Influence of OA, Gender, Co-authorship on Citation

Science & Technology Indicators 2017 (Paris)

September 6–8, 2017
Obligatory warm-up slide

- Order of operations for today:
  - Context
  - Methodology
  - Analysis
  - Some notes on ongoing follow-up work
  - Discussion
  - Conclusions, wild speculation

- Your divided attention:
Study context

- Citations used to evaluate impact of research (impact w/in academic community)
- Citation rates normalized by year, subfield, doc type to provide a level playing field
- Other parameters affect citation rates as well:
  - International collaboration
  - OA citation advantage (maybe?)
  - Gender
- These parameters inter-related as well:
  - International collabs more likely to be available in OA.
  - Women participate less often in international collab than men do.
- This project aims to disentangle this knot.
Study methodology

- Data sources:
  - Scopus
  - 1science
  - NamSor API

- Scoping out a sample:
  - NamSor very reliable for US context
  - 1science database also covers US very well (~95% recall)
  - Citation windows + OA backfilling effect exclude most recent years
  - Not all subfields have similar gender composition, so gender dynamics might vary between them
Study methodology

- Sample selected:
  - US publications (including their co-authors abroad)
  - Only pubs where all authors tagged by NamSor w/100% confidence
  - Publication year 2010 (follow-up work covers 2010–2012)
  - Subfields: Developmental Bio, Cardio. System & Hematology (follow-up work covers all subfields, except under Arts & Hum.)

- Filtering and bias:
  - Each of the two subfields represented ~2% of initial article pop. (US).
  - After filtering for NamSor, 1science coverage: 1.9% of population.
  - Dev. Bio sample: 3 000 papers, 32% of initial pub count
  - Cardio. System & Hema. sample: 3 500 papers, 36% of initial count
Study methodology

- Coding variables, for each paper:
  - Citations: log (cite + 1), *pace* Thelwall & Wilson 2014
  - OA status: binary (no gold/green)
  - Gender:
    - Any women involved in team: binary
    - Share of women in team: scalar [0–1]
    - Lead author is female: binary
  - Collaboration
    - Number of authors: scalar [1–∞[ (actual max ≈ 20)
    - Number of institutions: scalar [1–∞[ (actual max ≈ 20)
    - Number of countries: scalar [1–∞[ (actual max ≈ 10)
    - International collaboration: binary
Analysis—Developmental biology

- Inter-correlations between variables ($R^2$):

```
<table>
<thead>
<tr>
<th>DEV_BIO</th>
<th>OA</th>
<th>Female$</th>
<th>Prop_F</th>
<th>Lead_F$</th>
<th>n_authors</th>
<th>n_addresses</th>
<th>n_countries</th>
<th>Int$</th>
<th>nb_cit</th>
<th>log(cite)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA</td>
<td>100%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>4%</td>
<td>3%</td>
<td>1%</td>
<td>1%</td>
<td>0%</td>
<td>10%</td>
</tr>
<tr>
<td>Female$</td>
<td>2%</td>
<td>100%</td>
<td>67%</td>
<td>25%</td>
<td>16%</td>
<td>5%</td>
<td>2%</td>
<td>2%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td>Prop_F</td>
<td>0%</td>
<td>67%</td>
<td>100%</td>
<td>47%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Lead_F$</td>
<td>0%</td>
<td>25%</td>
<td>47%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>n_authors</td>
<td>4%</td>
<td>16%</td>
<td>2%</td>
<td>0%</td>
<td>100%</td>
<td>43%</td>
<td>23%</td>
<td>17%</td>
<td>0%</td>
<td>7%</td>
</tr>
<tr>
<td>n_addresses</td>
<td>3%</td>
<td>5%</td>
<td>0%</td>
<td>0%</td>
<td>43%</td>
<td>100%</td>
<td>32%</td>
<td>25%</td>
<td>1%</td>
<td>6%</td>
</tr>
<tr>
<td>n_countries</td>
<td>1%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>23%</td>
<td>32%</td>
<td>100%</td>
<td>80%</td>
<td>0%</td>
<td>3%</td>
</tr>
<tr>
<td>Int$</td>
<td>1%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>17%</td>
<td>25%</td>
<td>80%</td>
<td>100%</td>
<td>0%</td>
<td>3%</td>
</tr>
<tr>
<td>nb_cit</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>30%</td>
</tr>
<tr>
<td>log(cite)</td>
<td>10%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>7%</td>
<td>6%</td>
<td>3%</td>
<td>3%</td>
<td>30%</td>
<td>100%</td>
</tr>
</tbody>
</table>
```
### Coefficient tables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Coefficient</th>
<th>Stat. significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Access (OA) status</td>
<td>Binary</td>
<td>0.341</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Women involved in authorship</td>
<td>Binary</td>
<td>0.158</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Woman as corresponding author</td>
<td>Binary</td>
<td><strong>not significant</strong></td>
<td></td>
</tr>
<tr>
<td>Proportion of women in research team</td>
<td>Scalar</td>
<td>-0.249</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Number of authors</td>
<td>Scalar</td>
<td>0.034</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Number of institutions</td>
<td>Scalar</td>
<td>0.040</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Number of countries</td>
<td>Scalar</td>
<td><strong>not significant</strong></td>
<td></td>
</tr>
<tr>
<td>International co-authorship</td>
<td>Binary</td>
<td><strong>not significant</strong></td>
<td></td>
</tr>
<tr>
<td>Model constant</td>
<td>n/a</td>
<td>0.749</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Model overall</td>
<td>n/a</td>
<td>n/a</td>
<td>p&lt;0.001</td>
</tr>
</tbody>
</table>
Analysis—Developmental biology

- Predicted citation scores (5 authors, 2 institutions):
## Inter-correlations between variables ($R^2$):

<table>
<thead>
<tr>
<th>CARDIO</th>
<th>OA</th>
<th>Female$</th>
<th>Prop_F</th>
<th>Lead_F$</th>
<th>n_authors</th>
<th>n_addresses</th>
<th>n_countries</th>
<th>Internat$</th>
<th>nb_cit</th>
<th>log(cite)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA</td>
<td>100%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>3%</td>
<td>2%</td>
<td>1%</td>
<td>1%</td>
<td>5%</td>
<td>11%</td>
</tr>
<tr>
<td>Female$</td>
<td>2%</td>
<td>100%</td>
<td>62%</td>
<td>20%</td>
<td>18%</td>
<td>8%</td>
<td>2%</td>
<td>3%</td>
<td>1%</td>
<td>3%</td>
</tr>
<tr>
<td>Prop_F</td>
<td>0%</td>
<td>62%</td>
<td>100%</td>
<td>40%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td>Lead_F$</td>
<td>0%</td>
<td>20%</td>
<td>40%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>n_authors</td>
<td>3%</td>
<td>18%</td>
<td>1%</td>
<td>0%</td>
<td>100%</td>
<td>47%</td>
<td>18%</td>
<td>13%</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>n_addresses</td>
<td>2%</td>
<td>8%</td>
<td>0%</td>
<td>0%</td>
<td>47%</td>
<td>100%</td>
<td>28%</td>
<td>18%</td>
<td>5%</td>
<td>7%</td>
</tr>
<tr>
<td>n_countries</td>
<td>1%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>18%</td>
<td>28%</td>
<td>100%</td>
<td>71%</td>
<td>3%</td>
<td>4%</td>
</tr>
<tr>
<td>Internat$</td>
<td>1%</td>
<td>3%</td>
<td>0%</td>
<td>0%</td>
<td>13%</td>
<td>18%</td>
<td>71%</td>
<td>100%</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>nb_cit</td>
<td>5%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
<td>5%</td>
<td>3%</td>
<td>2%</td>
<td>100%</td>
<td>53%</td>
</tr>
<tr>
<td>log(cite)</td>
<td>11%</td>
<td>3%</td>
<td>1%</td>
<td>0%</td>
<td>10%</td>
<td>7%</td>
<td>4%</td>
<td>3%</td>
<td>53%</td>
<td>100%</td>
</tr>
</tbody>
</table>
### Coefficient tables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Coefficient</th>
<th>Stat. significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Access (OA) status</td>
<td>Binary</td>
<td>0.317</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Women involved in authorship</td>
<td>Binary</td>
<td><strong>not significant</strong></td>
<td></td>
</tr>
<tr>
<td>Woman as corresponding author</td>
<td>Binary</td>
<td><strong>not significant</strong></td>
<td></td>
</tr>
<tr>
<td>Proportion of women in research team</td>
<td>Scalar</td>
<td>0.069</td>
<td>p=0.033</td>
</tr>
<tr>
<td>Number of authors</td>
<td>Scalar</td>
<td>0.039</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Number of institutions</td>
<td>Scalar</td>
<td>0.022</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Number of countries</td>
<td>Scalar</td>
<td><strong>not significant</strong></td>
<td></td>
</tr>
<tr>
<td>International co-authorship</td>
<td>Binary</td>
<td>0.084</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Model constant</td>
<td>n/a</td>
<td>0.576</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Model overall</td>
<td>n/a</td>
<td>n/a</td>
<td>p&lt;0.001</td>
</tr>
</tbody>
</table>
Predicted citation scores (5 authors, 2 institutions):
Looking at different modeling approaches:

- Robust modeling: better suited to input and output variables that are non-normally distributed, which we know to be the case
- Binning: looking at 5 bins for gender composition of research teams
- Citation data as a category variable: trying to address challenges posed by non-normal distribution

Applied models to all subfields, for 2010–2012, still US pubs

Prepping data to include number of pubs from researchers & institutions involved, to figure out how much of a role these play in determining citation outcomes.
Discussion

- OA associated with higher citation scores—though reasons still not clear!
  - Selection bias?
  - Early availability bias?
  - Prestigious institutions/researchers having funds for APCs?
- International collaboration promotes higher citation scores; best parametrised as binary variable, not scalar.
- Larger number of authors & institutions promotes citation.
- Mixed-gender teams—leaning male—seem to be optimal for promoting citations.
- Gender of lead author does not seem to have an effect.
Discussion

- OA and international collab advantages robust across areas of research, and models. OA > international collab.
- Gender dynamics quite even across domains using approach presented today; but across subfields and across models their impact is much less consistent.
- All the models have strong statistical significance (p<0.001), but low fit ($R^2 \approx 0.15$): good predictor of aggregate results.
- Potential concern: with a low $R^2$, models might be picking up on different underlying patterns.
Conclusions, wild speculation

- Citation scores partially determined by OA and international collaboration, and (seemingly) gender balance of research teams.
- Influence of each is independent of the others.
- Each can be considered a strategy for increasing citation.
- If citation scores supposed to measure quality or “excellence” of research content—distinct from visibility or uptake—should we be normalising for these strategies when assessing excellence?
- Would be interesting to inspect effect of these strategies on quality and dissemination, independent of each other.
- Ultimately, is citation-based evaluation primarily about quality or uptake? This should guide testing, interpretation.
Shameless pitch

- If you liked the presentation, consider following on Twitter.

  @DBS truck  @ScienceMetrix

- We also blog about bibliometrics, data mining and science policy at ScienceMetrics.org: check it out, sign up!
CONTACT

Brooke Struck
brooke.struck@science-metrix.com

Guillaume Roberge
guillaume.roberge@science-metrix.com

Matthew Durning
matt.durning@science-metrix.com

David Campbell
david.campbell@science-metrix.com

Thank you!

Science-Metrix
Montréal – Ottawa – Washington

WEBSITE
www.science-metrix.com

PHONE
1.514.495.6505
1.800.994.4761