Cultural transmission bias in the spread of voter fraud conspiracy theories on Twitter during the 2020 US election

(Preregistration of methods)

Mason Youngblood[1,2,*], Alberto Acerbi[3], Ryan Glassman, Olivier Morin[4,5], & Joseph Stubbersfield[6]

[1]Dept. of Psychology, Graduate Center, City University of New York, USA
[2]Dept. of Biology, Queens College, City University of New York, USA
[3]Center for Culture and Evolution, Brunel University London, UK
[4]Minds and Traditions Group, Max Planck Institute for the Science of Human History, DE
[5]Institut Jean Nicod, ENS, EHESS, PSL University, CNRS, FR
[6]Dept. of Psychology, University of Winchester, UK
*masonyoungblood@gmail.com

Introduction

The aim of the proposed study is to investigate whether retweet frequencies among proponents of voter fraud conspiracy theories on Twitter during the 2020 US election are consistent with content bias, demonstrator bias, and/or frequency bias. To do this, we will use an agent-based model (ABM) with parameters that correspond to each form of bias. The output of this ABM will be fit to the real data using the random forest version of approximate Bayesian computation (ABC) to infer the parameter values that likely generated the observed retweet frequencies.

Additionally, we aim to conduct secondary analyses to assess the potential targets of content or demonstrator biases. The emotional content of tweets will be measured using sentiment analysis, whereas demonstrator attractiveness will be based on follower count and whether they are verified. A large body of research suggests that content with negative sentiment has an advantage over content with positive sentiment across several domains. A bias towards negative sentiment has been found in recall, perception, and impression formation (Baumeister et al., 2001; Rozin & Royzman, 2001), recall-based social transmission (Bebbington et al., 2017; Walker & Blaine, 1991), credulity (Fessler, 2014; Hilbig, 2009, 2011, 2012) and cultural artefacts such as song lyrics (Brand et al., 2019). In digital media, evidence of negative bias has been suggested for "fake news" articles (Acerbi, 2019) and tweets about a climate change summit (Hansen et al, 2011), or, more generally, about political events (Schöne et al., 2021).

Some research, however, has demonstrated an advantage for content with positive sentiment when sharing information with others. Van Leeuwen and colleagues (2018) found an advantage for positive vignettes in decisions to share with strangers (but not friends), and Ferrara and Yang (2015) found that, while tweets with negative sentiment were retweeted more quickly, they received fewer retweets than tweets with positive sentiment. Studies have suggested that the strength of emotion evoked by information content influences the transmission of that content. In general, stronger emotional content (regardless of sentiment direction) is associated with increased attention (Fernández-Martín & Calvo, 2015), improved recall-based transmission (Kashima et al., 2020; Stubbersfield et al, 2017) and increased choice to transmit to others (Berger, 2011; Berger & Milkman, 2010; Steiglitz & Dang-Xuan, 2013). Similarly, other studies evidenced that emotional language, in general, influences the diffusion of content in social media (Brady et al., 2107).

Based on this research, if content bias is detected then we hypothesize that it will be targeted towards stronger emotional content, but we remain agnostic as to the direction of the emotion (positive or negative). If demonstrator bias is detected then we hypothesize that it will be targeted towards individuals that are verified and have more followers.

Data

The data for this study comes from the *VoterFraud2020* dataset[1], collected between October 23, 2020 and December 16, 2020 by Abilov et al. (2021). This dataset includes 7.6 million tweets that were collected in real time using Twitter's streaming API. Abilov et al. (2021) started out with a set of keywords and hashtags that co-occurred with "voter fraud" and "#voterfraud" between July 21, 2020 and October 22, 2020, and expanded their search with additional keywords and hashtags as they emerged (e.g. "#discardedballots" and #stopthesteal"). They estimate that their dataset includes at least 60% of tweets that included their search terms. Abilov et al. (2021) also applied the infomap clustering algorithm to the directed retweet network to identify different communities that engaged with the voter fraud conspiracy theory. We will run our analysis using only the user and tweet data from cluster #2, the "proponent" community that tweets primarily in English and does not have significant connections to members of the "detractor" community, so that retweet events are more indicative of cultural transmission.

Methods

ABM and ABC

The agent-based model (ABM) we will use has elements from Carrignon et al. (2019), Lachlan et al. (2018), and Youngblood and Lahti (2021). The ABM is initialized with a fully-connected population of N users and is run for 216 timesteps, each of which correspond to a six-hour interval in the real dataset (the highest resolution possible given computational limits). Each user is assigned a follower count (T) and an activity level (r) drawn randomly from the observed data. T is scaled with a mean of 1 and a standard deviation of 1. Follower counts greater than or equal to 100,000 (0.087%) will be excluded, as they flatten nearly all variation in T after scaling. The ABM is also initialized with a set of tweets with retweet frequencies drawn randomly from the first timestep in the observed data. Each tweet is assigned an attractiveness (M). At the start of each timestep, a pseudo-random subset of users becomes active (weighted by their values of r) and tweets according to the observed overall level of activity in the same timestep. All active users have the same

[1] https://voterfraud2020.io/

probability of tweeting an original tweet (μ) as opposed to retweeting an existing tweet ($1 - \mu$), based on the proportion of original tweets in the real dataset. New original tweets are assigned an attractiveness of M, while retweets occur with probability P_x:

$$P_x = F_x^a \cdot T_x^d \cdot M_x^c \cdot \frac{1}{age_x^g}$$

F is the number of times that a tweet has been previously retweeted, and is raised by the level of frequency bias (a). a is the same across all agents, where values > 1 simulate conformity bias and values < 1 simulate novelty bias. T is raised by the level of demonstrator bias (d). d is the same across all agents, where values of 0 simulate neutrality by removing variation in follower count and values > 0 simulate increasing levels of demonstrator bias. M is the attractiveness of the tweet, and is drawn from a truncated normal distribution with a mean of 1, a standard deviation of 1, and a lower bound of 0. M is raised by the level of content bias (c). c is the same across all agents, where values of 0 simulate neutrality by removing variation in the attractiveness of content and values > 0 simulating increasing levels of content bias. Lastly, the final term simulates the decreasing probability that a tweet is retweeted as it ages, where g controls the rate of decay. Once the active users are done each tweet increases in age by 1 and the next timestep begins.

In summary, the following are the dynamic parameters in this ABM that we will estimate using approximate Bayesian computation (ABC):

- a - level of frequency bias
- d - variation in the salience of follower count
- c - variation in the salience of the attractiveness of content
- g - rate of decay in tweet aging

All other parameters in the ABM will be assigned static values based on the real dataset. The output of this ABM is a distribution of retweet frequencies (see Figure 1), which will be used to calculate the following summary statistics: (1) the proportion of tweets that only appear once, (2) the proportion of the most common tweet, (3) the Hill number when $q = 1$ (which emphasizes more rare tweets), and (4) the Hill number when $q = 2$ (which emphasizes more common tweets). We will use Hill numbers rather than their traditional diversity index counterparts (Shannon's and Simpson's diversity) because they are measured on the same scale and better account for relative abundance (Chao et al., 2014; Roswell et al., 2021).

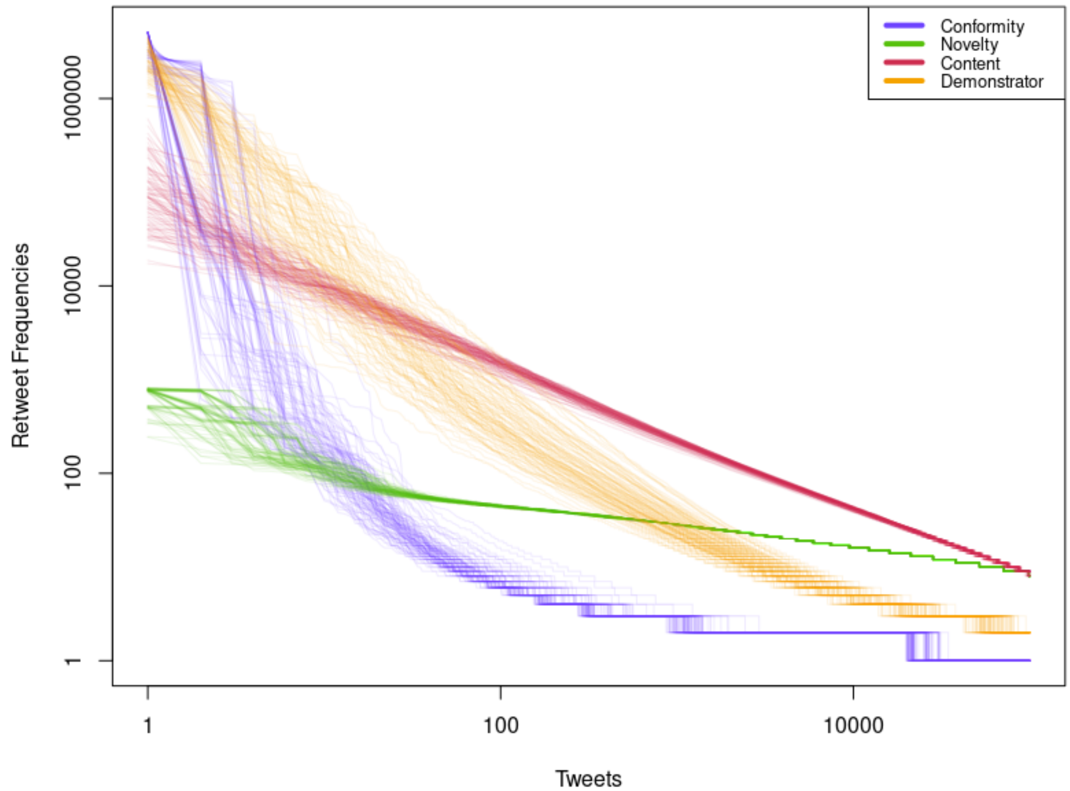

Figure 1. The retweet distributions resulting from conformity, novelty, content, and demonstrator bias using this ABM (100 iterations each). Biases were modelled with the following parameter values: a = 1.4 (conformity), a = 0.6 (novelty), c = 1 (content), and d = 1 (demonstrator). The x-axis (the identity of each tweet) and the y-axis (the number of times each tweet was retweeted) have been log-transformed.

The same summary statistics will be calculated from the observed retweet distribution of the real dataset. For purposes of the summary statistic calculations quote tweets will be treated like original tweets, as they themselves can be retweeted. Then, the random forest version of ABC (Raynal et al., 2018) will be conducted with the following steps:

- 200,000 iterations of the ABM will be run to generate simulated summary statistics for different values of the parameters: c, a, d, and g. More iterations may be needed depending on levels of out-of-bag error during the final step.
- The output of these simulations will be combined into a reference table with the simulated summary statistics as predictor variables, and the parameter values as outcome variables.
- A random forest of 1,000 regression trees will be constructed for each of the four parameters using bootstrap samples from the reference table.
- Each trained forest will be provided with the observed summary statistics, and each regression tree will be used to predict the parameter values that likely generated the data.

We estimate that 200,000 iterations of the ABM would take about 134 days to run in serial, so the analysis will be conducted in parallel using the High Performance Computing Center at the College of Staten Island, City University of New York.

Sentiment Analysis

Sentiment analysis will be conducted using the valence aware dictionary for sentiment reasoning (VADER), a model specifically designed for use with social media posts from platforms like Twitter (Hutto & Gilbert, 2014). VADER is available through the natural language toolkit in Python, and outputs both the strength (low to high) and the direction (positive to negative) of emotion in a text.

General Linear Modeling

To determine the potential targets of content and demonstrator biases we will conduct Bayesian general linear modeling (GLM). Retweet count will be used as the outcome variable, and the following will be used as predictor variables: strength of tweet sentiment (low to high), direction of tweet sentiment (positive to negative), follower count, and verification status (T/F).

References

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications, 5*(1), 1–7. https://doi.org/10.1057/s41599-019-0224-y

Abilov, A., Hua, Y., Matatov, H., Amir, O., & Naaman, M. (2021). VoterFraud2020: a Multi-modal Dataset of Election Fraud Claims on Twitter. *arXiv*. http://arxiv.org/abs/2101.08210

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology, 5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Bebbington, K., MacLeod, C., Ellison, T. M., & Fay, N. (2017). The sky is falling: Evidence of a negativity bias in the social transmission of information. *Evolution and Human Behavior, 38*(1), 92–101. https://doi.org/10.1016/j.evolhumbehav.2016.07.004

Berger, J. (2011). Arousal Increases Social Transmission of Information. *Psychological Science, 22*(7), 891–893. https://doi.org/10.1177/0956797611413294

Berger, J., & Milkman, K. (2010). Social transmission, emotion, and the virality of online content. *Wharton Research Paper, 106*.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28), 7313-7318. https://doi.org/10.1073/pnas.1618923114

Brand, C. O., Acerbi, A., & Mesoudi, A. (2019). Cultural evolution of emotional expression in 50 years of song lyrics. *Evolutionary Human Sciences, 1*. https://doi.org/10.1017/ehs.2019.11

Carrignon, S., Bentley, R. A., & Ruck, D. (2019). Modelling rapid online cultural transmission: evaluating neutral models on Twitter data with approximate Bayesian computation. *Palgrave Communications, 5*(83). https://doi.org/10.1057/s41599-019-0295-9

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. and Ellison, A.M. (2014), Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs, 84*, 45–67. https://doi.org/10.1890/13-0133.1

Fernández-Martín, A., & Calvo, M. G. (2015). Extrafoveal capture of attention by emotional scenes: Affective valence versus visual saliency. Visual Cognition, 23(9–10), 1061–1071. https://doi.org/10.1080/13506285.2016.1139026

Ferrara, E., & Yang, Z. (2015). Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science, 1*, e26. https://doi.org/10.7717/peerj-cs.26

Fessler, D. M. T., Pisor, A. C., & Navarrete, C. D. (2014). Negatively-Biased Credulity and the Cultural Evolution of Beliefs. *PLoS ONE, 9*(4). https://doi.org/10.1371/journal.pone.0095167

Fisher, R., Corbet, A., & Williams, C. (1943). The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology, 12*(1), 42-58. https://doi.org/10.2307/1411

Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., & Etter, M. (2011). Good Friends, Bad News—Affect and Virality in Twitter. In J. J. Park, L. T. Yang, & C. Lee (Eds.), *Future Information Technology* (pp. 34–43). Springer. https://doi.org/10.1007/978-3-642-22309-9_5

Hilbig, B. E. (2009). Sad, thus true: Negativity bias in judgments of truth. *Journal of Experimental Social Psychology, 45*(4), 983–986. https://doi.org/10.1016/j.jesp.2009.04.012

Hilbig, B. E. (2011). Good Things Don't Come Easy (to Mind). *Experimental Psychology, 59*(1), 38–46. https://doi.org/10.1027/1618-3169/a000124

Hilbig, B. E. (2012). How framing statistical statements affects subjective veracity: Validation and application of a multinomial model for judgments of truth. *Cognition, 125*(1), 37–48. https://doi.org/10.1016/j.cognition.2012.06.009

Hutto, C. J., & Gilbert, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–225.

Kandler, A., & Crema, E. R. (2019). Analysing Cultural Frequency Data: Neutral Theory and Beyond. In A. M. Prentiss (Ed.), *Handbook of Evolutionary Research in Archaeology* (pp. 83–108). Springer International Publishing. https://doi.org/10.1007/978-3-030-11117-5_5

Kashima, Y., Coman, A., Pauketat, J. V. T., & Yzerbyt, V. (2020). Emotion in Cultural Dynamics. *Emotion Review, 12*(2), 48–64. https://doi.org/10.1177/1754073919875215

Lachlan, R. F., Ratmann, O., & Nowicki, S. (2018). Cultural conformity generates extremely stable traditions in bird song. *Nature Communications*, *9*. https://doi.org/10.1038/s41467-018-04728-1

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2018). ABC random forests for Bayesian parameter inference. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty867

Roswell, M., Dushoff, J. and Winfree, R. (2021). A conceptual guide to measuring species diversity. *Oikos, 130*, 321–338. https://doi.org/10.1111/oik.07202

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review, 5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Schöne, J., Parkinson, B., & Goldenberg, A. (2021, January 2). Negativity Spreads More than Positivity on Twitter after both Positive and Negative Political Situations. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/x9e7u

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems, 29*(4), 217–248. https://doi.org/10.2753/MIS0742-1222290408

Stubbersfield, J. M., Tehrani, J. J., & Flynn, E. G. (2017). Chicken Tumours and a Fishy Revenge: Evidence for Emotional Content Bias in the Cumulative Recall of Urban Legends. *Journal of Cognition and Culture, 17*(1–2), 12–26. https://doi.org/10.1163/15685373-12342189

van Leeuwen, F., Parren, N., Miton, H., & Boyer, P. (2018). Individual Choose-to-Transmit Decisions Reveal Little Preference for Transmitting Negative or High-Arousal Content. *Journal of Cognition and Culture*, *18*(1–2), 124–153. https://doi.org/10.1163/15685373-12340018

Walker, C. J., & Blaine, B. (1991). The virulence of dread rumors: A field experiment. *Language & Communication, 11*(4), 291–297. https://doi.org/10.1016/0271-5309(91)90033-R

Youngblood, M., & Lahti, D. (2021). Content bias in the cultural evolution of house finch song. *bioRxiv*, 1–14. https://doi.org/10.1101/2021.03.05.434109