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

Université de Montréal

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

Introduction

- increase of Twitter use
  - 230 million active users, 500 million tweets per day
  - 39% increase of users from 09/2012 to 09/2013[1]
  - 16% of US, 3% of world population in 2013[1]
  - 19% of US internet users 01/2014[2]

- 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 <u>automated Twitter accounts</u>!
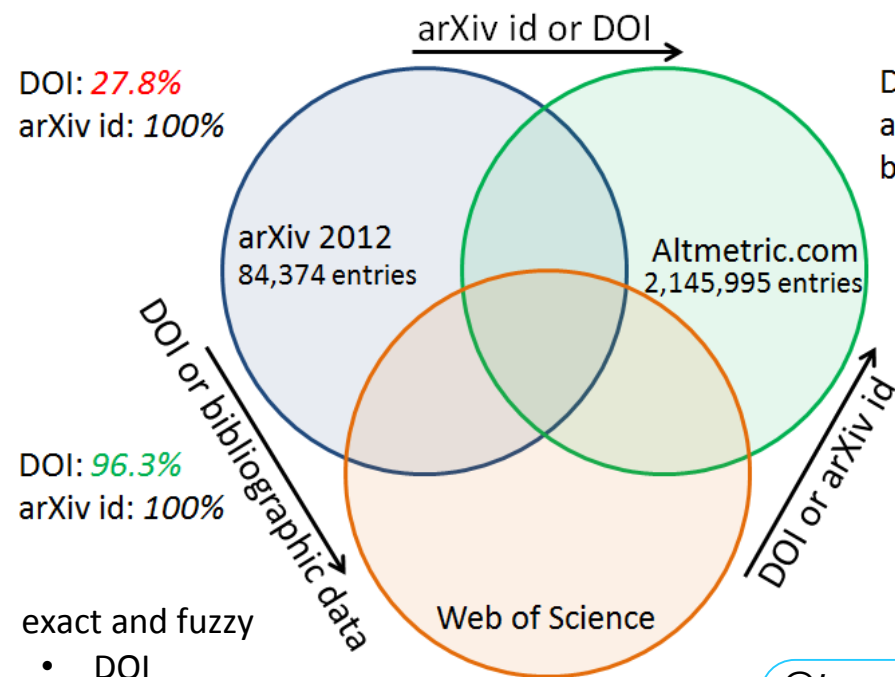
⬤ arXiv hep-ex @hep_ex · 14h
[1406.5171] CMS Collaboration : Search for
excited quarks in the photon+jet final state
in proton-proton collisi...
arxiv.org/abs/1406.5171

High Energy Physics – Experiment

**Search for excited quarks in the
photon+jet final state in proton–
proton collisions at sqrt(s) = 8 TeV**

CMS Collaboration

arXiv id or DOI

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

DOI: *85.9%*
arXiv id: *8.5%*
both: *1.7%*

arXiv 2012
84,374 entries

Altmetric.com
2,145,995 entries

DOI or bibliographic data

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

DOI or arXiv id

Web of Science

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

@hep_ex
@hep_lat
@hep_ph
@hep_th

arXiv hep-ex
@hep_ex

arXiv new submissions on hep-ex.
Feedbacks to @Misho are welcome.
[README] bit.ly/arxiv_twitter [See also]
@hep_lat @hep_ph @hep_th

🔗 en.misho-web.com/phys/hep_tools...
🕐 Joined June 2012

arXiv primary (sub)category	Papers	Indirect match	
		Matched	**Tweeted (%)**
Computer Science	2,055	1,018	44.9%
Mathematics	7,723	2,711	29.2%
Physics	33,782	18,614	48.5%
Astrophysics	10,187	5,896	53.9%
Condensed Matter	10,964	4,819	32.5%
GR & QC	1,439	948	63.1%
HEP Experiment	1,573	1,320	81.1%
HEP Lattice	451	343	69.2%
HEP Phenomenology	2,392	1,728	69.6%
HEP Theory	1,225	914	72.7%
Nonlinear Sciences	543	284	40.3%
Nuclear Experiment	432	243	47.7%
Nuclear Theory	384	122	28.6%
Physics (subcategory)	2,689	1,255	40.4%
Quantum Physics	1,503	742	44.4%
Quantitative Biology	289	267	64.7%
Quantitative Finance	51	28	54.9%
Statistics	263	84	23.2%
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?

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

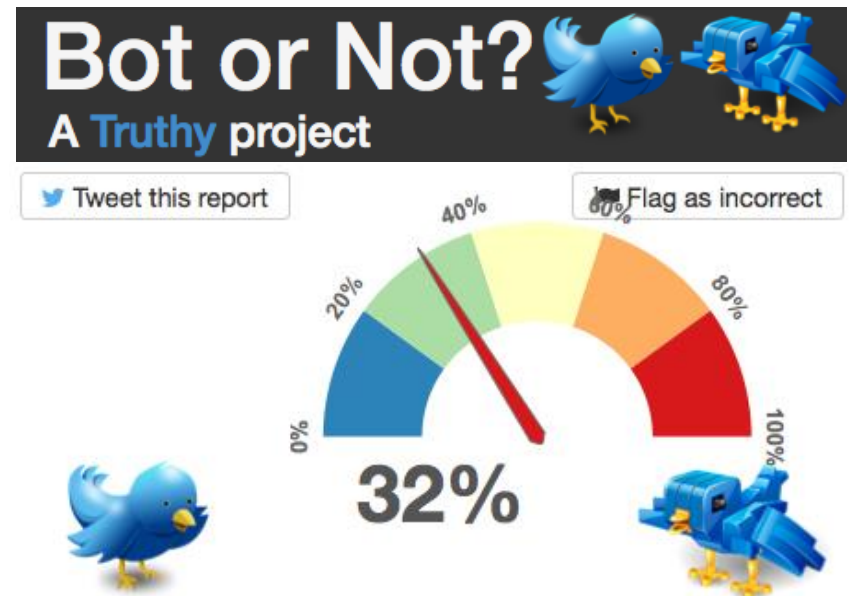**topic feed**: automated feed of papers relevant to a certain topic; keyword-based feeds, triggered by keyword specific searches

selective/qualitative: some sort of qualitative selection; human selection of "interesting" papers
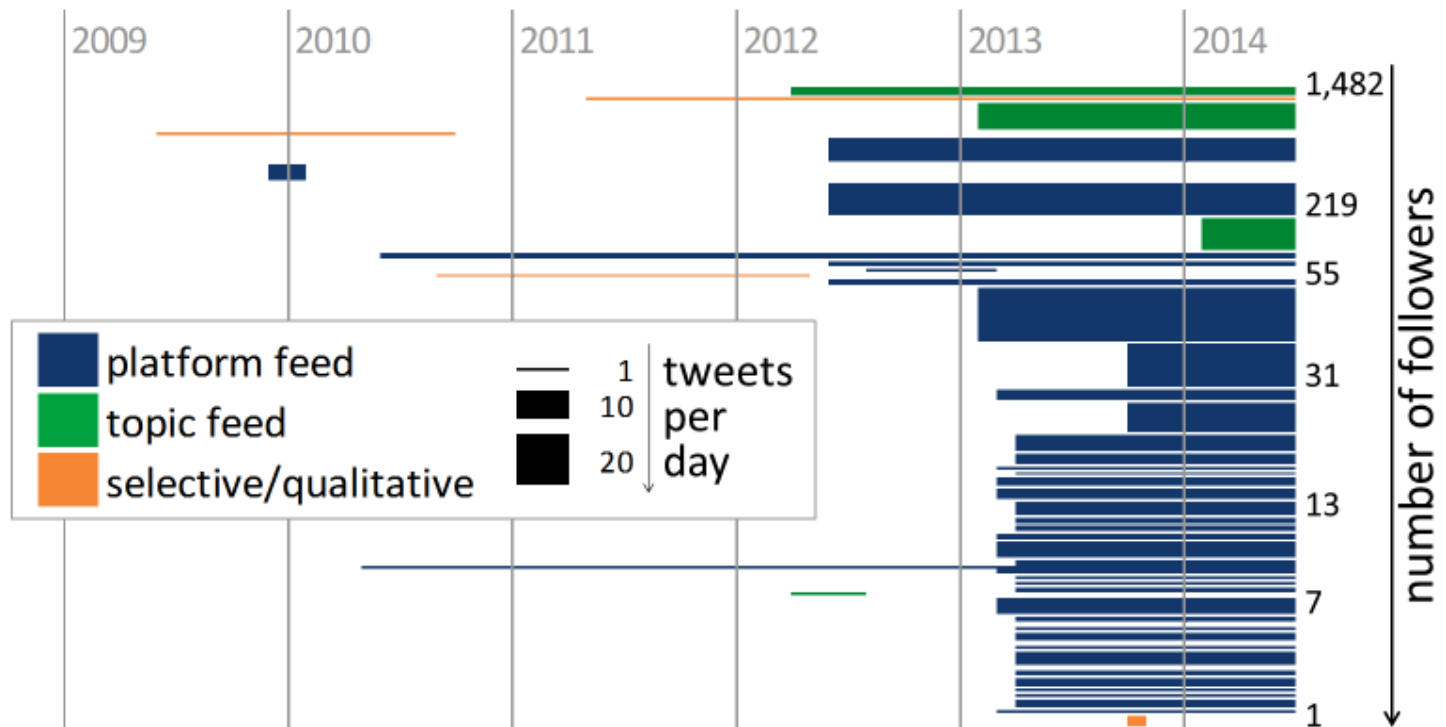
not related to arXiv

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

Results

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

Results

- platform and topic feed bots do not behave like regular Twitter bots
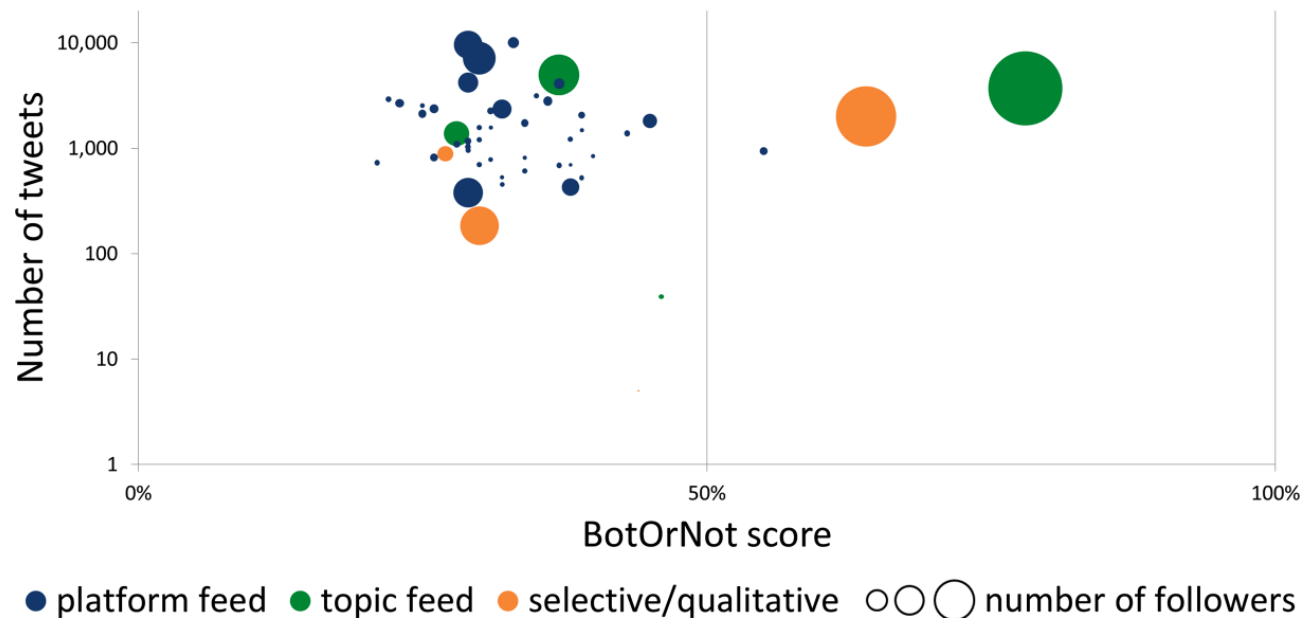
- more automated accounts possible:
  - journals
  - publishers
  - societies / associations
  - institutions
  - authors

Data & Methods

➢ identifying and quantifying different kinds of user groups among Twitter users tweeting links to scientific papers
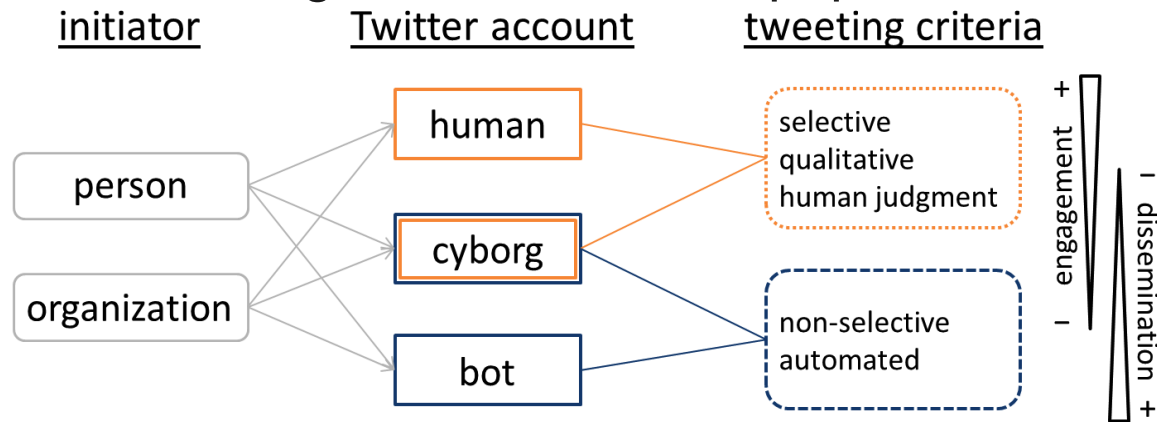
- 19,804 WoS papers with arXiv eprint submitted in 2012, tweeted at least once

- 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

TWITTER PROFILE INFO

Name	AIP Publishing
Screen Name	AIP_Publishing
URL	http://journals.aip.org
Description	Highly cited, highly respected, and highly relevant, AIP's archival journals are the cornerstone of many physics collections.

AIP

CODEBOOK

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

- ○ Person
- ◉ Organization
- ○ Unknown

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

Tweets   Tweets & replies

AIP **AIP Publishing** @AIP_Publishing · 1h
Editorial: Building on excellence: A vision for the future ow.ly/BN61N #AIP_APL
↩   ⟲1   ★   ···

AIP **AIP Publishing** @AIP_Publishing · 7h
Job Posting: Nuclear and Analytical Science Division Leader | Lawrence Livermore National Laboratory bit.ly/1v2OUJs #PTJOBS
↩   ⟲   ★   ···

AIP **AIP Publishing** @AIP_Publishing · 15h
Improved self-consistent and resolution-of-identity approximated Becke'05 density functional model of…ow.ly/BLGga #AIP_JCP
↩   ⟲1   ★   ···

AIP **AIP Publishing** @AIP_Publishing · 18h
Engineering of optical polarization based on electronic band structures ofA-plane ZnO layers under biaxial…ow.ly/BLFcN #AIP_JAP
↩   ⟲   ★   ···

AIP **AIP Publishing** @AIP_Publishing · 20h
Targeting Tumors With Electricity @insidescience ow.ly/BLDSQ
↩   ⟲   ★   ···

AIP **AIP Publishing** @AIP_Publishing · 22h
PlasmaTalks: Ron Davidson interviews Cédric Villani ow.ly/BLD7P #AIP_POP
↩   ⟲1   ★1   ···

AIP **AIP Publishing** @AIP_Publishing · 24h
Congress Passes Short Term Funding Bill; NASA's Near-Earth Objects

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   high engagement

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

100% A unification of RDE model and XCDM model
arxiv.org/abs/1212.5790

36% Burkert & Hartmann on star formation thresholds, should be good.
arxiv.org/abs/1212.4543 with nod to newly be-doctored @███████ too!

100% On the calculation of percentile-based bibliometric indicators ow.ly/h8Kzp
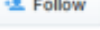
7 Jan 2013
Richard Ellis about possible detection of z=11.9 galaxy in Hubble data: "While definitively real, we remain cautious of it's nature" #AAS221
Details

7 Jan 2013
@███ Any links to learn more about the galaxy?
Details

87% Hysteretic response characteristics and dynamic phase transition via site dilution in the kinetic...(arXiv:1206.5425)
arxiv.org/abs/1206.5425

16% .@████ 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

Université de Montréal

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

ALFRED P. SLOAN FOUNDATION

References

Adie, E. & Roe, W. (2013). Altmetric: Enriching Scholarly Content with Article-level Discussion and Metrics. *Learned Publishing*, 26(1), 11-17.

Costas, R., Zahedi, Z. & Wouters, P. (2014). Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Sciences and Technology*. arxiv: 1401.4321

Haustein, S. Bowman, T.D., Macaluso, B., Sugimoto, C.R. & Larivière, V. (2014a): Measuring Twitter activity of arXiv e-prints and published papers. altmetrics14 http://dx.doi.org/10.6084/m9.figshare.1041514

Haustein, S., Peters, I., Sugimoto, C.R., Thelwall, M., & Larivière, V. (2014b). Tweeting Biomedicine: An Analysis of Tweets and Citations in the Biomedical Literature. *Journal of the Association for Information Sciences and Technology*, 65(4), 656-669. doi: 10.1002/asi.23101

Priem, J., & Costello, K. L. (2010). How and why scholars cite on Twitter. *Proceedings of the 73th Annual Meeting of the American Society for Information Science and Technology*, Pittsburgh, USA.

Pscheida, D., Albrecht, S., Herbst, Minet, C. & Köhler, T. (2013). Nutzung von Social Media und onlinebasierten Anwendungen in der Wissenschaft. Erste Ergebnisse des Science 2.0-Survey 2013 des Leibniz-Forschungsverbunds „Science 2.0" available from: http://www.qucosa.de/fileadmin/data/qucosa/documents/13296/Science20_Datenreport_2013_PDF_A.pdf

Rowlands, I., Nicholas, D., Russell, B., Canty, N., & Watkinson, A. (2011). Social media use in the research workflow. *Learned Publishing*, 24, 183–195.

Zahedi, Z., Costas, R. & Wouters, P. (2014). How well developed are altmetrics? cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*. doi: 10.1007/s11192-014-1264-0