Network Analysis of Data Reuse in the Social Sciences

Kathleen Fear
IASSIST 2012
June 6, 2012
Intro

• Knowledge transfer within and between fields
  – Citation analysis and bibliometrics

• Studying data reuse in the social sciences

• My goal: using citation analysis to characterize reuse patterns within the social sciences
What do I mean by reuse?

• Using data you did not collect yourself
  – Secondary analysis (or reanalysis or replication)
    (Chao & Weber 2011)

• What about interdisciplinary reuse?

  “one of the most productive and inspiring of human pursuits”

  (Committee on Facilitating Interdisciplinary Research, National Academy of Sciences, National Academy of Engineering, Institute of Medicine, 2004, p. 1)
Multidisciplinary
reuse

Integrative reuse
Hypotheses

• *H1:* Multidisciplinarity will be more common than interdisciplinarity.

• *H2:* For the case of both multidisciplinarity and interdisciplinarity, citations will be primarily to nearby disciplines rather than distant ones.
About ‘distance’

• Citation studies of interdisciplinarity show that while interdisciplinarity is increasing, most interdisciplinary citations are to neighboring fields.
Source data: ICPSR bibliography

- Citation analysis requires a corpus of citation data

- ICPSR closely tracks reuse of data holdings
Source data: ICPSR bibliography

• Comprised of citations from books, book sections, conference publications, journal articles, reports, and theses

• Each entry includes: authors, title, dataset(s) cited, year published, etc.

• Some caveats:
  – Non-standard citation practices increase possibility of missing citations
  – Limited to journal citations to avoid duplication
  – Excluded analysis of author discipline
## Citations to ICPSR data, 2006 - 2011

<table>
<thead>
<tr>
<th></th>
<th>Full Dataset</th>
<th>Study sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of items</td>
<td>7973</td>
<td>5850</td>
</tr>
<tr>
<td>Number of data citations</td>
<td>15,787</td>
<td>12,532</td>
</tr>
<tr>
<td>Datasets cited per publication</td>
<td>Mean = 1.98</td>
<td>Mean = 2.14</td>
</tr>
<tr>
<td></td>
<td>(range: 1 – 121)</td>
<td>(range: 1 – 121)</td>
</tr>
</tbody>
</table>
Metrics

• Multidisciplinarity
  – How many disciplines are represented among the citing papers?

• Integrative index
  – How often is the dataset co-cited with other datasets?

• Distance
  – 0 or 1
Method: Network analysis

• Full network:
  – Nodes: publications and datasets
  – Edges: citation from a publication to a dataset

This publication cites these datasets
Method: Network analysis

- Data co-citation network
  - Nodes: datasets
  - Edges: co-citation in a single publication

These datasets are cited together in the same publication.
Findings: Reuse overall

• Of 1684 datasets represented in the sample, 928 (55%) were cited more than once
  – Mean: 8.37
  – Median: 2, 3rd Q: 5

• 660 are in giant connected component
Findings: Multidisciplinary reuse

- 715 datasets cited in papers belonging to more than one discipline
  - Mean: 2.81; median: 2
## Findings: Multidisciplinary reuse

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Dataset Name</th>
<th>Total Degree</th>
<th>Multidisciplinarity Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3202</td>
<td>Panel Study of Income Dynamics, 1968-1999: Supplemental Files</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>D2521</td>
<td>Monitoring the Future: A Continuing Study of American Youth (8th- and 10th-Grade Surveys), 1991</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>D4075</td>
<td>Early Childhood Longitudinal Study [United States]: Kindergarten Class of 1998-1999, Third Grade</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>D8506</td>
<td>National Youth Survey [United States]: Wave III, 1978</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>D3672</td>
<td>Treatment Episode Data Set -- Admissions (TEDS-A), 2000</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>D3884</td>
<td>Treatment Episode Data Set -- Admissions (TEDS-A), 2001</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>D3894</td>
<td>Survey of Income and Program Participation (SIPP) 2001 Panel</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>
Findings: Integrative reuse of datasets from different disciplines

- 657 datasets have been co-cited with at least one other dataset
- On average, 43% of a dataset’s edges are interdisciplinary

These datasets are cited together in the same publication but are associated with different disciplines.
Findings: Distance

- Average distance between disciplines involved in interdisciplinary reuse is 0.07

Diagram:

- Different disciplines, both ‘core’: distance = 0
- Different disciplines, one ‘core’, one external: distance = 1
Discussion

• High level of multidisciplinary reuse
  – 42.5% of datasets receive citations from two or more disciplines

• Integrative reuse is also common (but probably not as common as it looks here)
Discussion

• But reuse tends to be local, especially for integrative reuse
  – 59% of multidisciplinary citations are to core disciplines
  – 93% of integrative citations are to core disciplines
Wrap-up

• Multidisciplinary data reuse is an established practice
  – Datasets with many multidisciplinary citations may be exemplars for curating data for reuse in multiple communities

• Integrative reuse may still be a challenge

• Good data citation practices enable new (and maybe interesting) research
Future work

• Solving problems: separating reuse from courtesy/review citations; better assignment of discipline; strengthen metrics

• What are the primary characteristics of datasets associated with high multidisciplinarity scores? What makes a dataset more likely to be integrated with another?

• How does the network evolve over time?
Thank you!

• Photo credits:
  – orange/apple hybrid: http://www.flickr.com/photos/horas18/3957537071/

• Data provided by Elizabeth Moss and ICPSR.

• Thanks to Eytan Adar and Jessica Hullman for their input on this project.