Information retention in multi-platform discussions of science

Sohyeon Hwang[1], Emőke-Ágnes Horvát[1], and Daniel M Romero[2]
[1]Northwestern University; [2]University of Michigan
Questions can be sent to the corresponding author: sohyeonhwang@u.northwestern.edu

Online spread of science across platforms

Public interest in science communication [1], highlighted by recent public health crises, underscores how content loses critical pieces of information as it spreads [2]. Yet, multi-platform analyses remain limited due to challenges in reliable data collection. In this work, we leverage a large dataset to examine information retention in online discussions of scientific research findings "in the wild" across 5 platforms. We ask two main questions:

RQ1. How is information retained over time?

RQ2. As different types of platforms present different constraints about text, content, and posting, how does information retention differ across platforms?

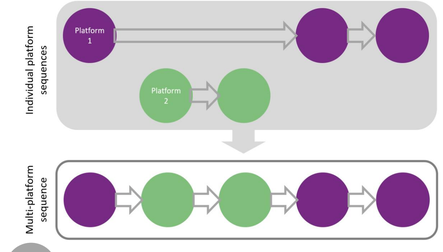
Research Design

Data. We leverage the 4+ million online mentions of 9,765 research articles tracked by Altmetric LLC [3] on blogs, Facebook, News, Twitter, and Wikipedia.

Measure development. We construct a keyword-based measure of "information retention", extracting keywords using the TextRank algorithm[4]; we validated the measure via a survey collecting expert labels on which mentions have more information retention:

$$\frac{sum(importance\ of\ abstract\ keyphrases\ found\ in\ post\ text)}{sum(importance\ of\ all\ abstract\ keyphrases)}$$

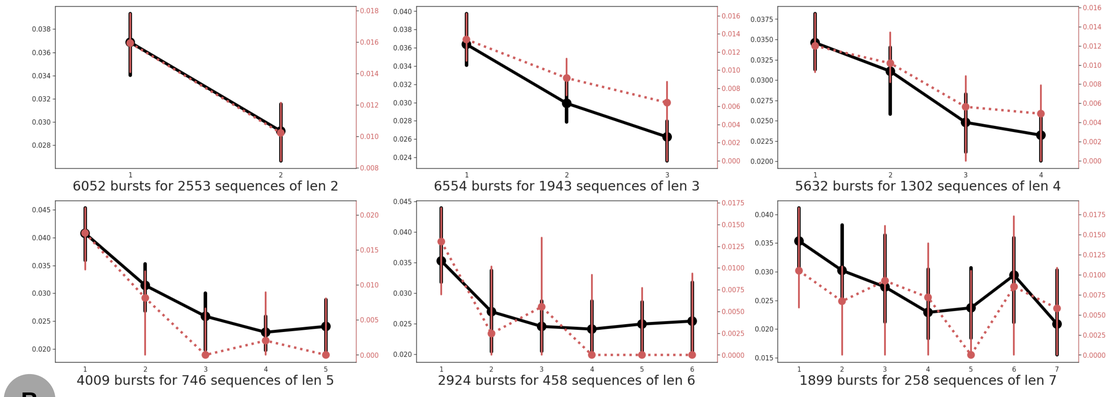
Burst-based framework. We present and use a burst-based [5] framework to identify meaningful aggregated cross-platform moments of attention to science online.



A

FIG A. shows a conceptual model of the burst-based framework.

RQ1. Information retention over time

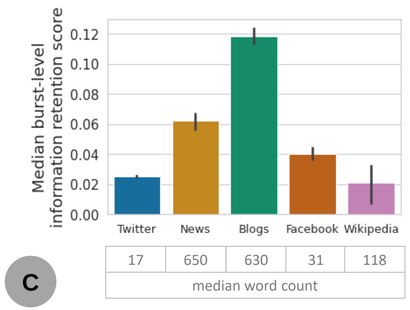


B

- Examining median information retention scores at the burst level, we found a strong propensity towards information loss in online mentions of science over time, for sequences of multiple lengths.
- However, sequences that started on social media platforms tended to start with and maintain low information retention.

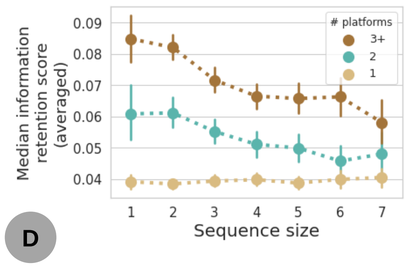
FIG B. shows the median information retention score at each sequential point over burst sequences of lengths 2-7; in dotted red is a robustness check, with our analysis replicated with RAKE-extracted keywords [6].

RQ2. Information retention across platforms



C

- Median **information retention** scores **varied across platforms**, and differences were not simply attributable to text length differences.



D

- **Sequences containing more platforms had higher median scores**, at all sequence lengths.

FIG C. shows the median information retention scores for mentions of different platform categories.
FIG D. shows median scores stratified by number of platforms, for sequence lengths 2-7.

Implications + Future work

- Patterns of information *loss* over time underscore a need to devise ways to mitigate such loss and test potential mechanisms driving it, such as research relevance and platform effects.
- Science discussions on *more* platforms tend to have *higher* information retention scores, suggesting multi-platform strategies can improve information retention. Future work should examine how to improve and synchronize information retention across platforms.

References
[1] National Science Board. 2020. Science and Technology: Public Attitudes, Knowledge, and Interest. Technical Report NSB-2020-7, National Science Foundation.
[2] Ribeiro, M.; Gligoric, K.; and West, R. 2019. Message Distortion in Information Cascades. In Proceedings of WWW 2019, 681–692. San Francisco, CA. ACM Press. doi:10.1145/3308558.3313531
[3] Altmetric Support. 2021. About Our Data: Our Sources. https://www.altmetric.com/about-our-data/our-sources/
[4] Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing Order into Texts. In Proc. of the 2004 Conference on EMNLP, 404–411. Barcelona, Spain. ACL:
[5] Barabasi, A.-L. 2005. The Origin of Bursts and Heavy Tails in Human Dynamics. *Nature* 435(7039): 207–211.
[6] Rose, S.; Engel, D.; Cramer, N.; and Cowley, W. 2010. Automatic Keyword Extraction from Individual Documents. In Berry, M. W.; and Kogan, J., eds., *Text Mining*, 1–20. Chichester, UK: John Wiley & Sons.

Acknowledgements. We thank students Amanda Hardy and Joshua Jacobs for their data assistance, Hao Peng for generously sharing code, and the Community Data Science Collective for early feedback. This work was supported by the NSF (DGE-1842165, IIS-2133963) and by the Air Force Office of Scientific Research under award number FA9550-19-1-0029.