Has Much Potential but Biased: Exploring the Scholarly Landscape in Twitter

Haewoon Kwak
Telefonica Research Lab
Barcelona, Spain
kwak@tid.es

Jong Gun Lee
Pulse Lab Jakarta - UN Global Pulse
Jakarta, Indonesia
jonggun.lee@un.or.id

ABSTRACT

We explore how research papers are shared in Twitter to understand its potential and limitation of the current practice that measures or predicts the scientific impact of research papers from the web. We track 54 second-level domains offering the top 100 journals listed in Google Scholar and collect 403,165 tweets sharing 75,677 unique research papers by 142,743 users over the course of 135 days. Our findings show the great potential of Twitter as a platform for paper sharing, but at the same time, indicate the limitations of measuring scientific impact through the lens of social media mainly due to the highly skewed and limited attention to few number of top journals.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Twitter, Altmetric, Google Scholar, research paper sharing

1. INTRODUCTION

Along with the growing popularity of social media, measuring or predicting the scientific impact of research papers from the web have received much attention [1, 2, 3]. For instance, PLOS ONE shows the volume of social responses, such as links from Twitter or Facebook, in their web pages, and Altmetrics[1], an independent service to measure the paper-level online impacts, has been building reputation through the collaboration with famous scholarly journals. Although those efforts are supported by recent findings that online social responses to research papers and their future citation counts are positively correlated in PLOS ONE [2] and arXiv [4], they are not fully validated across disciplines.

In this work we explore paper sharing dynamics in Twitter to answer i) how actively social media reacts to research papers; and ii) how well online social responses reflect the established reputation in research communities. We collect 403,165 tweets over the course of 135 days by tracking 54 second-level domain names offering the top 100 journals across discipline listed in Google Scholar as of 20th January 2013. We crawl social graphs, lists of followers and followees, of those who share at least one paper for measuring the reach of each research paper.

We surprisingly discover that the huge scholarly landscape already is represented on Twitter; 98,308,648 unique users are *exposed* to at least one paper published in any of 54 scholarly domains and 29,946,873 unique users to those published in the top 100 journals. The number of users exposed to papers is two orders of magnitude greater than those who tweet about said papers. We find that papers are more actively shared via grassroot efforts than official journal accounts. However, we also reveal the limitations of Twitter in sharing research papers due to its huge bias towards a few top journals. Furthermore, we find that higher ranking journals do not guarantee a higher volume of social responses in general. Our findings show both the potential and the limitation of current practice that measures or predicts the scientific impact of research papers across disciplines and journals through the lens of social media. Much care must be taken to apply it to the entire scholarly landscape.

2. DATA COLLECTION

We use the Twitter Streaming API to obtain longitudinal traces of tweets sharing research papers. Although it offers only 1% samples of all public tweets, and sometimes fails to represent the overall activities in Twitter, it is currently the most efficient way of monitoring tweets.

As tracking all papers shared via Twitter is infeasible, we focus only on those from high quality journals. We do this by finding the second level domain names of the top 100 journals according to Google scholar as of the 20th January 2013. For example, the second-level domain of ACM portal, `portal.acm.org`, is `acm.org`. As a result, the 54 second-level domain names we tracked include: i) prestigious journals, such as `nature.com`, `sciencemag.org`, and `pnas.org`; ii) open access journals including `plosone.org`; iii) preprint services, such as `arXiv.org` and `papers.ssrn.com`; and iv) online library sites where many journals are available online, such as `jstor.org` and `sciencedirect.com`. By extracting unique identifiers of each paper to resolve an issue that different URLs point to the same article, we collect 403,165 tweets written by 142,743 users between 21st January 2013

[1] `http://www.altmetric.com/`

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576956.

and 4th June 2013. We crawl the whole social graph among them and user profiles as well.

3. HUGE SCHOLARLY LANDSCAPE AND PROMINENT ROLE OF GRASSROOTS

Among tweets of 54 scholarly domains, we find that 136,139 tweets (33.77%), written by 57,699 users, shared 31,661 unique research papers published in the top 100 journals. Surprisingly, on the reconstructed follow network, we discover that 98,308,648 unique users are exposed to at least one paper published on 54 scholarly domains, with 29,946,873 users exposed to a paper from the top 100 journals. This wide reach to the audience shows the strong potential of Twitter as a platform for disseminating papers since there is already a custom of sharing not only interesting news or gossip, but also interesting research papers.

We then look into who drives these dynamics. We find 211 official accounts related to the journals and scholarly domains we collected, e.g. `@NatureMagazine` and `@PLOSONE`, by detecting domain names in the homepage field of user profiles. They have 6,814 tweets for 4,042 papers published on 54 second-level domain; among them there are 2,750 tweets for 2,160 papers published in the top 100 journals, which is about 6.8% of the top-100 papers shared in our dataset. Furthermore, on the reconstructed network, we discover that a set of followers of the official accounts of top journals, which are Nature, Science, SSRN, PLOS ONE, and British Medical Journal, reaches only 2.86% of who are exposed to papers published in those journals, even considering all the retweets. Grassroot efforts, as opposed to official accounts of scholarly journals, are more active in sharing papers and reach a wider audience. This is another evidence to support that Twitter has a strong foundation for paper sharing.

4. DO HIGHER RANKING JOURNAL PAPERS LEAD TO MORE RESPONSES?

We first discover skewed social responses towards a few number of top journals via the Gini coefficient, which is widely used to summarize the inequalities of a distribution as a value between 0 (perfect equality) and 1 (maximal inequality). We compute the Gini coefficients for the distributions of the number of tweets, sharers, and readers of the top 100 journals: .8133 for tweets, .8015 for sharers, and .8018 for readers. In other words, online popularity, in terms of tweets, sharers, and readers, is highly skewed even among the top 100 journals listed by the research communities. It implies that the current practice of measuring the scientific impact of research papers through social media does not work well, even with major journals, except a few number of prestigious journals, such as Nature or Science.

We then investigate a correlation between a journal ranking based on citations and social responses in Twitter. We rank journals based on the h-5 citation index from Google Scholar and the average number of tweets sharing papers published in the corresponding journal. We find that the Spearman correlation coefficient is -.2385. Although the weak coefficient seems to imply that higher ranking journals lead to more social responses to some extent, it is an illusion coming from just a few of the top journals. If we filter out top 10 journals, the coefficient computed from the remaining 90 journals becomes -.0670. It indicates that, except for a few number of top journals, established reputation in research communities does not correlate to the volume of responses in social media at all. Higher ranking journals do not lead to more social responses.

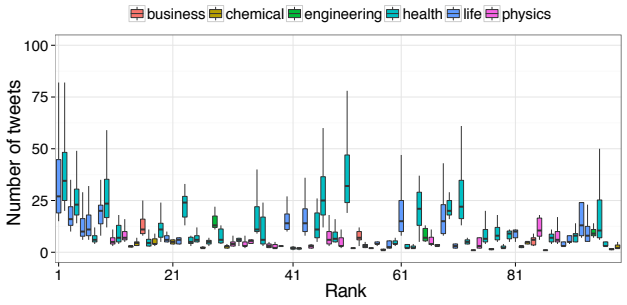

Figure 1: Tweets from the top 10% of papers in each journal (categories are assigned by Google Scholar)

The inconsistency between research communities and social media is also observed even when we focus on only popularly shared papers. Figure 1 shows the box plots for the number of tweets about the top 10% most shared papers in each journal. We find hardly any meaningful tendency between journal rank and the popularity of top papers represented by tweet volume. Moreover, we find that i) among the top 1 papers of 100 journals, 24 are not tweeted more than 10 times; ii) among the top 1%, the median of tweets per paper of 42 journals is less than 10; and iii) among the top 10%, the median of tweets per paper of 69 journals is less than 10. This low volume of social responses, even to many of the top 100 journals, shows the limited attention to research papers in social media.

Through our findings, which are the skewed popularity distribution of journals, the inconsistency between social responses and established reputation in research communities, and limited attention to research papers, we emphasize that the current practice to measure the volume of responses to research papers in social media should be carefully considered. Twitter users are highly biased towards a few number of top journals, and there is no association between social responses and the established reputation in research communities when we look at the entire scholarly landscape.

5. ACKNOWLEDGMENT

We thank Jeremy Blackburn and Jaimie Yejean Park for comments on earlier version of this manuscript.

6. REFERENCES

[1] J. Priem and B. H. Hemminger. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First Monday*, 15(7), July 2010.

[2] J. Priem, H. A. Piwowar, and B. M. Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv*, 1203.4745, 2012.

[3] R. C. Roemer and R. Borchardt. From bibliometrics to altmetrics: A changing scholarly landscape. *College & Research Libraries News*, 73(10):596–600, 2012.

[4] X. Shuai, A. Pepe, and J. Bollen. How the scientific community reacts to newly submitted preprints: Article downloads, Twitter mentions, and citations. *PLoS ONE*, 7(11):e47523, 2012.