The MESUR project: an update from the trenches.

Johan Bollen
Digital Library Research & Prototyping Team
Los Alamos National Laboratory - Research Library

jbollen@lanl.gov

Acknowledgements:
Herbert Van de Sompel (LANL), Marko A. Rodriguez (LANL), Lyudmila L. Balakireva (LANL)
Wenzhong Zhao (LANL), Aric Hagberg (LANL)

Research supported by the Andrew W. Mellon Foundation.

NISO Usage Data Forum • November 1-2, 2007 • Dallas, TX
Citations and the journal impact factor.

Citation data:
- Golden standard of scholarly evaluation
- Citation = scholarly influences.
- Extracted from published materials.
- Main bibliometric data source for scholarly evaluation.

**Journal Impact Factor: mean 2-year citation rate**
2003 citations to 2001 and 2002 articles in X divided by number of articles published in X in 2001 and 2002

**Widely applied**
- Fair approximation of journal “status”,…but
- Used to rank authors, departments, institutions, regions, nations, etc.
- Now common in tenure, promotion and other evaluation procedures!
From ranking to assessment

Ranking 0.6:
- single criterion of value
- single data source, e.g. citation data
- impact: frequentist, e.g. average citation rates (ISI Impact Factor)

“Assessment 3.0”:
- situate item in value landscape
- multiple sources of scholarly information, e.g. citation, usage and bibliographic data.
- impact: relational indicators: trust, reputation, centrality, etc.

1 significant problem: which dimensions to choose?

- best
- good
- not so good
- bad

Mean citation rate
Many possible facets of scholarly impact
The MESUR project.

Johan Bollen (LANL): Principal investigator.
Herbert Van de Sompel (LANL): Architectural consultant.
Aric Hagberg (LANL): Mathematical and statistical consultant.
Lyudmila Balakireva (LANL): Database management and development.
Wenzhong Zhao (LANL): Data processing, normalization and ingestion.

“The Andrew W. Mellon Foundation has awarded a grant to Los Alamos National Laboratory (LANL) in support of a two-year project that will investigate metrics derived from the network-based usage of scholarly information. The Digital Library Research & Prototyping Team of the LANL Research Library will carry out the project. The project's major objective is enriching the toolkit used for the assessment of the impact of scholarly communication items, and hence of scholars, with metrics that derive from usage data.”
Usage data

Citation data pertain to 4 levels in the scholarly communication process:
- Community: authors of journal articles.
- Artifacts: journal articles.
- Data: citation data (+1 year publication delay).
- Metrics: mean citation rate rules supreme.
- Scale: expensive to extract.

However, for usage data:
- Community: all users including most authors.
- Artifacts: all that is accessible.
- Data: recorded upon publication.
- Metrics: a range of web and web2.0 inspired metrics, e.g. clickstream and datamining.
- Scale: automatically recorded at point of service.

Hence, various initiatives focused on usage data: COUNTER, IRS, SUSHI, CiteBase. But where are the metrics?
Divergence and convergence.

Convergence!

• Open research questions:
  o Is this guaranteed?
  o To what? A common-baseline?
• What we do know:
  o Institutional perspective can be contrasted to baseline.
  o As aggregation increases in size, so does value.
  o Cross-validation is key.
MESUR\textsuperscript{1}: Metrics from Scholarly Usage of Resources.

Andrew W. Mellon Foundation funded study of usage-based metrics (2006-2008)

Executed at the Digital Library Research and Prototyping team, Los Alamos National Laboratory Research Library

Objectives:
1. Create a model of the scholarly communication process.
2. Create a large-scale reference data set (semantic network) that relates all relevant bibliographic, citation and usage data according to (1).
4. **Survey usage-based metrics on basis of reference data set.**
Project data flow and work plan.

1. publishers aggregators institutions
   - raw usage events
   - processing
   - usage data

2. aggregation
3. ingest
4. metrics survey
   - reference data set
Project timeline.

We are here!
Presentation structure:
an update on the MESUR project

1) Usage data characterization
2) Analysis
   1) Usage graphs
   2) Metrics analysis
3) Results
4) Discussion
Presentation structure:
an update on the MESUR project

1) Usage data characterization
2) Analysis
   1) Usage graphs
   2) Metrics analysis
3) Results
4) Discussion
Usage data providers

Publisher recorded usage data:
- COUNTER statistics
- Item-level data

Aggregator recorded usage data:
- COUNTER statistics
- Item-level data

Institutionally recorded usage data:
- SFX logs
- Custom log formats (XML)
Usage data collected

Data providers:
- Publishers: 8
- Aggregators: 3
- Institutions: 4

- Data: > 1B usage events and 1B citations
  - At this point, 247,083,481 usage events loaded
  - Another +1,000,000,000 on the way
- Documents: > 50M documents
- Journals: 326,000
- Community: > 100M users and authors combined
Data acquired: timelines

Span:
- Majority: -1 years
- Some minor 2002-2003 data

Sharing models:
- Historical data for period \([t-x,t]\)
- Periodical updates

Main issues:
- Restoration of archives
- Digital preservation issues
  - All data fields intact
  - Integration of various sources of usage data
Presentation structure: an update on the MESUR project

1) Usage data characterization
2) Analysis
   1) Usage graphs
   2) Metrics analysis
3) Results
4) Discussion
MESUR’s usage data representation framework.

Assumptions:
1. *Sessions* identify sequences of events (same user - same document)
2. *Documents* tied to aggregate *request objects*
3. *Request objects* consist of *series of service requests* and the date and time at which request took place.

Implications:
1. Sequence preserved.
2. Most usage data and statistics can be reconstructed from framework
3. Lends itself to XML and RDF format
4. Permits request type filtering

Example:
1. COUNTER stats= aggregate request counts for each document (journal) grouped by date/time (month)
2. Usage graph: overlay sessions with same document pairs

• Out of 13 MESUR providers so far, only 3 natively follow this model.
• The usage data of another 8 contains the necessary information for conversion
Implications for structural analysis of usage data

- Sequence preservation allows:
  - Reconstruction of user behavior
  - Usage graphs!
- Statistics do not allow this type of analysis BUT are useful for:
  - Validating results
  - Rankings
How to generate a usage graph.

Documents are associated by co-occurrence in same session
• Same session, same user: common interest
• Frequency of co-occurrence in same session: degree of relationship

Usage data:
• Works for journals and articles
• Anything for which usage was recorded

Options:
• Strict pair-wise sequence
• All within session

Note: not something we invented. Association rule learning in data mining. Beer and diapers!
Usage graphs

MESUR graph created:
- 200M usage events
- Usage restricted to 2006
- Journals clipped to 7600 2004 JCR journals
- Pair-wise sequences
  - Within session, only consecutive pairs
  - Raw frequency weights
Tracing the flow of information.

Marko Rodriguez
Presentation structure:
an update on the MESUR project

1) Usage data characterization
2) Analysis
   1) Usage graphs
   2) Metrics analysis
3) Results
4) Discussion
Metric types

Note:
- Metrics can be calculated both on citation and usage data
- Structural metrics require graphs
  - Citation graph, e.g. 2004 JCR
  - Usage graph, e.g. created by MESUR

Wasserman (1994) Social network analysis
Frequentist metrics

Raw cites:
- Count number of citations to document or journal
- Count number of times document or journal was accessed

Normalized:
- Journal Impact factor:
  - Number of citations to journal
  - Divided by number of articles published in journal
- Usage Impact Factor
  - Number of request for journal or article
  - Divided by number of articles published in journal

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, i.e. an unweighted and weighted variant.
Social network metrics: different aspects of impact I

- **In-degree/IF**
- **Degree metrics**
- **Degree centrality**
- **Closeness centrality**
- **Shortest path metrics**
- **Betweenness centrality**

---

Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
@ November, 2007, NISO - Dallas, TX
Social network metrics: different aspects of impact II

Random walk
Metrics, e.g. PageRank

Basic idea:
- Random walkers follow edges
- + Probability of random teleportation
- Visitation numbers converge ~ PageRank
- “Stationary Probability distribution”

\[
PR(p_i) = \frac{1 - d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}
\]

From wikipedia.org
Presentation structure: an update on the MESUR project

1) Usage data characterization
2) Analysis
   1) Usage graphs
   2) Metrics analysis
3) Results
4) Discussion
Set of metrics calculated on MESUR data set

List of metrics:
- 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
Overlaps and discrepancies: Usage Impact Factor
Overlaps and discrepancies

Citation PageRank vs. citation IF

For Computer Science only

- Map metrics themselves
- We can examine structure of metric correlations and arrive at a structural model of metric relationships
- Metric semantics -> what do they tell us about scholarly impact?
Citation rankings

<table>
<thead>
<tr>
<th>2004 Impact Factor</th>
<th>Citation Pagerank</th>
<th>betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>value</strong></td>
<td><strong>journal</strong></td>
<td><strong>value</strong></td>
</tr>
<tr>
<td>1 49.794</td>
<td>CANCER</td>
<td>1 0.0116</td>
</tr>
<tr>
<td>2 47.400</td>
<td>ANNU REV IMMUNOL</td>
<td>2 0.0111</td>
</tr>
<tr>
<td>3 44.016</td>
<td>NEW ENGL J MED</td>
<td>3 0.0108</td>
</tr>
<tr>
<td>4 33.456</td>
<td>ANNU REV BIOCHEM</td>
<td>4 0.0101</td>
</tr>
<tr>
<td>5 31.694</td>
<td>NAT REV CANCER</td>
<td>5 0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Closeness</th>
<th>In-Degree</th>
<th>In-degree entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>value</strong></td>
<td><strong>journal</strong></td>
<td><strong>value</strong></td>
</tr>
<tr>
<td>1 7.02e-05</td>
<td>PNAS</td>
<td>1 3448</td>
</tr>
<tr>
<td>2 6.72e-05</td>
<td>LECT NOTES COMPUT SC</td>
<td>2 3182</td>
</tr>
<tr>
<td>3 6.43e-05</td>
<td>NATURE</td>
<td>3 2913</td>
</tr>
<tr>
<td>4 6.37e-05</td>
<td>SCIENCE</td>
<td>4 2190</td>
</tr>
<tr>
<td>5 6.37e-05</td>
<td>J BIOL CHEM</td>
<td>5 2160</td>
</tr>
</tbody>
</table>
## Usage rankings

### 2004 Impact Factor

<table>
<thead>
<tr>
<th>Value</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CANCER</td>
</tr>
<tr>
<td>2</td>
<td>ANNU REV IMMUNOL</td>
</tr>
<tr>
<td>3</td>
<td>NEW ENGL J MED</td>
</tr>
<tr>
<td>4</td>
<td>NNU REV BIOCHEM</td>
</tr>
<tr>
<td>5</td>
<td>NAT REV CANCER</td>
</tr>
</tbody>
</table>

### Citation Pagerank

<table>
<thead>
<tr>
<th>Value</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SCIENCE</td>
</tr>
<tr>
<td>2</td>
<td>NATURE</td>
</tr>
<tr>
<td>3</td>
<td>PNAS</td>
</tr>
<tr>
<td>4</td>
<td>LECT NOTES COMPUT SC</td>
</tr>
<tr>
<td>5</td>
<td>J BIOL CHEM</td>
</tr>
</tbody>
</table>

### Betweenness

<table>
<thead>
<tr>
<th>Value</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SCIENCE</td>
</tr>
<tr>
<td>2</td>
<td>NATURE</td>
</tr>
<tr>
<td>3</td>
<td>PNAS</td>
</tr>
<tr>
<td>4</td>
<td>LECT NOTES COMPUT SC</td>
</tr>
<tr>
<td>5</td>
<td>LANCET</td>
</tr>
</tbody>
</table>

### Closeness

<table>
<thead>
<tr>
<th>Value</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SCIENCE</td>
</tr>
<tr>
<td>2</td>
<td>NATURE</td>
</tr>
<tr>
<td>3</td>
<td>PNAS</td>
</tr>
<tr>
<td>4</td>
<td>LECT NOTES COMPUT SC</td>
</tr>
<tr>
<td>5</td>
<td>BIOCHEM BIOPH RES CO</td>
</tr>
</tbody>
</table>

### In-Degree

<table>
<thead>
<tr>
<th>Value</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SCIENCE</td>
</tr>
<tr>
<td>2</td>
<td>NATURE</td>
</tr>
<tr>
<td>3</td>
<td>PNAS</td>
</tr>
<tr>
<td>4</td>
<td>LECT NOTES COMPUT SC</td>
</tr>
<tr>
<td>5</td>
<td>J BIOL CHEM</td>
</tr>
</tbody>
</table>

### In-degree entropy

<table>
<thead>
<tr>
<th>Value</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MED HYPOTHESES</td>
</tr>
<tr>
<td>2</td>
<td>PNAS</td>
</tr>
<tr>
<td>3</td>
<td>LIFE SCI</td>
</tr>
<tr>
<td>4</td>
<td>LANCET</td>
</tr>
<tr>
<td>5</td>
<td>INT J BIOCHEM CELL B</td>
</tr>
</tbody>
</table>
Metrics relationship
Metrics relationships

- Citation and usage metrics reveal an entirely different pattern
  - Citation is split in 2 sections:
    - Degree metrics (right)
    - Shortest path and random walk (left)
  - Usage is split in 4 clusters:
    - Degree metrics
    - PageRank and entropy
    - Closeness
    - Betweenness

- Usage pattern can be caused
  - Noise in usage graph
  - Higher density of usage/nodes
Hierarchical analysis
Presentation structure:
an update on the MESUR project

1) Usage data characterization
2) Analysis
   1) Usage graphs
   2) Metrics analysis
3) Results
4) Discussion
MESUR: an update

Usage data:
• Creation of single largest reference data set of usage, citation and bibliographic data
• +1,000,000,000 usage events loaded in next month
• Usage data obtained from multiple publishers, aggregators and institutions
• Infrastructure for a continued research program in this domain
• Results will guide scholarly evaluation and may help produce standards for usage data representation

Usage graphs:
• Adequate data model for item-level usage data naturally leads to this
• Reduced distortion compared to raw usage: structure counts, not raw hits
• Several option on how to create: MESUR investigates option

Metrics:
• Frequentist and structural metrics
• Each can represent different facets of scholarly impact
• Simple metrics can produce adequate results, e.g. in-degree vs. PageRank. How far should we go?
• Note increasing convergence of usage-metrics to citation metrics as sample increases.
Some relevant publications.


