**Submitted for presentation** at ASIS&T SIG/MET Metrics 2015 workshop

# Identifying Twitter audiences: who is tweeting about scientific papers?

Stefanie Haustein*,[1] & Rodrigo Costas[2]

*stefanie.haustein@umontreal.ca

[1] École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, QC. H3C 3J7 (Canada)

[2] Center for Science and Technology Studies, Leiden University, Wassenaarseweg 62A, 2333 AL Leiden (The Netherlands)

## Introduction

Due to the high uptake of Twitter by the general public, the number of tweets mentioning scientific articles is discussed as a potential indicator of their impact on society at large. About one fifth of current journal papers receive at least one tweet (Haustein, Costas, & Larivière, 2015). At the same time, as few as 10% to 15% of researchers use Twitter for work (Rowlands, Nicholas, Russell, Canty, & Watkinson, 2011; Van Noorden, 2014) and the share of their tweets linking to scientific publications is quite low (Priem & Costello, 2010). The increasing share of journal articles on Twitter and the low uptake by researchers suggests that scientific articles might be discussed by users that are not part of the scholarly community. Although a study reported that almost half of the Twitter users linking to *Science*, *Nature*, *PNAS* and *PLOS ONE* papers had an academic background (Tsou, Bowman, Ghazinejad, & Sugimoto, 2015), the typology of users tweeting scientific articles remains largely unknown. Particularly in the context of altmetrics, user types and engagement should be identified to specify the type of impact tweet counts capture. The objective of this paper is to identify groups of users tweeting about scientific papers by analyzing their Twitter account descriptions (i.e., Twitter "bios"), number of followers as well as the degree to which they engage with the tweeted papers (Haustein, Bowman, & Costas, 2015).

## Methods

The study focuses on users tweeting papers published in 2012 in journals covered by the Web of Science. Tweets were obtained from Altmetric.com and matched via the DOI as described in Haustein, Costas, and Larivière (2015). Retweets were excluded to focus on original contributions on Twitter. Twitter account descriptions (i.e., 160 character account descriptions for users to present themselves), number of followers and other profile information were retrieved from Twitter in April 2015 using the Twitter handle from Altmetric.com. 8.9% of accounts had to be excluded as they could not be found because they had either been deleted or users had changed handles. The set of remaining 115,053 Twitter accounts were further filtered to 89,768 accounts with an English setting ('en', 'en-gb') to reduce the number of non-English descriptions, which could not be properly processed in the co-word analysis. Following Haustein, Bowman, and Costas (2015), accounts were divided into four quadrants (categories A, B, C, D; see Figure 2) according to their number of followers and the dissimilarity between the tweet text and the title of the tweeted paper. While the followers reflect the potential size of audience or exposure, the dissimilarity estimates the engagement of the Twitter users based on the degree to which the tweet differs from the title of the paper. Many accounts, including automated bots (Haustein, Bowman, Holmberg, et al., 2015), were found to simply tweet paper titles without discussing their content (Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013).

Account descriptions were analyzed, for those 80,939 account descriptions containing some text, by extracting noun phrases using VOSviewer (van Eck, Waltman, Noyons, & Buter, 2010). As the extraction is based on a linguistic part-of-speech tagger for English texts, it did not work properly for non-English account descriptions. The algorithm extracted 185,824 unique terms (merging English regular singular and plural forms) from 78,991 accounts. VOSviewer was also used for visualizing and clustering the most frequent terms as shown in Figure 1. The visualization was restricted to terms occurring at least 100 times resulting in a network of 325 co-occurring terms[1]. A clustering resolution of 0.9 and minimum cluster size of 5 resulted in three clusters indicated by the color and labeling in Figure 1A. In Figure 1B node size was changed from the number of Twitter accounts including the term in their description to the mean number of followers and node color indicates the average engagement of accounts associated with a particular term.

[1] The restriction to 100 occurrences included 327 terms, but "http" and "tco" were removed manually as they referred to URLs mentioned in the account descriptions.

## Preliminary results and discussion

As shown in Figure 1A, cluster 3 can be clearly identified as 'academic' based on terms such as *university*, *science*, *professor* and *PhD*, which reflect that academics often identify themselves professionally on Twitter (Bowman, 2015). Cluster 1 consists of terms that describe users with a focus on 'personal' attributes such as *life*, *lover*, *father*, *husband*, *fan* or *geek* as well as non-academic professional terms such as *consultant*, *advocate*, *work*, *founder*, *co-founder* or *entrepreneur*. Overlapping with cluster 3, cluster 1 also contains terms such as *scientist*, *student* and *biologist*. As the network structure is based on the co-occurrence of terms, it can be inferred that Twitter descriptions are often used to identify professionally but reveal also private interests. The terms in cluster 2 seem to focus on 'topics and collectives', suggesting descriptions for interest groups, organizations or journals, particularly on health-related topics. As shown by the node colors in Figure 1B, there is a clear tendency of the academic and personal cluster to show more engagement (red) than the organizational cluster (blue). This might be explained by accounts that frequently distribute only paper titles, for example, those of scientific journals, publishers, or even automated accounts (Haustein, Bowman, Holmberg, et al., 2015).

Figure 2 highlights terms according to the classification based on the number of followers and average engagement of the accounts that they are mostly associated with. It can be seen that accounts classified as A or B use mostly terms from the personal and academic clusters, while terms from the cluster 2 appear mostly in Twitter descriptions of accounts with a high exposure but low engagement (category C).

## Conclusions

The analysis of terms used in Twitter descriptions suggests that scientific papers are tweeted by individuals, who identify professionally, personally or both, as well as organizations or interest groups. While accounts with organizational descriptions (cluster 2) seemed to have a more disseminative role, accounts with academic or personal terms (cluster 1 & 3) exhibited higher engagement (as measured by the similarity between paper title and tweet text).

## Acknowledgments

This work was partly funded by the Alfred P. Sloan Foundation Grant #G-2014–3–25. The authors would like to thank Timothy D. Bowman for the retrieval of Twitter data.

## References

Bowman, T. D. (2015, July). Investigating the use of affordances and framing techniques by scholars to manage personal and professional impressions on Twitter (Dissertation). Indiana University, Bloomington, IN, USA. Retrieved from http://www.tdbowman.com/pdf/2015_07_TDBowman_Dissertation.pdf

Haustein, S., Bowman, T. D., & Costas, R. (2015). "Communities of attention" around scientific publications: who is tweeting about scientific papers? Presented at the Social Media & Society 2015 International Conference, Toronto, Canada.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2015). Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology*. http://doi.org/10.1002/asi.23456

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLoS ONE*, 10(3), e0120495. http://doi.org/10.1371/journal.pone.0120495

Priem, J., & Costello, K. L. (2010). How and why scholars cite on Twitter. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1–4. http://doi.org/10.1002/meet.14504701201

Rowlands, I., Nicholas, D., Russell, B., Canty, N., & Watkinson, A. (2011). Social media use in the research workflow. *Learned Publishing*, 24(3), 183–195. http://doi.org/10.1087/20110306

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., & Haustein, S. (2013). Tweeting Links to Academic Articles. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometric*s, 17(1), 1–8.

Tsou, A., Bowman, T. D., Ghazinejad, A., & Sugimoto, C. R. (2015). Who tweets about science? In P*roceedings of the 2015 International Society for Scientometrics and Informetrics*. Istanbul, Turkey.

van Eck, N. J., Waltman, L., Noyons, E. C. M., & Buter, R. K. (2010). Automatic term identification for bibliometric mapping. Scientometrics, 82(2), 581–596.

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. Nature, 512(7513), 126–129. http://doi.org/10.1038/512126a

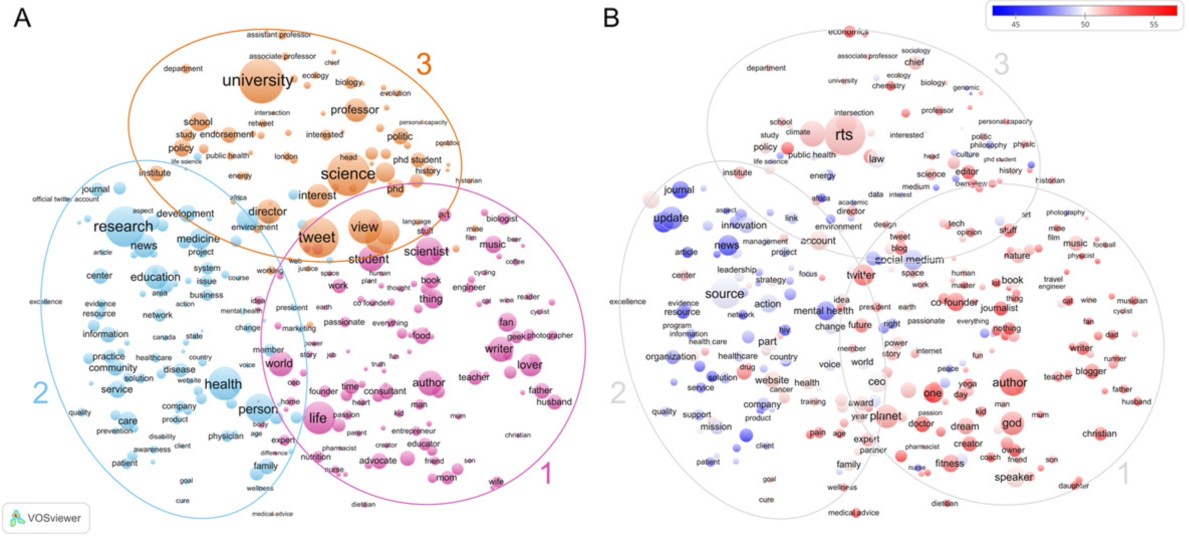

**Figure 1**. Co-occurrence network of most frequent terms in Twitter account descriptions. **A**: Node size represents number of accounts associated with a term, node color represents cluster affiliation. **B**: Node size represents average exposure of accounts associated with a term, node color represents average engagement of accounts associated with a term from low (blue) to high (red).

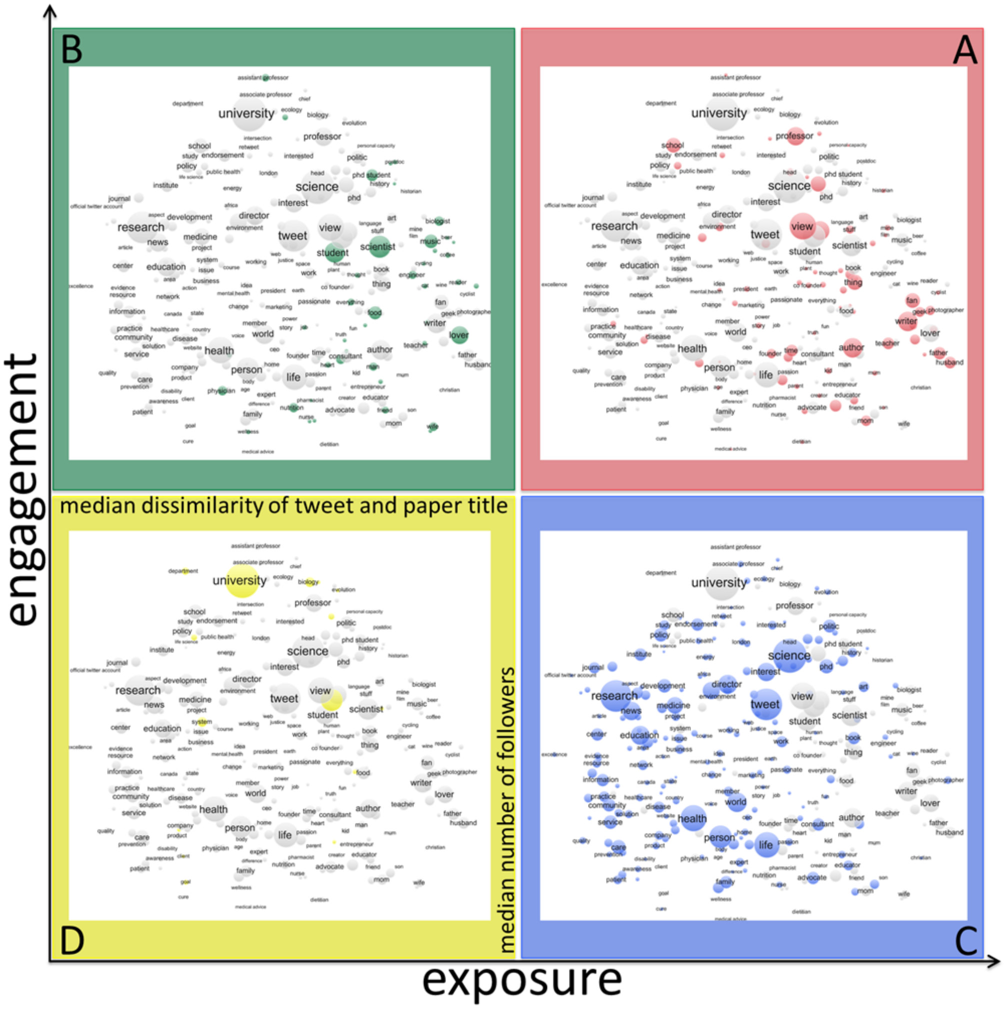

**Figure 2**. Classification of accounts according to exposure and engagement highlighting terms associated with accounts with high engagement and high exposure (A), high engagement and low exposure (B), low engagement and high exposure (C) and low engagement and low exposure. In each quadrant terms associated mostly with accounts classified in the particular category are highlighted.