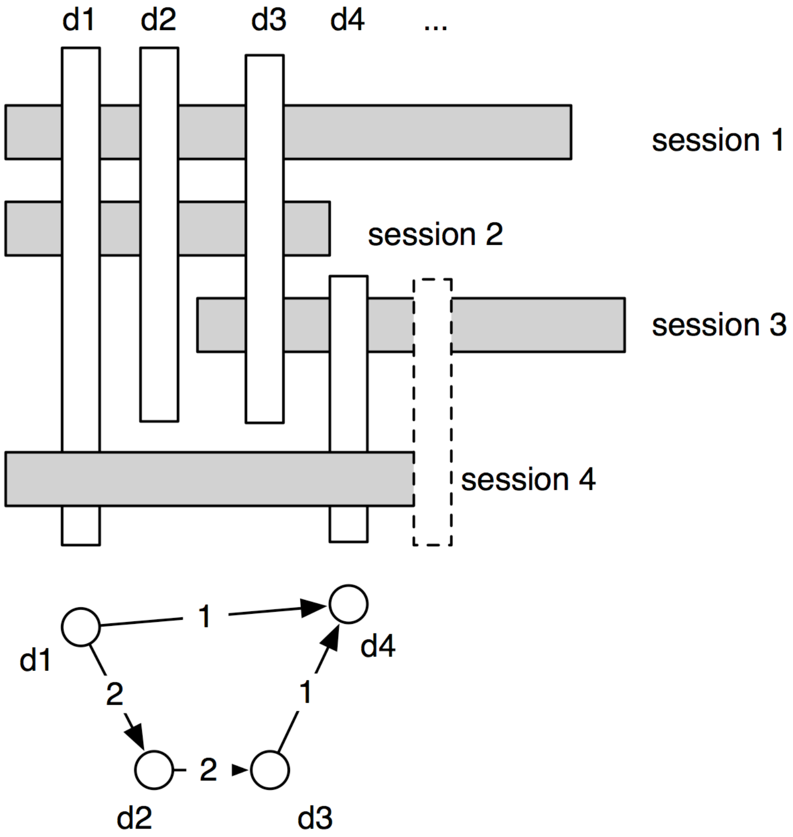
```
2007   9   1   0   0   1    CFA   cffoe   A172080.N1.Vanderbilt.Edu     unknown AST    A                                   .com
2007   9   1   0   0   1    CFA   cffoe   210.94.41.89   unknown PHY    A     2007ApPhL.90
2007   9   1   0   0   1    CFA   cffoe   24-196-228-125.dhcp.gwnt.ga.charter.com unknown                                   cholar.goo
2007   9   1   0   0   4    CFA   cffoe   163.152.35.114  4700387eae     PHY    A     1993W
2007   9   1   0   0   6    CFA   cffoe   pd9e980fc.dip0.t-ipconnect.de   45f0c69881     AST
2007   9   1   0   0   1    CFA   cffoe   A172080.N1.Vanderbilt.Edu     unknown AST    A                                   ogle.com
2007   9   1   0   0   1    CFA   cffoe   210.94.41.89   unknown PHY    A     2007ApPhL.90
2007   9   1   0   0   1    CFA   cffoe   24-196-228-125.dhcp.gwnt.ga.charter.com unknown                                   ://scholar.
2007   9   1   0   0   4    CFA   cffoe   163.152.35.114  4700387eae     PHY    A     1993W                                n
2007   9   1   0   0   6    CFA   cffoe   pd9e980fc.dip0.t-ipconnect.de   45f0c69881     AST
2007   9   1   0   0   6    CFA   cffoe   foel25144.4u.com.gh    47002f8eda     PHY    A                                   /www.goog
2007   9   1   0   0   6    CFA   cffoe   66-215-171-214.dhcp.ccmn.ca.charter.com 4681d22a                                  =2001P%
2007   9   1   0   0   7    CFA   cffoe   nat-ptouser3.uspto.gov  unknown PHY    A     2005                                 .com
2007   9   1   0   0   7    CFA   cffoe   cpe-71-65-25-115.ma.res.rr.com  unknown PHY    A                                 ogle.com
2007   9   1   0   0   7    CFA   cffoe   customer3491.pool1.unallocated-106-0.orangehomed                                  ...19.883V
2007   9   1   0   0   8    CFA   cffoe   Uranus.seas.ucla.edu    46672d96b2     PHY    A                                  om
2007   9   1   0   0   9    CFA   cffoe   75-121-173-37.dyn.centurytel.net      46cf1fd8a6                                   http://foea
2007   9   1   0   0   13    CFA   cffoe   foel17-18.kln.forthnet.gr      unknown AST    A     1
2007   9   1   0   0   15    CFA   cffoe   hades.astro.uiuc.edu    46f707564d     PRE    A                                 edu
2007   9   1   0   0   17    CFA   cffoe   ool-43554752.dyn.optonline.net  unknown PHY    A                               ogle.com
2007   9   1   0   0   17    CFA   cffoe   c-68-33-176-222.hsd1.md.comcast.net     unknown G                                  http://ww
2007   9   1   0   0   19    CFA   cffoe   74-36-139-46.dr02.brvl.mn.frontiernet.net      unknow                           code=200
    amp;nr_to_return=500&amp;data_and=YES&amp;toc_link=YES       http://foeabs.edu
2007   9   1   0   0   19    CFA   cffoe   c-76-16-53-120.hsd1.il.comcast.net      46f667b71b                             ?1916PA..
    whole_paper=YES&amp;type=PRINTER&amp;filetype=.pdf       http://foeabs.edu
2007   9   1   0   0   20    CFA   cffoe   74-39-37-62.nas03.roch.ny.frontiernet.net      unknown PHY     E      2007JSTEd.tmp..29B     http://dx.doi.org/10.1007/s10972-007-9067-2     http://foe
2007   9   1   0   0   22    ANU    bio-mirror     uatu-virtual1.anu.edu.au       46f9e8f87f      AST     A      2006ApJ..647.128E     http://foe.grangenet.net/abs/2006ApJ..647.128E       http://foe.g
2007   9   1   0   0   22    CFA   cffoe   fw.hia.nrc.ca   46f1531d59     AST    A      2002P&SS..50.745H    http://foeabs.edu/abs/2002P%26SS..50.745H    http://foeabs.edu
2007   9   1   0   0   22    CFA   cffoe   24-117-0-220.cpe.cableone.net   unknown AST     A      1984BITA..15.268S    http://foeabs.edu/abs/1984BITA..15.268S    http://www.google.com
2
```

Presentation structure

1) Usage data acquisition
2) **Science mapping from usage graphs**
3) Metrics survey
4) Services
5) Discussion



How to generate a usage network.



Same session ~ documents relatedness

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

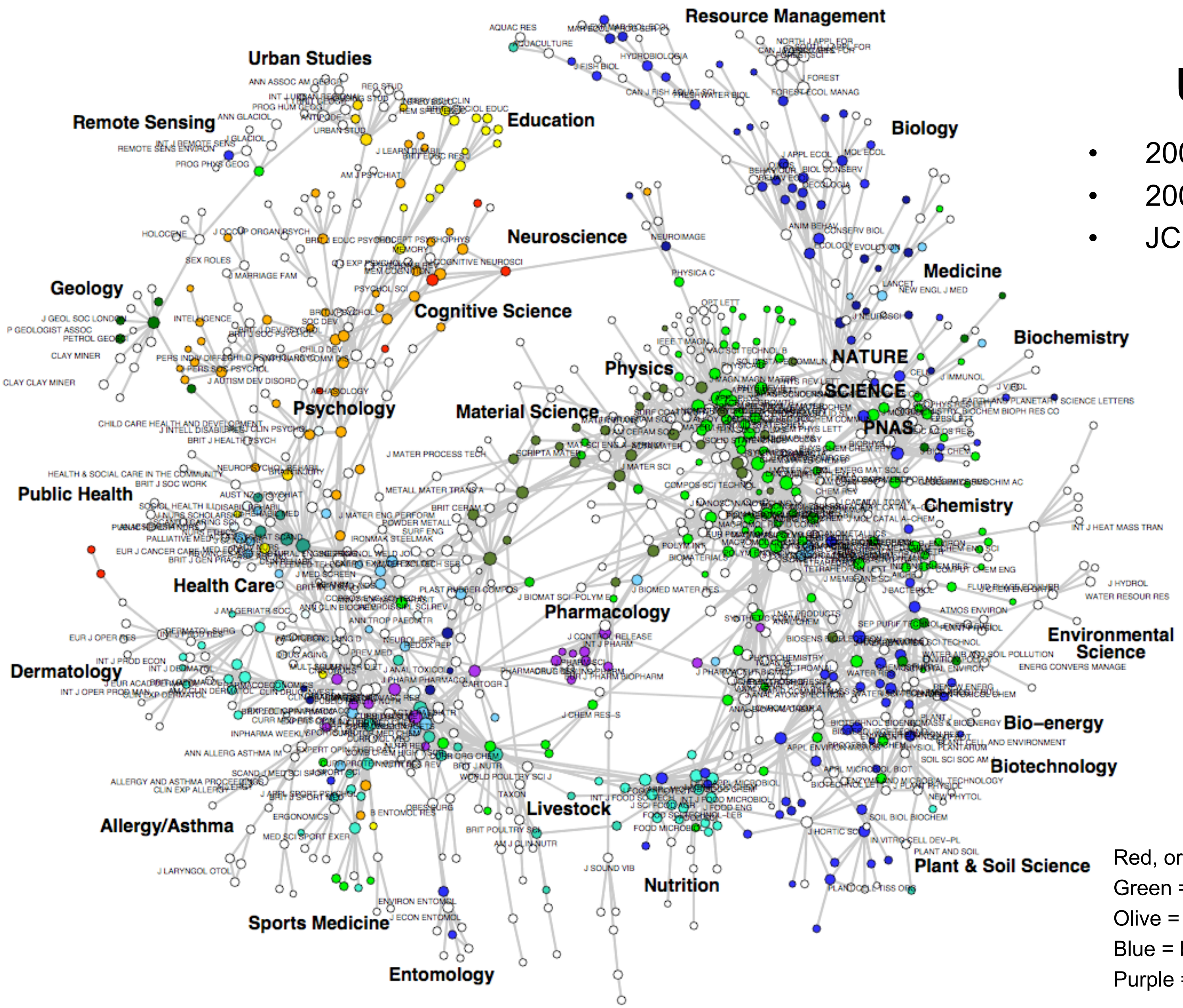
Usage data is on article level:

- Works for journals **and** articles
- **Anything** for which usage was recorded

Note: not something we invented: association rule learning in data mining. Beer and diapers! Netflix, amazon.com, etc.

Usage map

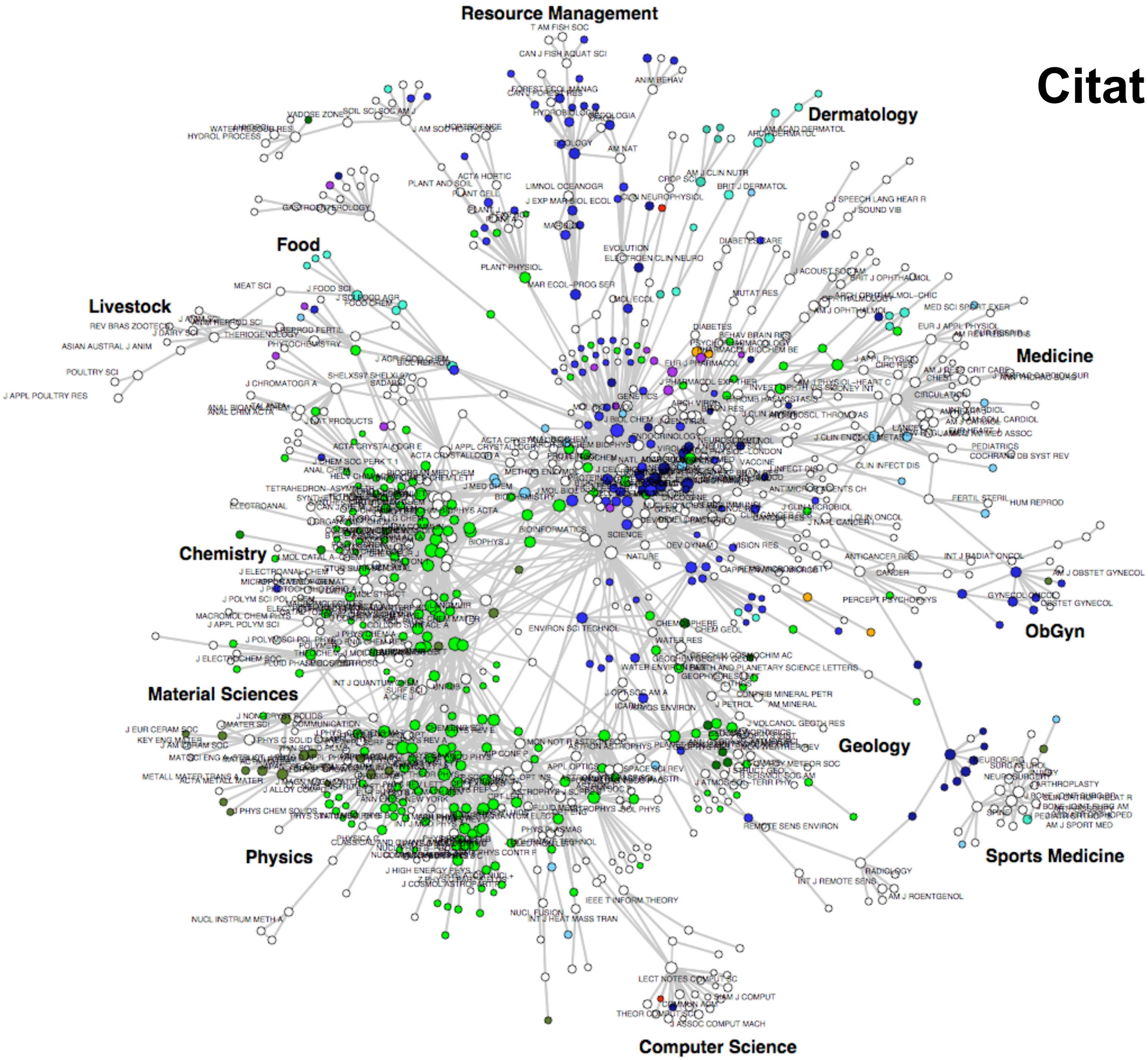
- 200M usage events
- 2006 usage only
- JCR journals (+-7600)



Red, orange= psych, cogn
Green = phys, chem
Olive = material science
Blue = biology
Purple = pharma

Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
@ Minneapolis, CNI 2008, April 2008

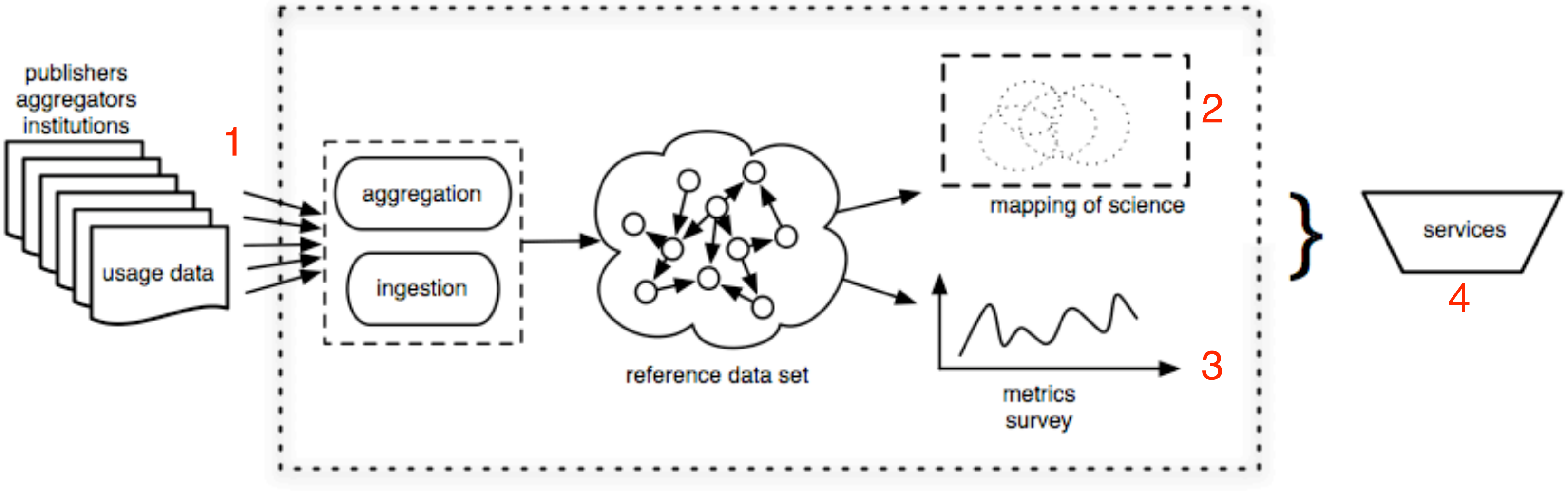
Citation map



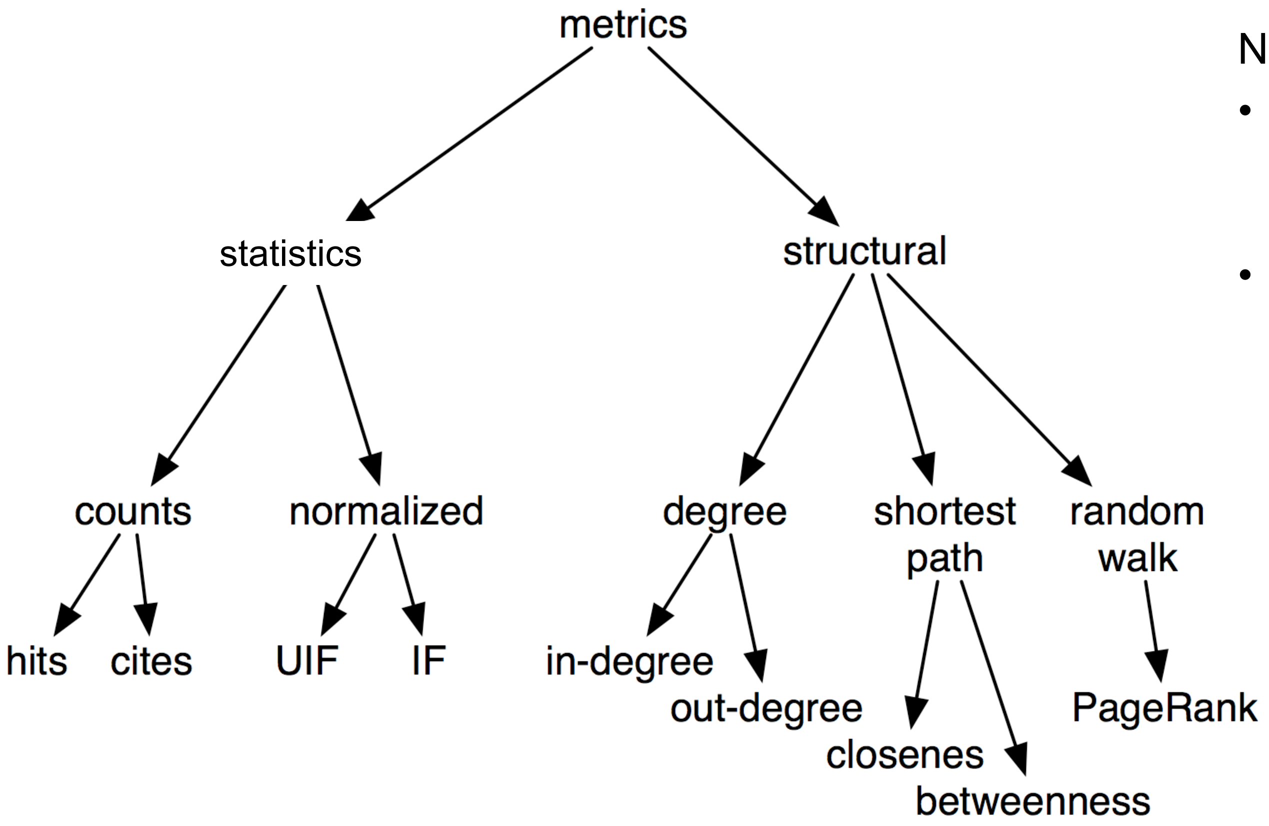
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
@ Minneapolis, CNI 2008, April 2008

Presentation structure

1) Usage data acquisition
2) Science mapping from usage graphs
3) **Metrics survey**
4) Services
5) Discussion



Metric types



Note:

- Metrics can be calculated both on citation and usage data
- Structural metrics require networks:
  - E.g. JCR
  - Usage: created from MESUR data

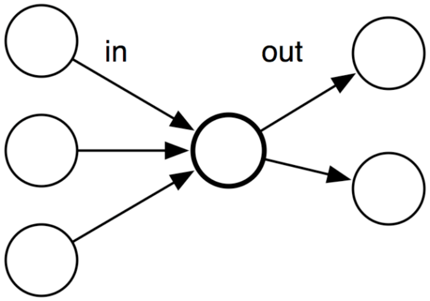
Structural metrics calculated from usage graph

Classes of metrics:

- Degree
- Shortest path
- Random walk
- Distribution

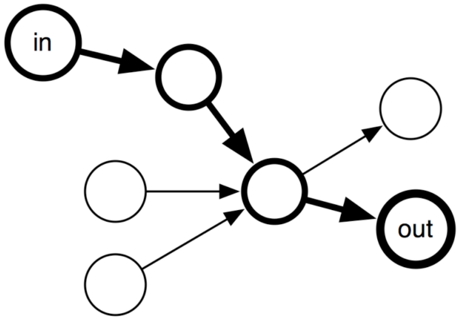
Degree

- In-degree
- Out-degree



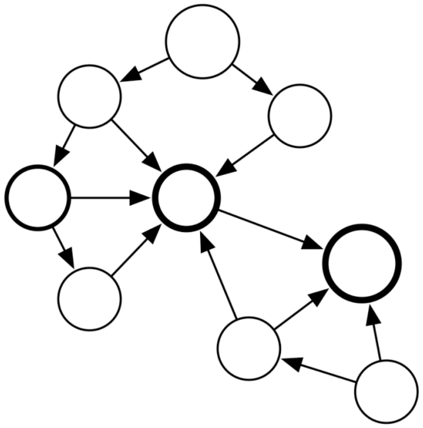
Shortest path

- Closeness
- Betweenness
- Newman



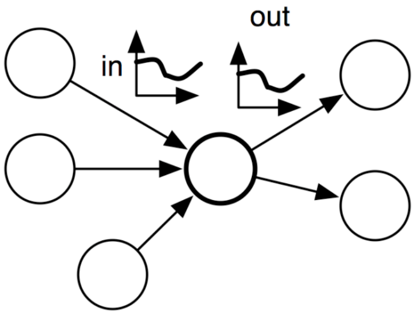
Random walk

- PageRank
- Eigenvector



Distribution

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



Each can be defined to take into account weights by e.g. means of weighted shortest path definition

Set of metrics calculated on MESUR data set

List of metrics:
Citation network from JCR 2004

- CITE-BE
- CITE-ID
- CITE-IE
- CITE-IF
- CITE-OD
- CITE-OE
- CITE-PG
- CITE-UBW
- CITE-UBW-UN
- CITE-UCL
- CITE-UCL-UN
- CITE-UNM
- CITE-UNM-UN
- CITE-UPG
- CITE-UPR
- CITE-WBW
- CITE-WBW-UN
- CITE-WCL
- CITE-WCL-UN
- CITE-WID
- CITE-WNM
- CITE-WNM-UN
- CITE-WOD
- CITE-WPR

Usage-based metrics:
MESUR 2006

- USES-BE,
- USES-ID
- USES-IE
- USES-OD
- USES-OE
- USES-PG
- USES-UBW
- USES-UBW-UN
- USES-UCL
- USES-UCL-UN
- USES-UNM
- USES-UNM-UN
- USES-UPG
- USES-UPR
- USES-WBW
- USES-WBW-UN
- USES-WCL
- USES-WCL-UN
- USES-WID
- USES-WNM
- USES-WNM-UN
- USES-WOD
- USES-WPR

Usage graph creation: Wenzhong Zhao
Metrics: Marko Rodriguez and Aric Hagberg

Citation network rankings

2004 Impact Factor

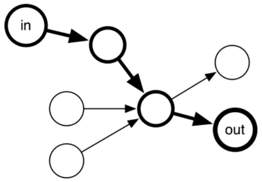
value	journal
1 49.794	CANCER
2 47.400	ANNU REV IMMUNOL
3 44.016	NEW ENGL J MED
4 33.456	ANNU REV BIOCHEM
5 31.694	NAT REV CANCER

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

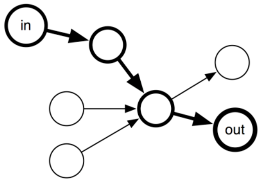
betweenness

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



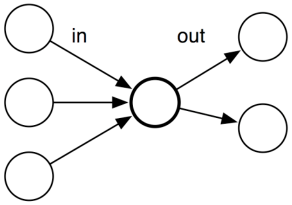
Closeness

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



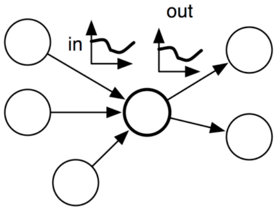
In-Degree

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

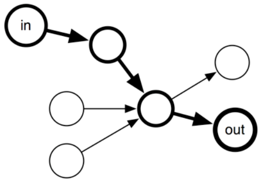
value	journal
1 49.794	CANCER
2 47.400	ANNU REV IMMUNOL
3 44.016	NEW ENGL J MED
4 33.456	ANNU REV BIOCHEM
5 31.694	NAT REV CANCER

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

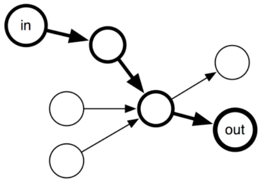
betweenness

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



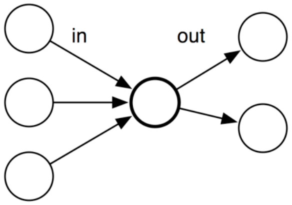
Closeness

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



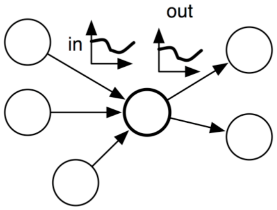
In-Degree

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

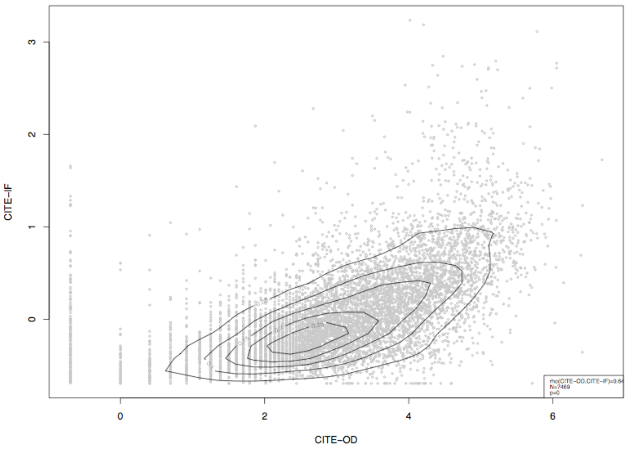


In-degree entropy

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

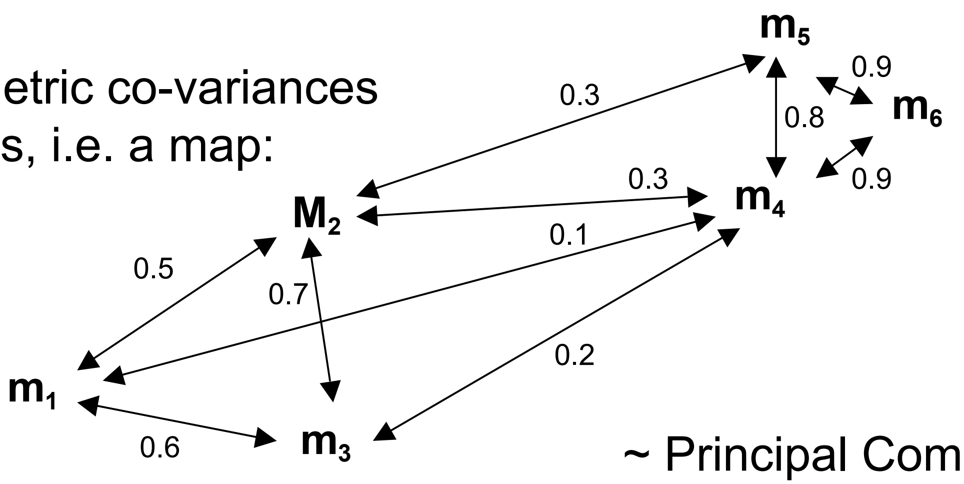


Metric correlations: metric maps



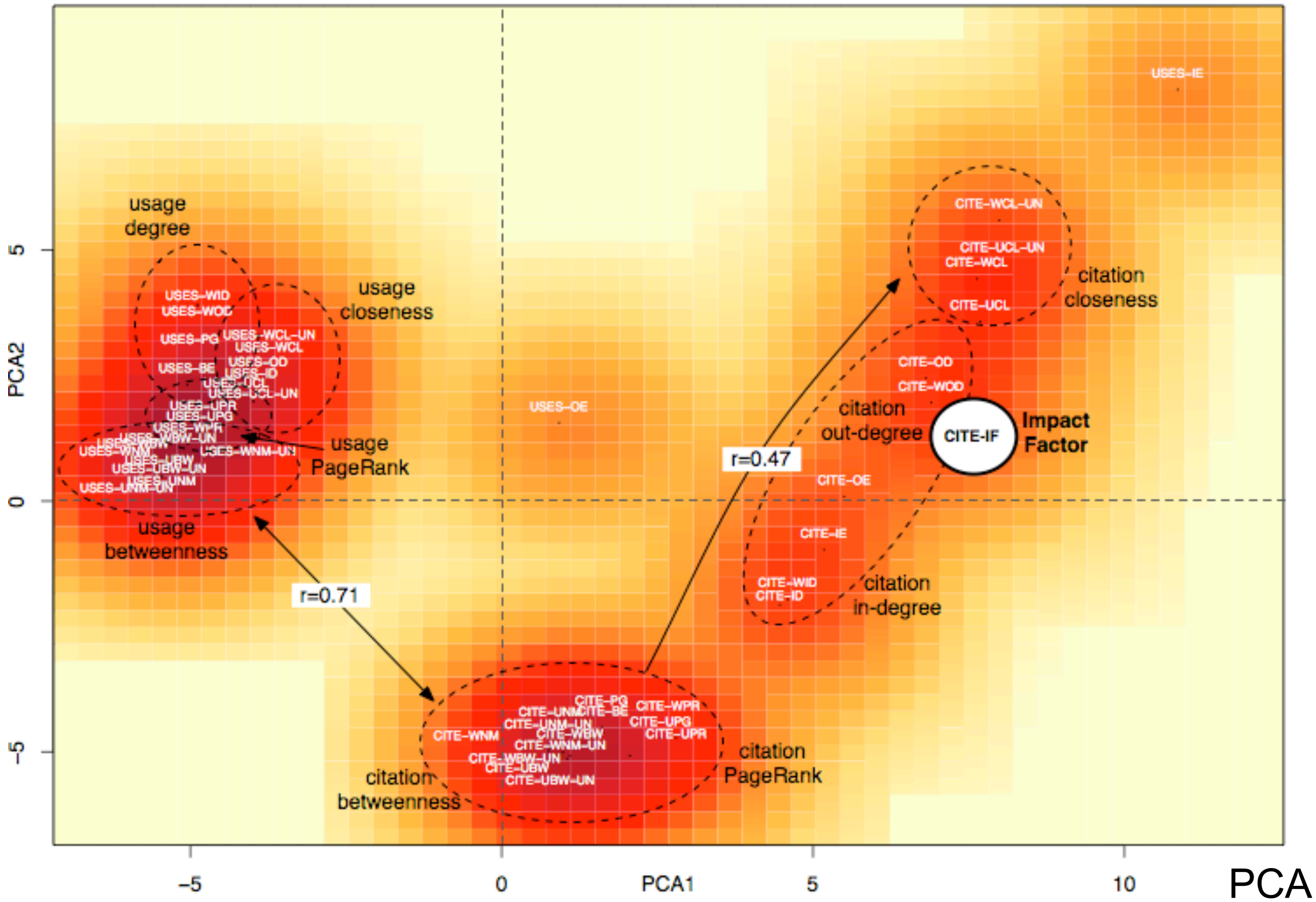
```
    m1   m2   m3   m4   m5   m6   m7   m8  m9   m10
m1 1.00 0.75 0.67 0.61 0.46 0.57 0.99 0.79 0.79 0.40
m2 0.75 1.00 0.96 0.81 0.82 0.83 0.73 0.68 0.69 0.77
m3 0.67 0.96 1.00 0.77 0.77 0.81 0.65 0.62 0.63 0.72
m4 0.61 0.81 0.77 1.00 0.64 0.67 0.60 0.50 0.51 0.64
m5 0.46 0.82 0.77 0.64 1.00 0.92 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
m1 0.40 0.77 0.72 0.64 0.89 0.77 0.39 0.54 0.55 1.00
 0
```

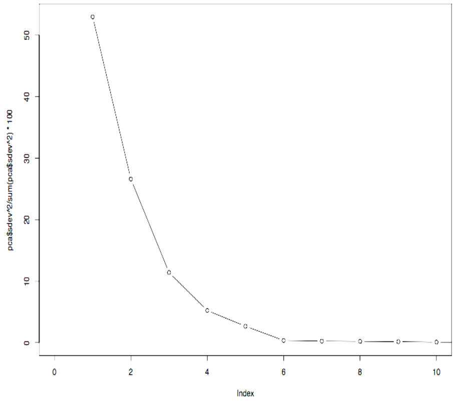
Mapping n x n metric co-variances to 2D coordinates, i.e. a map:



~ Principal Component Analysis

Metrics relationship

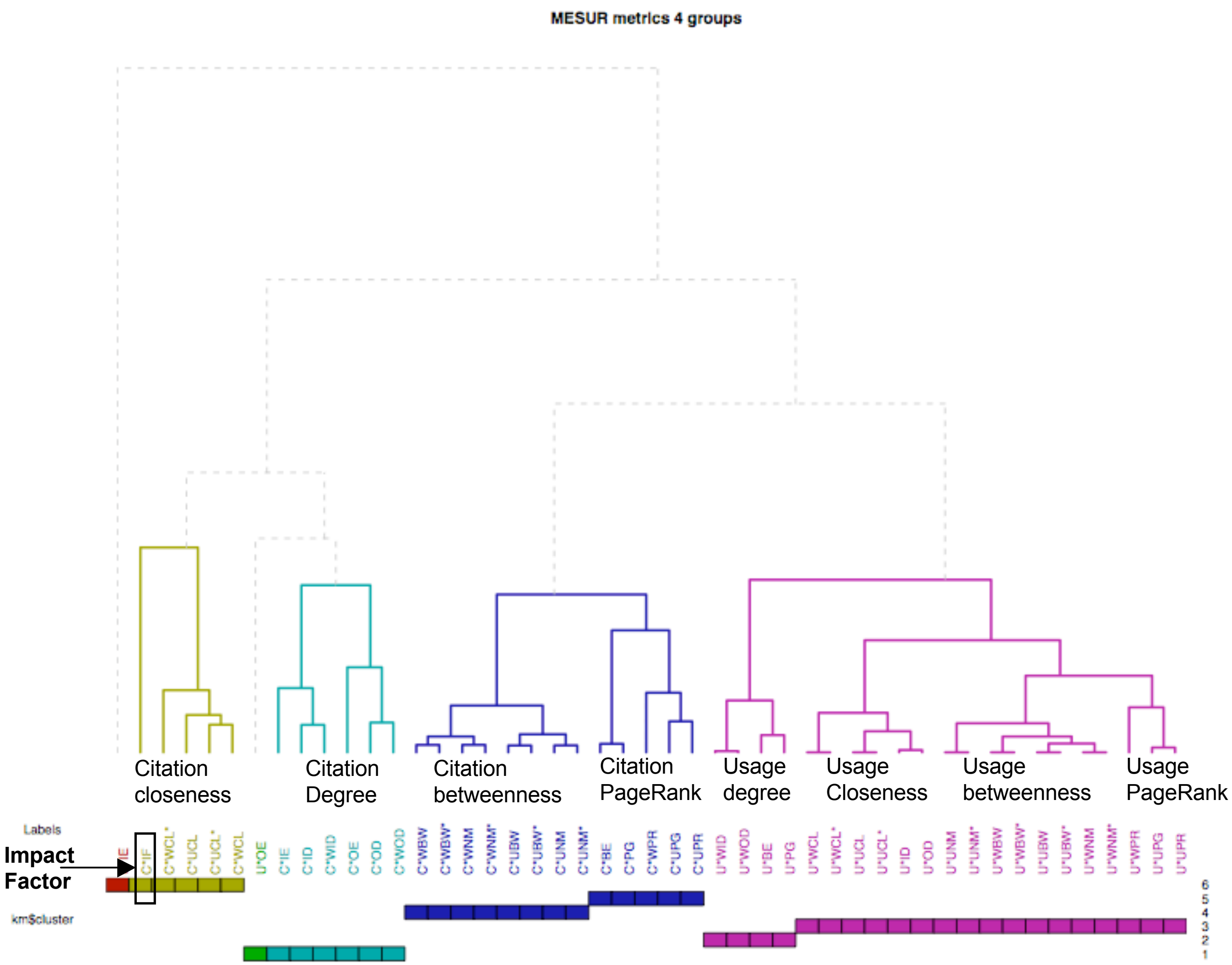




PCA1 ~ Usage vs. Citation

PCA ~ Prestige vs. Popularity

Hierarchical cluster analysis



Citation closeness

Citation Degree

Citation betweenness

Citation PageRank

Usage degree

Usage Closeness

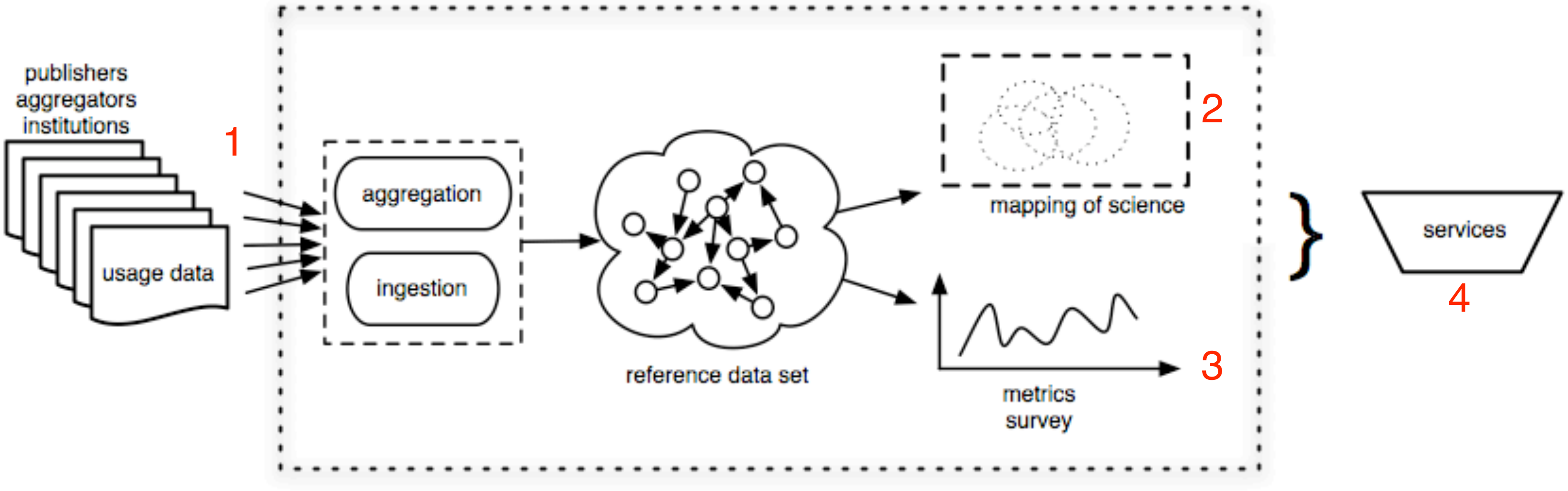
Usage betweenness

Usage PageRank

Impact Factor

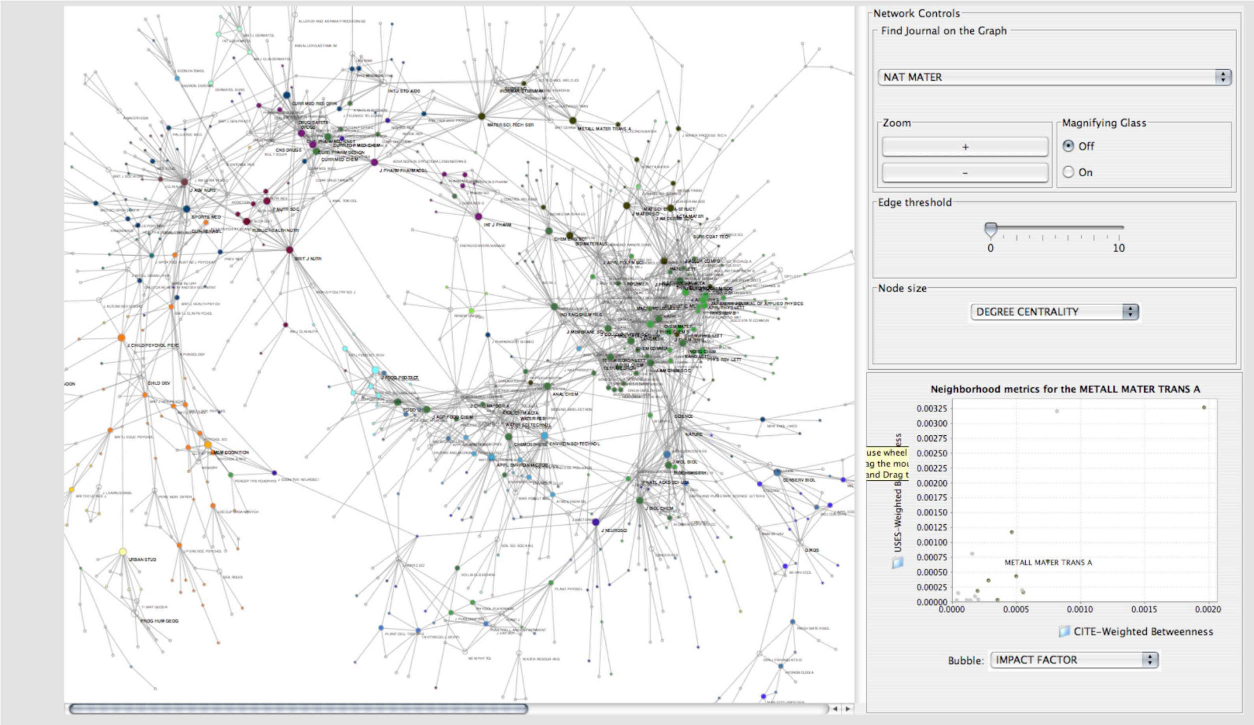
Presentation structure

1) Usage data acquisition
2) Science mapping from usage graphs
3) Metrics survey
4) **Services**
5) Discussion



MESUR explorer prototype

- Based on MESUR usage data collection
- Explore large-scale usage maps of science
- Explore journal rankings according to multiple metrics of interest



Network Controls

Find Journal on the Graph: NAT MATER

Zoom: + -

Magnifying Glass: ● Off ○ On

Edge threshold: 0 – 10

Node size: DEGREE CENTRALITY

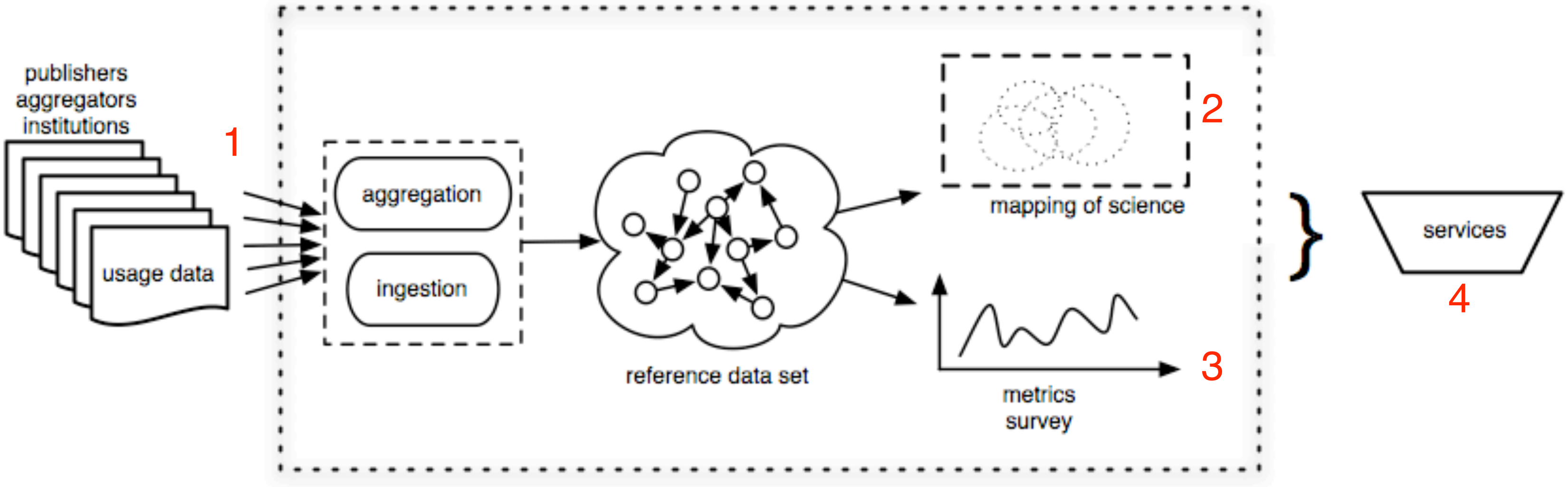
Neighborhood metrics for the METALL MATER TRANS A

Bubble: IMPACT FACTOR

http://www.mesur.org/mesurexplorer_jbollen042008.mov

Presentation structure

1) Usage data acquisition
2) Science mapping from usage graphs
3) Metrics survey
4) Services
5) **Discussion**



MESUR: conclusions.

After 1.5 year of MESUR:

- First scientific exploration of new paradigm in scholarly evaluation
- Creation of single largest reference data set of usage, citation and bibliographic data
- Infrastructure for a continued research program

Conclusions:

- Beyond discussion of merits and validity
- Usage data works! Mapping, metrics, …

Challenges:

- Standardization: recording, aggregating, normalization
- Applications in other domains: IRs/ORs
- Community acceptance and advocacy: simple is not always better, smart is always better.

Some relevant publications.

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.