Automated arXiv feeds on Twitter: On the role of bots in scholarly communication

Stefanie Haustein, Kim Holmberg, Timothy D. Bowman, Andrew Tsou, Cassidy R. Sugimoto & Vincent Larivière

@stefhaustein
Introduction

- increase of Twitter use
  - 230 million active users, 500 million tweets per day
  - 39% increase of users from 09/2012 to 09/2013
  - 16% of US, 3% of world population in 2013
  - 19% of US internet users 01/2014
- uptake by researchers
  - 1 in 40 university faculty member in US and UK have Twitter account (Priem & Costello, 2010)
  - 9% of researchers use Twitter for work (Rowlands et al., 2011)
  - 15% of German university faculty members, 70% of which at least occasionally in professional context (Pscheida et al., 2013)

1 Twitter statistics calculated based on data from: http://www.sec.gov/Archives/edgar/data/1418091/000119312513400028/d564001ds1a.htm and http://www.census.gov/population/international/data
2 Pew Research Center’s Internet Project surveys, 2010-2014 http://www.pewresearch.org/fact-tank/2014/06/11/can-twitter-survive-in-a-facebook-world-the-key-is-being-different/
Introduction

- increasing presence of tweets as impact metrics
Introduction

- 5% to 10% monthly growth of social media activity related to scientific articles (Adie & Roe, 2013)
- scholarly documents on Twitter
  - 1.6% of WoS papers with DOIs 2005-2011 (Zahedi, Costas & Wouters, 2014)
  - 13.3% of WoS papers with DOIs 07-12/2011 (Costas, Zahedi & Wouters, 2014)
  - 20.4% of PubMed/WoS 2012 (Haustein et al., 2014b)
  - 21.5% of WoS papers with DOIs 2012 (Costas, Haustein & Larivière, in prep.)
- high Twitter coverage (44.9%) for set of arXiv papers (Haustein et al., 2014a)
  - high presence of automated Twitter accounts!
Introduction: tweets to arXiv eprints and published papers

- **DOI:** 27.8% arXiv id: 100%
- **DOI:** 96.3% arXiv id: 100%

- **exact and fuzzy**
  - DOI
  - titles
  - author names
  - abstracts
  - title length

**arXiv primary (sub)category**

<table>
<thead>
<tr>
<th>Category</th>
<th>Papers Matched</th>
<th>Indirect match tweeted (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science</td>
<td>2,055</td>
<td>44.9%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>7,723</td>
<td>29.2%</td>
</tr>
<tr>
<td>Physics</td>
<td>33,782</td>
<td>48.5%</td>
</tr>
<tr>
<td>Astrophysics</td>
<td>10,187</td>
<td>53.9%</td>
</tr>
<tr>
<td>Condensed Matter</td>
<td>10,964</td>
<td>32.5%</td>
</tr>
<tr>
<td>GR &amp; QC</td>
<td>1,439</td>
<td>63.1%</td>
</tr>
<tr>
<td>HEP Experiment</td>
<td>1,573</td>
<td>81.1%</td>
</tr>
<tr>
<td>HEP Lattice</td>
<td>451</td>
<td>69.2%</td>
</tr>
<tr>
<td>HEP Phenomenology</td>
<td>2,392</td>
<td>69.6%</td>
</tr>
<tr>
<td>HEP Theory</td>
<td>1,225</td>
<td>72.7%</td>
</tr>
<tr>
<td>Nonlinear Sciences</td>
<td>543</td>
<td>40.3%</td>
</tr>
<tr>
<td>Nuclear Experiment</td>
<td>432</td>
<td>47.7%</td>
</tr>
<tr>
<td>Nuclear Theory</td>
<td>384</td>
<td>28.6%</td>
</tr>
<tr>
<td>Physics (subcategory)</td>
<td>2,689</td>
<td>40.4%</td>
</tr>
<tr>
<td>Quantum Physics</td>
<td>1,503</td>
<td>44.4%</td>
</tr>
<tr>
<td>Quantitative Biology</td>
<td>289</td>
<td>64.7%</td>
</tr>
<tr>
<td>Quantitative Finance</td>
<td>51</td>
<td>54.9%</td>
</tr>
<tr>
<td>Statistics</td>
<td>263</td>
<td>23.2%</td>
</tr>
</tbody>
</table>

**arXiv-WoS subset** 44,163 22,722 44.9%
Research questions

• How can automated accounts be defined and identified?

• How much tweets to scientific papers are produced by automated Twitter accounts?
  1) systematic search for automated arXiv Twitter accounts
  2) coding of a representative set of Twitter accounts which have tweeted at least one arXiv preprint or published paper

➢ Do automated accounts affect the validity of tweets as impact measures?
Systematic search for arXiv Twitter accounts

Data & Methods

- Twitter online search for “arXiv” in Twitter handle, display name, or account description (05/2014)
- Manual coding of 90 accounts by two researchers:
  - Platform feed: automated feed of papers from arXiv section or subsection; platform-based feeds tweeting everything published in an arXiv subject area, triggered by arXiv RSS feed
  - Topic feed: automated feed of papers relevant to a certain topic; keyword-based feeds, triggered by keyword specific searches
  - Selective/qualitative: some sort of qualitative selection; human selection of “interesting” papers
  - Not related to arXiv
Systematic search for arXiv Twitter accounts

Data & Methods

- collection of Twitter statistics
  - number of tweets
  - date of first tweet
  - mean number of tweets per day
  - number of followers
  - number of following
  - BotOrNot score
### Systematic search for arXiv Twitter accounts

#### Results

<table>
<thead>
<tr>
<th>account type</th>
<th>number (%) of accounts</th>
<th>tweets</th>
<th>mean followers</th>
<th>mean following</th>
<th>% of 50,068 tweets</th>
<th>mean Truthy BotOrNot score</th>
</tr>
</thead>
<tbody>
<tr>
<td>platform feed (bot)</td>
<td>43 (84.3%)</td>
<td>87,389</td>
<td>34.9</td>
<td>0.6</td>
<td>8.8%</td>
<td>33%</td>
</tr>
<tr>
<td>topic feed (bot)</td>
<td>4 (7.8%)</td>
<td>10,040</td>
<td>527.0</td>
<td>491.5</td>
<td>0.1%</td>
<td>40%</td>
</tr>
<tr>
<td>selective/qualitative</td>
<td>4 (7.8%)</td>
<td>3,081</td>
<td>361.8</td>
<td>50.5</td>
<td>1.0%</td>
<td>46%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>51 (100%)</strong></td>
<td><strong>100,510</strong></td>
<td><strong>99.1</strong></td>
<td><strong>43.0</strong></td>
<td><strong>9.9%</strong></td>
<td><strong>33%</strong></td>
</tr>
</tbody>
</table>

![Graph showing the number of followers over time for different account types]
Systematic search for arXiv Twitter accounts

Results

- platform and topic feed bots do not behave like regular Twitter bots

    ![Graph showing distribution of BotOrNot scores vs. number of tweets]

    - **BotOrNot score**
    - **Number of tweets**
    - Platform feed
    - Topic feed
    - Selective/qualitative
    - Number of followers

- more automated accounts possible:
  - journals
  - publishers
  - societies / associations
  - institutions
  - authors
Coding of Twitter accounts

Data & Methods

- identifying and quantifying different kinds of user groups among Twitter users tweeting links to scientific papers

- 19,804 WoS papers with arXiv eprint submitted in 2012, tweeted at least once
- 50,068 tweets
- 10,384 unique Twitter accounts
  - coding of random sample of 1,000 accounts by three researchers
  - 100 accounts to test inter-rater reliability
## Coding of Twitter accounts

### Data & Methods

<table>
<thead>
<tr>
<th><strong>TWITTER PROFILE INFO</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td><strong>Screen Name</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>

### CODEBOOK

Is the account (based on its description) maintained by a person or organization?

- [ ] Person
- [x] Organization
- [ ] Unknown
Coding of Twitter accounts

Data & Methods

What type of organization?
- Corporation
- Government
- Journal
- University
- Library
- Non-profit
- Publisher
- Research Center
- Scientific Society/Assoc.
- Other

Who is behind tweet content?
- Human
- Bot
- Cyborg
- Unknown
- Private
- Does Not Exist

<table>
<thead>
<tr>
<th>Tweets</th>
<th>Tweets &amp; replies</th>
</tr>
</thead>
</table>
| AIP Publishing @AIP_Publishing - 1h
Editorial: Building on excellence: A vision for the future ow.ly/BN61N #AIP_APL |
| AIP Publishing @AIP_Publishing - 7h
Job Posting: Nuclear and Analytical Science Division Leader | Lawrence Livermore National Laboratory bit.ly/1v2OUJs #PTJOBS |
| AIP Publishing @AIP_Publishing - 15h
Improved self-consistent and resolution-of-identity approximated Becke'05 density functional model of...ow.ly/BLGga #AIP_JCP |
| AIP Publishing @AIP_Publishing - 18h
Engineering of optical polarization based on electronic band structures ofA-plane ZnO layers under biaxial...ow.ly/BLFncN #AIP_JAP |
| AIP Publishing @AIP_Publishing - 20h
Targeting Tumors With Electricity @insidescience ow.ly/BLDSQ |
| AIP Publishing @AIP_Publishing - 22h
PlasmaTalks: Ron Davidson interviews Cédric Villani ow.ly/BLD7P #AIP_POP |
| AIP Publishing @AIP_Publishing - 24h
Congress Passes Short Term Funding Bill; NASA's Near-Earth Objects Program Extends |

---
Data & Methods

Coding of Twitter accounts

The tweets are mostly RETWEETS:  □ Yes

The tweets contain NON-ENGLISH text:  □ Yes

Is the type of automation clarified in Twitter account profile?

□ Yes

□ No

What is the automated content?

□ Platform Based (arXiv, PubMed, SSRN, etc.)

□ Journal Based (papers of one journal)

□ Publisher Based (papers of one publisher)

□ Institution/Association Based (relevant to one society/assoc.)

□ Topic Based (research field, keyword, discipline, etc.)

□ Other
How many tweets seem to be automated?

- Complete (all or majority)
- Selective (only a fraction)
- Unknown

Random Notes: tweets are paper title plus hashtag for particular AIP journal, e.g. #AIP_JCP for Journal of Chemical Physics, #AIP_PT for Physics Today
Outlook

• Who is posting links to scientific papers on Twitter and what is their motivation?
• How many tweets are generated by humans, cyborgs and bots?
• How many tweets show actual engagement, how much distribution only?
• Do user groups and tweets differ between arXiv eprint and journal paper?

➢ What do tweets to scientific papers indicate?
   ➢ impact or distribution?
   ➢ among the scientific community or the general public?
Outlook

- distinguishing type of tweet based on content e.g., similarity with article title (%)

**low engagement**

- percentage (%) represents similarity between paper title and tweet text

**high engagement**

- Burkert & Hartmann on star formation thresholds, should be good. arxiv.org/abs/1212.4543 with nod to newly be-doctored @xxxxxxxx too!

- Richard Ellis about possible detection of z=11.9 galaxy in Hubble data: “While definitively real, we remain cautious of its nature” AAS221

- Any links to learn more about the galaxy?

- @xxxxxxx Read Ellis’ paper from 2012 arxiv.org/abs/1211.6804, and the recently submitted one by Brammer et al: arxiv.org/abs/1301.0317 #AAS221
Thanks to Euan Adie and Altmetric for access to their Twitter data!

Please submit to *Aslib Journal of Information Management*
Special Issue *Social Media Metrics in Scholarly Communication: exploring tweets, blogs, likes and other altmetrics*
Submission Deadline 15.11.2014
http://www.emeraldgrouppublishing.com/products/journals/call_for_papers.htm?id=5754

Thank you for your attention!
Questions?

Stefanie Haustein
stefanie.haustein@umontreal.ca
@stefhaustein

Canada Research Chair on the Transformations of Scholarly Communication
École de bibliothéconomie et des sciences de l'information

Alfred P. Sloan Foundation
References


