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

Mason Youngblood[1,2,3,*], Joseph M. Stubbersfield[4], Olivier Morin[3,5], Ryan Glassman[6], Alberto Acerbi[7]

[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]Minds and Traditions Group, Max Planck Institute for the Science of Human History, DE
[4]Dept. of Psychology, University of Winchester, UK
[5]Institut Jean Nicod, ENS, EHESS, PSL University, CNRS, FR
[6]IBM Watson, 51 Astor Place, New York, NY, USA
[7]Centre for Culture and Evolution, Brunel University London, UK
*youngblood@shh.mpg.de

Abstract

During the 2020 US presidential election, conspiracy theories about large-scale voter fraud were widely circulated on social media platforms. Given their scale, persistence, and impact, it is critically important to understand the mechanisms that caused these theories to spread so rapidly. The aim of this study was to investigate whether retweet frequencies among proponents of voter fraud conspiracy theories on Twitter during the 2020 US election are consistent with frequency bias, demonstrator bias, and/or content bias. To do this, we conducted generative inference using an agent-based model of cultural transmission on Twitter and the *VoterFraud2020* dataset. The results show that the observed retweet distribution is consistent with a strong content bias and demonstrator bias, likely targeted towards negative emotion and follower count, respectively. Based on the confounding effects of the timeline algorithm and population structure, we are most confident in concluding that the differential spread of voter fraud claims among proponents of voter fraud conspiracy theories on Twitter during and after the 2020 US election was partly driven by a content bias causing users to preferentially retweet tweets with more negative emotion.

Keywords: voter fraud, conspiracy theory, social media, transmission bias, negative emotion

Introduction

Allegations of malign acts, carried out in secret by powerful groups, have been offered as explanations for major events throughout history, from ancient Rome, through the medieval period, to the present day (Brotherton, 2015; Pagán, 2020; Zwierlein, 2020). People across the globe have shared these conspiracy theories (Butter & Knight, 2020; West & Sanders, 2003). Conspiracy theories have been a part of North American culture since the colonial period, with beliefs about conspiring, "un-American" groups of witches, enslaved Africans, Masons, Catholics, and Jews dominating early versions (Olmsted, 2018). Later, in the twentieth century, the focus shifted towards the US government itself as the source of conspiring agents (Olmsted, 2018).

Conspiracy theories are typically defined as explanations of important events which allege secret plots by powerful actors as salient causes (Douglas et al., 2019; Goertzel, 1994; Keeley, 1999). Belief is not inherently irrational (as conspiracies do occur; see Dentith, 2014; Pigden, 1995), but conspiracy *theories* (as opposed to simply conspiracies) are allegations that survive and spread despite a lack of reliable evidence (Douglas et al., 2019; Keeley, 1999). Belief in conspiracy theories is associated with reduced engagement with mainstream politics (Imhoff et al., 2020; Jolley & Douglas, 2014), increased support for political violence and extremism (Imhoff et al., 2020; Uscinski & Parent, 2014), and increased prejudice towards minority groups (Jolley et al., 2020; Kofta et al., 2020).

A range of recent and ongoing conspiracy theories allege that the result of the 2020 US presidential election was achieved through large-scale electoral fraud (Enders et al., 2021). Building on allegations of voter fraud made prior to the 2016 election (Cottrell et al., 2018) and years of Republican messaging about electoral fraud and illegal voting (Edelson et al., 2017), these conspiracy theories were widely circulated on social media platforms like Twitter. Major political and public figures, including US President Donald Trump, boosted these theories using hashtags like #stopthesteal (Sardarizadeh & Lussenhop, 2021) and eventually had their accounts suspended for incitement of violence following the January 6th attack on the US Capitol (Conger & Isaac, 2021). More specific claims, such as hacked voting machines being programmed in favor of then-Presidential Candidate Joe Biden and large numbers of ballots being thrown out in trash bags (Cohen, 2021; Spring, 2020) have been used to justify election audits and tighter voting laws in states like Arizona (Cooper & Christie, 2021) and Georgia (Corasaniti & Epstein, 2021). The Justice Department has found "no evidence of widespread voter fraud" (Balsamo, 2020), and the Cybersecurity and Infrastructure Security Agency concluded that 2020 was "the most secure election ever" (Tucker & Bajak, 2020). Despite this, polls suggest that

up to a third of Americans (Cillizza, 2021) and the majority of Republicans (Skelley, 2021) believe that Biden won the election illegitimately through voter fraud. Exposure to such claims has been shown to reduce confidence in democratic institutions (Albertson & Guiler, 2020) and is thought to have contributed to motivating the US Capitol attack (Beckett, 2021). Given the scale, persistence, and impact of voter fraud conspiracy theories, it is critically important to understand the mechanisms that caused them to spread so rapidly and widely.

While conspiracy theories, as everything else, are disseminated through social media, the nature of the association between social media usage and conspiracy theory belief is an open question (Enders et al., 2021; Hall Jamieson & Albarracín, 2020; Min, 2021; Stempel et al., 2007). Social media does provide, however, a source of data that can be used to test theories about their spreading. Most studies have focused on the content of social media posts, highlighting how negative content (Schöne et al., 2021), emotional content (Brady et al., 2017), or out-group derogation (Osmundsen et al., 2021; Rathje et al., 2021) tend to be associated with their spreading. However, content is only one of the possible features that influence the success of a social media post. In what follows, we use a framework inspired by cultural evolution that allows us to distinguish among various features and assess their relative importance.

Broadly, cultural evolution adopts an evolutionary framework to research the stability, change and diffusion of cultural traits (Mesoudi, 2011). Transmission biases—biases in social learning that cause individuals to adopt some cultural variants over others—are thought to be some of the most important factors driving cultural evolutionary patterns (Kendal et al., 2018). According to this perspective, the probability that a behavior will be adopted is influenced by various cues. Frequency bias, which includes conformity and novelty bias, is when the frequency of a variant in the population disproportionately affects its probability of adoption. Content bias is when the inherent characteristics of a variant affects its probability of adoption. Demonstrator bias is when some characteristic of the individuals expressing a variant affects its probability of adoption (see review in Kendal et al., 2018). Importantly, transmission biases can lead to discernible changes in the cultural frequency distributions of populations (Lachlan et al., 2018). For example, in the context of Twitter, a positive frequency bias (i.e. conformity) would cause users to be more likely to retweet content that has already been heavily retweeted by other users, thus increasing the right skew of the overall retweet distribution. This framework allows us to consider both individual susceptibility and the influence of social context on wider population level patterns.

Using generative inference, it is possible to infer the underlying cognitive biases of individuals in a population from the cultural frequency distribution that they generate. Generative inference is a statistical procedure in which a model is run many times with varying parameter values to generate large quantities of simulated data. This simulated data is then compared to real data using approximate Bayesian computation (ABC) to infer the parameter values that likely generated it (Kandler & Powell, 2018). Carrignon et al. (2019) recently applied generative inference to the spread of confirmed and debunked information on Twitter and found that the retweet distributions of both were more consistent with random copying than with conformity. However, their model did not include parameters for demonstrator bias and did not explore the influence of content bias due to computational limitations (Carrignon et al., 2019).

The aim of this study is to investigate whether retweet frequencies among proponents of voter fraud conspiracy theories on Twitter during the 2020 US election are consistent with frequency bias, demonstrator bias, and/or content bias. To do this, we conducted generative inference using an agent-based model (ABM) of cultural transmission on Twitter that combines elements from Carrignon et al. (2019), Lachlan et al. (2018), and Youngblood and Lahti (2021). Our ABM simulates a fully-connected population of Twitter users with randomly-assigned follower counts and activity levels from the real data. Every six hours, a subset of users become active and either compose a new tweet or retweet an existing tweet. The probability of an existing tweet being retweeted is based on four factors: (1) how many times it has already been retweeted, (2) the attractiveness of the user who tweeted it (e.g. follower count or verification status), (3) the attractiveness of the content in the tweet (e.g. emotional valence and/or intensity), and (4) the age of the tweet. The influence of each of these factors on retweet probability is controlled by separate parameters which correspond to frequency bias, demonstrator bias, content bias, and age dependency, which we fitted to real data using the random forest version of ABC (Raynal et al., 2019).

The data used in this study comes from a team of researchers at Cornell Tech, who retrieved millions of tweets and retweets relating to voter fraud conspiracy theories between October 23 and December 16 of 2020 (Abilov et al., 2021). After iteratively building a set of search terms from the seeds "voter fraud" and #voterfraud and using them to collect data in real-time, they estimate that they collected ~60% of tweets about voter fraud conspiracy theories during that period. An anonymized version of the *VoterFraud2020* dataset is publicly available[1], and Abilov et al. (2021) generously provided us with access to their full disambiguated dataset. Importantly, this dataset includes tweets from users who were "purged" from Twitter following the US Capitol attack (Romm & Dwoskin, 2021).

Additionally, we conducted secondary analyses with general linear mixed models (GLMM) to assess the potential targets of content or demonstrator biases. The emotional content of tweets was measured using sentiment

[1] https://voterfraud2020.io/

analysis, whereas demonstrator attractiveness was based on follower count and whether the account holder was verified. Twitter verifies some accounts to make sure they are authorized by the person they claim to represent, but only undertakes this costly verification for high-profile accounts—in practice a small but highly influential minority, whose status is signaled by a "blue check mark" icon. Sentiment analysis was conducted using the valence aware dictionary and sentiment reasoner (VADER), a model trained for use with Twitter and other social media data (Hutto & Gilbert, 2014). A large body of research suggests that content with negative sentiment has an advantage over content with positive sentiment across several domains (Baumeister et al., 2001; Rozin & Royzman, 2001). In digital media, evidence of negative bias has been suggested for "fake news" articles (Acerbi, 2019), within online "echo chambers" (Asatani et al., 2021; Del Vicario et al., 2016), and for tweets about political events both from individual users (Schöne et al., 2021) and institutions (Bellovary et al., 2021). Other studies have suggested that just the strength of emotion influences the transmission of content on social media (Brady et al., 2017; Stieglitz & Dang-Xuan, 2013; but see critiques in Burton et al., 2021), and, in two experimental studies, van Prooijen et al. (2021) found that the appeal of conspiracy theories was associated with the intensity of emotion evoked, rather than the valence of that emotion.

Demonstrator bias often manifests as a tendency to copy or adopt the behaviors of successful and/or prestigious demonstrators (Henrich & McElreath, 2003; Kendal et al., 2018), with prestige typically operationalized as being indicated by increased attention or deference (Jiménez & Mesoudi, 2019). In social media, Bakshy et al. (2011) found, for example, that the success of a tweet was correlated with the number of followers of the author of the tweet. Other research suggests that high profile political and media figures played a central role in the diffusion of voter fraud conspiracy theories (Benkler et al., 2020).

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 accounts that are verified or have more followers (i.e. receive more attention).

Results

The *VoterFraud2020* dataset is divided into several sub-communities, including both detractors and proponents of the conspiracy theories. We chose to focus on cluster #2, the "proponent" community that tweets and retweets content in English and does not have significant connections to the "detractor" community (see Methods). After subsetting the *VoterFraud2020* data to only include user and tweet data from cluster #2, we ended up with 3,982,990 tweets from 341,676 users. Note that we calculated the number of users as all unique users that either tweeted or retweeted content from cluster #2. The agent-based model was initialized with a population size (N) of 341,676 and an original tweet probability (μ) of 0.45, based on the proportion of original tweets in the dataset. The model and methods were preregistered in advance of the analysis (see Methods).

The posterior distributions for content bias (c), demonstrator bias (d), frequency bias (a), and age dependency (g) can be seen in Figure 1 and Table 1. Higher values of c, d, and g are indicative of stronger effects of those parameters, where 0 is neutrality. Values of a that are lower and higher than 1 are indicative of novelty and conformity bias, respectively, where 1 is neutrality. The median estimate for content bias is 2.61, with a relatively wide 95% credible interval (CI) that spans from 0.4 to 4.68. This indicates that content bias plays a significant role in driving retweet frequencies, but that its effect is either difficult to estimate or varies among users. The posterior distribution for demonstrator bias, on the other hand, has a strong peak at 2.21 with a tight 95% CI and low prediction error. This indicates that demonstrator bias also drives retweet frequencies to a similar degree, and that its effect is much more predictable and stable. The median estimate for frequency bias is 0.29, with an extremely wide 95% CI that is strongly right-skewed towards zero. If we assume that neutrality in frequency bias is $a = 1$, where retweet probability is perfectly proportional to the number of times a tweet has already been shared, then this result is indicative of a strong novelty bias. However, we feel that a novelty bias of this magnitude is unrealistic, and that this result instead suggests that retweet probability is mostly decoupled from the number of times a tweet has already been shared (so neutrality is $a = 0$). This is much more consistent with our personal experiences on Twitter, where the timeline includes a balance of new tweets from followers and trending tweets that have already been heavily retweeted. According to this interpretation, the wide 95% CI for a could reflect departures from neutrality resulting from conformity bias, variation in user behavior, or how Twitter's timeline algorithm weights trending tweets. The posterior distribution for age dependency is wide and chaotic but generally left-skewed, suggesting that newer tweets may have an advantage in terms of retweet probability.

Importantly, the details of Twitter's timeline algorithm are not publicly available, which means that it is not statistically possible for us to differentiate between a bias produced by the algorithm or user behavior. This issue is most relevant for the results related to demonstrator and frequency bias, as follower and retweet count are the two cues most likely utilized by Twitter's timeline algorithm (Koumchatzky & Andryeyev, 2017).

	M	95% CI	NMAE
c	2.61	[0.40, 4.68]	1.22
d	2.21	[1.69, 2.50]	0.083
a	0.29	[0.014, 0.84]	1.64
g	5.61	[0.60, 7.87]	3.26

Table 1. The median, 95% credible interval, and out-of-bag normalized mean absolute error of the posterior distribution for each dynamic parameter in the agent-based model.

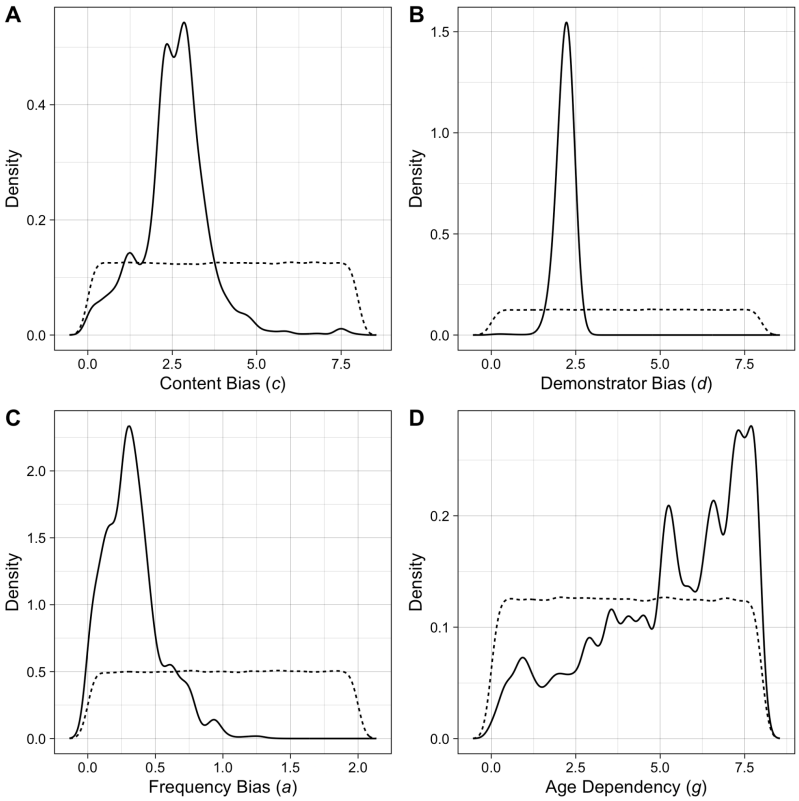

Figure 1. The prior (dotted lines) and posterior (solid lines) distributions for each of the four dynamic parameters from the ABM that were estimated using ABC.

Based on the null models for the GLMM, user appears to be the only grouping variable that explains a high level of variance in the data (ICC_{user} = 0.61, ICC_{date} = 0.11, ICC_{hour} = 0.04). As such, we chose to include user as a random effect in our base model. Adding tweet length as a control variable improved model fit (ΔAIC = 40803; LRT: χ^2 = 40805, $p < 0.0001$). Both follower count and verification status further improved model fit (ΔAIC > 2), but the model with follower account was significantly better (ΔAIC = 6217; LRT: χ^2 = 6217, $p < 0.0001$) so we updated our base model accordingly. All three content measures further improved model fit (ΔAIC > 2), but the model with the proportion of negative words was significantly better than the models with the proportion of positive words (ΔAIC = 460; LRT: χ^2 = 460, $p < 0.0001$) or the compound score (ΔAIC = 815; LRT: χ^2 = 815, $p < 0.0001$). All model specifications and AIC values for the primary GLMM are in Table S3, along with a partial specification curve analysis in Figure S5. The best fitting model included user as a random effect, and tweet length, follower count, and the proportion of negative words as predictor variables (see Table 2).

	IRR	95% CI
Tweet length	1.470	[1.468, 1.471]
Follower count	1.509	[1.499, 1.519]
Proportion negative	1.057	[1.055, 1.058]

Table 2. The incidence rate ratio (IRR) and 95% confidence interval for each predictor in the best fitting model. IRR, the exponentiated beta estimate, is interpreted as the rate at which the outcome variable is expected to change per unit increase in a predictor (one standard deviation for scaled and centered predictors). Wald confidence intervals were used due to the high sample size.

Tweet length, follower count, and the proportion of negative words all have significant effects on retweet frequency (Table 2). The incidence rate ratio (IRR) for tweet length is 1.470, indicating that if a tweet is one standard deviation (SD) longer then it is 47.0% more likely to be retweeted. Follower count has a similarly strong effect, where tweets from users with one SD more followers are 50.9% more likely to be retweeted. The IRR for the proportion of negative words is much lower but still significant. Tweets with a proportion of negative words that is one SD higher are 5.7% more likely to be retweeted. Pseudo R^2 values, calculated using log-normal approximation, indicate that the predictor variables alone account for about 10% of the variance in the data ($R^2 = 0.099$), whereas the predictor variables and random effects together account for about 68% of the variance in the data ($R^2 = 0.68$) (Nakagawa et al., 2017). A variance inflation factor test indicates that there are no significant issues with multicollinearity between predictors (VIFs < 2). Residual diagnostics for the best fitting model indicate that, while there are some extreme low and high outliers, the Poisson family is appropriate and there are no significant problems with overdispersion (see Figure S6).

An additional GLMM found that quote tweets tend to have reduced negative emotion relative to the original tweets that they are quoting (see SI). This lends support to our generative inference results, given our decision to treat quote tweets like original tweets when computing retweet distributions (see Methods). If quote tweets tend to be less attractive than original tweets, then our estimate for content bias is likely more conservative than it would have been if we had treated quote tweets like retweets instead of original tweets.

Discussion

Based on the results of generative inference, the observed retweet distribution is consistent with a strong content bias and demonstrator bias. Tweets with a higher ratio of negative words and from users with more followers are more likely to be retweeted, suggesting that these biases are targeted towards negative emotion and follower count, respectively. For frequency bias, the results of generative inference could be interpreted as evidence for either an extremely strong novelty bias or for frequency information being irrelevant. We feel that a novelty bias of this magnitude is unrealistic, and we interpret this result as evidence that retweet probability is mostly decoupled from the number of times a tweet has already been retweeted. Interestingly, we also found that quote tweets tend to contain less negative emotion than their targets. This means that users, despite having a content bias for negative emotion, do not tend to amplify negativity when commenting on a retweet.

Importantly, the results for demonstrator bias and frequency bias are difficult to separate from the influence of Twitter's timeline algorithm, which is heavily based on user characteristics and engagement (Koumchatzky & Andryeyev, 2017). Additionally, we followed Carrignon et al. (2019) in using a fully-connected population for the ABM so that, under neutral conditions, users in our model were not more likely to encounter content from users with high follower counts. As in previous studies, this means that it is difficult to disentangle the effect of "influence" from pure availability, or the fact that a tweet coming from a user with many followers will simply have more exposure (Bakshy et al., 2011; Benkler et al., 2020). Even if we had access to the follower network of these users, the relevance of it would depend upon user behavior. If users primarily share information that they see passively on their timeline, then population structure is much more important, whereas if users are searching with keywords and hashtags, then it is less so. Since we do not have access to the follower network, and since many users interested in voter fraud conspiracy theories were probably actively searching for content related to those theories with keywords and hashtags, we believe that a fully-connected population is a reasonable simplifying assumption for this study. Based on the confounding effects of the algorithm and population structure, we are most confident in concluding that the differential spread of voter fraud claims among proponents of voter fraud conspiracy theories on Twitter during and after the 2020 US election was partly driven by a content bias causing users to preferentially retweet tweets with more negative emotion.

Our results are consistent with previous work suggesting that emotionally negative content has an advantage on social media across a variety of domains, including "fake news" articles, climate change coverage, and political events (Acerbi, 2019; Asatani et al., 2021; Bellovary et al., 2021; Del Vicario et al., 2016; Schöne et al., 2021). Other studies, though, have shown that positive messages spread more slowly but reach more people (Ferrara & Yang, 2015b), that exposure to both positive and negative tweets increases the probability of a user tweeting content with similar emotional valence (Ferrara & Yang, 2015a), and that tweets with greater emotional intensity (independent of valence) are more likely to be retweeted (Brady et al., 2017; Stieglitz & Dang-Xuan, 2013). In such cases, there may be variation across domains and individuals. Messages about same-sex marriage, for example, are more likely to be retweeted if they use positive language, whereas messages about climate change are more likely to be retweeted if they use negative language (Brady et al., 2017). We suspect that conspiracy theory content generally falls into the latter category. Similarly, Ferrara and Yang (2015a) found that there is variation in how users respond to emotional content, where some "highly susceptible" users are more likely to be influenced by positive messages. To improve the robustness of modeling of emotional contagion on social media, Burton et al. (2021) recently came up with three recommendations for future studies: going beyond correlational evidence, analyzing the effect of specification decisions on model estimates, and preregistration. We fully agree with these recommendations, and we hope that we adequately addressed them by using a

preregistered generative inference framework to ensure that the data was consistent with transmission bias before conducting GLMM and ensuring that our estimates were robust across a reasonable range of modeling specifications.

Regarding the spread of conspiracy theories, previous research has proposed "herd behavior", in which rational individuals with limited information defer to the beliefs of the majority, to be a potential explanation (Sunstein, 2014a, 2014b). Our study addresses the differential sharing of conspiracist tweets among proponents, who presumably already believed some voter fraud claims before the election took place, but our lack of clear evidence for a frequency bias suggests that a disproportionate tendency to "follow the herd" may not be the primary driver of the spread of conspiracy theory messages. Rather, our study suggests that the content of conspiracy theory messages and the characteristics of the individuals sharing those messages are more salient cues for cultural transmission. While recognized as important, the transmission processes involved in the spread of conspiracy theories have received relatively little attention in research and are not well understood (Bangerter et al., 2020). This study demonstrates the value of cultural evolutionary approaches for understanding these transmission processes, and it highlights the importance of considering the roles of both the content of conspiracy theories and the context in which they are shared. Identifying and characterizing the biases influencing the transmission of conspiracy theories can help us to generate potential methods for countering the spread of harmful conspiracy theories and promoting the spread of genuine information (see Salali & Uysal, 2021).

A previous study using generative inference to investigate behavior on Twitter found that retweet patterns of both confirmed and debunked information were more consistent with unbiased random copying than with conformity (Carrignon et al., 2019). At first glance our study seems to contradict this result, but Carrignon et al. (2019) did not include a parameter for demonstrator bias in their agent-based model, and they assumed neutrality for content bias due to computational limitations. When we ran our agent-based model with neutral values for both content bias and demonstrator bias, we too found that the model best fit the observed data when copying was unbiased by frequency (see Figure S1 and Table S2). The discrepancy between the results when parameters are estimated together instead of individually highlights the importance of considering equifinality—the fact that different processes can lead to similar patterns at the population level (Barrett, 2019). If different processes lead to only subtle differences in retweet frequencies, then the effect of one could be mistakenly attributed to another if both are not considered simultaneously. Luckily, in our study we found that the observed retweet distribution was consistent with content bias when content bias was estimated both alongside other parameters (Figure 1 and Table 1) and in isolation (Figure S1 and Table S2).

One of our biggest takeaways from this study is the importance of algorithmic transparency and accountability (Matei et al., 2015; Shah, 2018). Without access to detailed information about recommendation algorithms, researchers will continue to face difficulty in constructing realistic null models and making inferences about behavior on social media. Given that Twitter's timeline algorithm uses deep learning (Koumchatzky & Andryeyev, 2017) it is likely impossible for them to simply make the details of it public. Instead, the company could try to infer how the algorithm boosts different kinds of content by running natural experiments on the platform, as has been recently done for racial and gender bias in the image cropping algorithm (Agrawal & Davis, 2020) and right-leaning political bias in the timeline algorithm (Huszár et al., 2021). For example, Twitter could publish the results of a model in which simulated users randomly share information from simulated timelines constructed by the algorithm to see how different kinds of content spread under neutral conditions. Luckily, algorithmic transparency and accountability are increasingly being viewed as public policy priorities by governments around the world[2][3]. We hope that future studies can take advantage of improved transparency to develop more effective policy recommendations for fighting the spread of conspiracy theories and disinformation on social media platforms.

In conclusion, our methodology, based on a cultural evolution framework, allowed us to weigh the relative importance of different features influencing the spread of voter fraud claims among conspiracy theorists on Twitter. Most importantly, we found that retweet frequencies of voter fraud messages posted during and after the 2020 US election are consistent with a demonstrator bias for users with many followers and a content bias for tweets with more negative emotion. We are most confident with the latter finding given the confounding effects of Twitter's proprietary timeline algorithm and the platform's population structure. While previous research focused a priori on the role of tweets' content, we were able to show that content is indeed central when compared with other possible mechanisms of social influence. The methods presented here can be easily applied to other datasets and even expanded with a wider range of possible biases depending on the model in question.

Methods

The data for this study comes from the *VoterFraud2020* dataset, collected between October 23 and December 16 of 2020 by Abilov et al. (2021). This dataset includes 7.6 million tweets and 25.6 million retweets that were collected in real time using Twitter's streaming API. The *VoterFraud2020* dataset was collected according to Twitter's Terms of Service and is

[2] https://www.congress.gov/bill/117th-congress/senate-bill/1896/text
[3] https://op.europa.eu/en/publication-detail/-/publication/8ed84cfe-8e62-11e9-9369-01aa75ed71a1

consistent with established academic guidelines for ethical social media data use (Abilov et al., 2021). Abilov et al. (2021) started out with a set of keywords and hashtags that co-occurred with "voter fraud" and #voterfraud between July 21 and October 22, 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 voter fraud conspiracy theories. We ran 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. We restricted our analysis to cluster #2 so that retweets would be indicative of the spread of the conspiracy theories among proponents, as opposed to discourse and debate between both proponents and detractors.

The agent-based model (ABM) we used has elements from Carrignon et al. (2019), Lachlan et al. (2018), and Youngblood and Lahti (2021), and is available as an R package on GitHub[4]. 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%) were 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 probability of tweeting an original tweet (μ) as opposed to retweeting an existing tweet (1 - μ), 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. 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. Lastly, we should note that we chose to exclude "top n" dynamics (e.g. trending topics) from our ABM, because we think they are unlikely to influence the spread of more fringe topics within a single community and they did not improve the fit of neutral models of cultural transmission on Twitter in Carrignon et al. (2019).

In summary, the following are the dynamic parameters in this ABM that we estimated 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 were assigned static values based on the real dataset. The output of this ABM is a distribution of retweet frequencies (see Figure 2), which was 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 used 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).

The same summary statistics were calculated from the observed retweet distribution of the real dataset. For purposes of the summary statistic calculations quote tweets were treated like original tweets, as they themselves can be retweeted. Then, the random forest version of ABC (Raynal et al., 2019) was conducted with the following steps:

[4] https://github.com/masonyoungblood/TwitterABM

- 200,000 iterations of the ABM were run to generate simulated summary statistics for different values of the parameters: c, a, d, and g.
- The output of these simulations were 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 was constructed for each of the four parameters using bootstrap samples from the reference table. Two summary statistics were randomly sampled for each split in the decision trees, and the optimal minimum node size yielding the lowest prediction error for each parameter was set using the *tuneRanger* package in R (Probst et al., 2019) (see Table S1).
- Each trained forest was provided with the observed summary statistics, and each regression tree was used to predict the parameter values that likely generated the data.

Uniform prior distributions were used for all four of the dynamic parameters: c = {0–8}, a = {0–2}, d = {0–8}, g = {0–8}. We plotted the first two principal components of the output from 10,000 iterations to ensure that we were capturing enough of the parameter space before running the full analysis (see Figure S3). We ran four additional rounds of the ABM, each with only a single term from the probability function included, to investigate the behavior of each parameter in isolation (see Figure S1, Figure S2, and Table S2). We also conducted posterior checks by running the agent-based model with parameter values drawn from the posterior distributions to see how closely the output matched the original data (see Figure S3 and Figure S4).

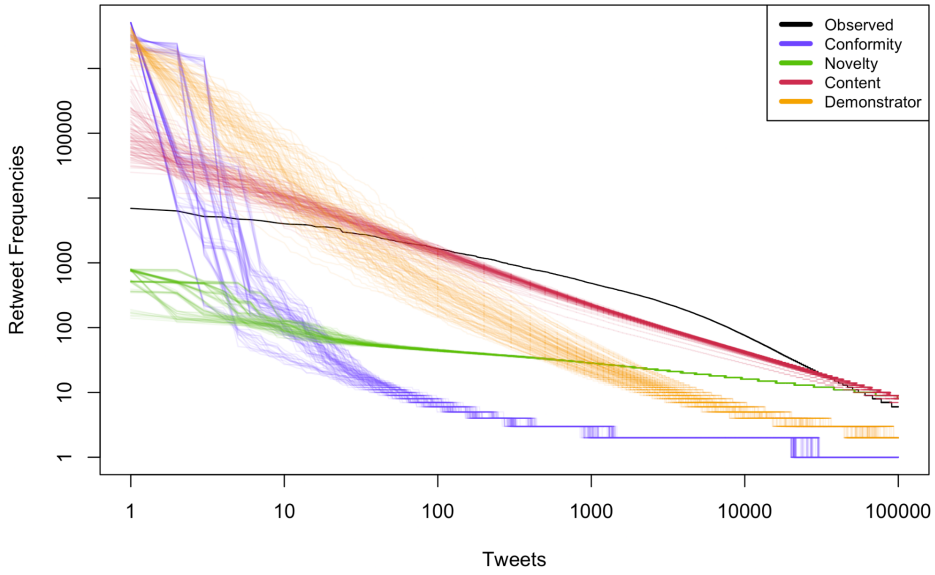

Figure 2. The retweet distributions resulting from conformity, novelty, content, and demonstrator bias using this ABM (100 iterations each), alongside the observed retweet distribution (in black). Biases were all modelled with a g of 0.25 and the following parameter values: a = 1.4 (conformity), a = 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.

Sentiment analysis was conducted using VADER from the natural language toolkit in Python, a model that performs similarly to human raters when applied to social media posts from platforms like Twitter (Hutto & Gilbert, 2014). VADER assigns a valence score to each word (and emoji or emoticon) in a tweet, and weights those scores according to a set of rules (e.g. negation, capitalization, punctuation). The main output of VADER is a compound score that sums and normalizes the weighted valences of the words in a tweet to give an overall score that indicates both the direction and the strength of emotion between -1 (strongly negative) and +1 (strongly positive). VADER also outputs the proportion of words in a tweet that are identified as neutral, positive, or negative. VADER was specifically trained to handle URLs, hashtags, and tagged users during sentiment analysis so we did not remove those from our dataset. Up-to-date details about VADER can be found in the GitHub repository[5].

To determine the potential targets of content and demonstrator biases we conducted GLMM using the *lme4* package in R (Bates et al., 2015). Retweet frequency was used as the outcome variable. To determine which grouping variables would be suitable as random effects we ran separate null models with each and calculated the intraclass correlation coefficient (ICC), or the proportion of the variance in retweet frequency explained by the grouping levels of

[5] https://github.com/cjhutto/vaderSentiment

each variable. Once random effects were chosen we added predictor variables in three stages, using the Akaike information criterion (AIC) and likelihood-ratio test (LRT) to choose between competing models. First, we determined whether tweet length would be an appropriate control variable. Then, we added follower count and verification status to see which measure of demonstrator attractiveness best improves the model. The 1.9% of tweets from users with missing verification statuses and follower counts were assigned verification statuses of "false" and follower counts of 0. Finally, we added the compound score, the proportion of negative words, and the proportion of positive words to see which measure of content best improves the model. All predictor variables were scaled and centered prior to analysis. Model choice and residual diagnostic tests were conducted using a random 10% of observations, but the best fitting model was run using the entire dataset. The Poisson family was used since our outcome variable was count data and did not appear to have over- or underdispersion issues.

To ensure that our decision to treat quote tweets like original tweets did not bias our results related to content, we did a second round of GLMM to determine whether quote tweets have different emotional content than the original tweets that they are quoting. We refer to original tweets that are quoted as target tweets (i.e. targets), and the tweets that quote them as quote tweets (i.e. quotes). Here we only considered target and quote tweets from cluster #2. Whether a tweet was a target (0) or a quote (1) was used as the outcome variable, and the identity of each target tweet was used as a random effect. In other words, each target and all of its quotes were assigned the same random effect. Like above, we first added tweet length as a control variable (for both targets and quotes). Then, we added the compound score, the absolute value of the compound score, the proportion of negative words, the proportion of positive words, and the proportion of neutral words as predictor variables to see which measure of content best improves the model. The absolute value of the compound score and the proportion of neutral words were included as indicators of a general reduction in the intensity of emotion independent of positive or negative valence. All predictor variables were scaled and centered prior to analysis, and model choice and residual diagnostic tests were conducted using all observations.

Preregistration

Our complete methods, model, and predictions were preregistered in advance of data analysis (https://osf.io/jnvyf/), except for the post hoc comparison between tweets and quote tweets.

Data & Code Availability

The agent-based model, analysis code, and processed data used in this study can be found on GitHub: https://github.com/masonyoungblood/TwitterABM. The full anonymized *VoterFraud2020* dataset can be found on Abilov et al.'s website (https://voterfraud2020.io/), and the full disambiguated dataset with tweet text is available from Abilov et al. upon request.

Author Contribution Statement

M.Y. developed the agent-based model and conducted all statistical modeling. R.G. conducted the sentiment analysis. A.A. and O.M. provided feedback on the agent-based model and statistical modeling. M.Y., J.S., A.A., and O.M. contributed to writing the manuscript.

Acknowledgments

We would like to thank Anne Kandler for providing us with feedback on our agent-based modeling. This research was supported, in part, under National Science Foundation Grants CNS-0958379, CNS-0855217, ACI-1126113 and the City University of New York High Performance Computing Center at the College of Staten Island. This work has received funding from the "Frontiers in Cognition" EUR grant, ANR-17-EURE-0017 EUR.

References

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

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, *5*(1), 15. https://doi.org/10.1057/s41599-019-0224-y

Agrawal, P., & Davis, D. (2020). *Transparency around image cropping and changes to come*. Twitter's Product Blog. https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping

Albertson, B., & Guiler, K. (2020). Conspiracy theories, election rigging, and support for democratic norms. *Research & Politics*, *7*(3), 2053168020959859. https://doi.org/10.1177/2053168020959859

Asatani, K., Yamano, H., Sakaki, T., & Sakata, I. (2021). Dense and influential core promotion of daily viral information spread in political echo chambers. *Scientific Reports*, *11*(1), 7491. https://doi.org/10.1038/s41598-021-86750-w

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 65–74. https://doi.org/10.1145/1935826.1935845

Balsamo, M. (2020, December). Disputing Trump, Barr says no widespread election fraud. *Associated Press*. https://apnews.com/article/barr-no-widespread-election-fraud-b1f1488796c9a98c4b1a9061a6c7f49d

Bangerter, A., Wagner-Egger, P., & Delouvée, S. (2020). How conspiracy theories spread. In M. Butter & P. Knight (Eds.), *Routledge Handbook of Conspiracy Theories* (pp. 206–218). Routledge.

Barrett, B. J. (2019). Equifinality in empirical studies of cultural transmission. *Behavioural Processes*, *161*, 129–138. https://doi.org/10.1016/j.beproc.2018.01.011

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1 SE-Articles), 1–48. https://doi.org/10.18637/jss.v067.i01

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
Beckett, L. (2021, April). Millions of Americans think the election was stolen. How worried should we be about more violence? *The Guardian*.
Bellovary, A. K., Young, N. A., & Goldenberg, A. (2021). Left- and right-leaning news organizations use negative emotional content and elicit user engagement similarly. *Affective Science*. https://doi.org/10.1007/s42761-021-00046-w
Benkler, Y., Tilton, C., Etling, B., Roberts, H., Clark, J., Faris, R., Kaiser, J., & Schmitt, C. (2020). Mail-in voter fraud: anatomy of a disinformation campaign. *Berkman Center Research Publication No. 2020-6*. https://doi.org/10.2139/ssrn.3703701
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 LP – 7318. https://doi.org/10.1073/pnas.1618923114
Brotherton, R. (2015). *Suspicious Minds: Why We Believe Conspiracy Theories*. Bloomsbury Sigma.
Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01133-5
Butter, M., & Knight, P. (2020). General introduction. In M. Butter & P. Knight (Eds.), *Routledge Handbook of Conspiracy Theories* (1st ed., pp. 1–8). Routledge. https://doi.org/10.4324/9780429452734-0
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., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45–67. https://doi.org/https://doi.org/10.1890/13-0133.1
Cillizza, C. (2021, June). 1 in 3 Americans believe the "Big Lie." *CNN*.
Cohen, L. (2021, January). 6 conspiracy theories about the 2020 election – debunked. *CBS News*.
Conger, K., & Isaac, M. (2021, January). Twitter permanently bans Trump, capping online revolt. *New York Times*.
Cooper, J. J., & Christie, B. (2021, April). Election conspiracies live on with audit by Arizona GOP. *Associated Press*. https://bit.ly/3k4soQD
Corasaniti, N., & Epstein, R. J. (2021, April). What Georgia's voting law really does. *New York Times*.
Cottrell, D., Herron, M. C., & Westwood, S. J. (2018). An exploration of Donald Trump's allegations of massive voter fraud in the 2016 General Election. *Electoral Studies*, *51*, 123–142. https://doi.org/10.1016/j.electstud.2017.09.002
Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: emotional contagion and group polarization on Facebook. *Scientific Reports*, *6*(1), 37825. https://doi.org/10.1038/srep37825
Dentith, M. (2014). *The Philosophy of Conspiracy Theories*. Palgrave Macmillan. https://doi.org/10.1057/9781137363169
Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*(S1), 3–35. https://doi.org/https://doi.org/10.1111/pops.12568
Edelson, J., Alduncin, A., Krewson, C., Sieja, J. A., & Uscinski, J. E. (2017). The effect of conspiratorial thinking and motivated reasoning on belief in election fraud. *Political Research Quarterly*, *70*(4), 933–946. https://doi.org/10.1177/1065912917721061
Enders, A. M., Uscinski, J. E., Klofstad, C. A., Premaratne, K., Seelig, M. I., Wuchty, S., Murthi, M. N., & Funchion, J. R. (2021). The 2020 presidential election and beliefs about fraud: continuity or change? *Electoral Studies*, *72*, 102366. https://doi.org/https://doi.org/10.1016/j.electstud.2021.102366
Ferrara, E., & Yang, Z. (2015a). Measuring emotional contagion in social media. *PLoS ONE*, *10*(11), 1–14. https://doi.org/10.1371/journal.pone.0142390
Ferrara, E., & Yang, Z. (2015b). Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, *1*(e26). https://doi.org/10.7717/peerj-cs.26
Goertzel, T. (1994). Belief in conspiracy theories. *Political Psychology*, *15*(4), 731–742. https://doi.org/10.2307/3791630
Hall Jamieson, K., & Albarracín, D. (2020). The telation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US. *Harvard Kennedy School Misinformation Review*, *1*. https://doi.org/10.37016/mr-2020-012
Hartig, F. (2020). *DHARMa: residual diagnostics for hierarchical regression models* (R package version 0.3.3.0). https://cran.r-project.org/package=DHARMa
Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *12*(3), 123–135. https://doi.org/https://doi.org/10.1002/evan.10110
Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2021). Algorithmic amplification of politics on Twitter. *Twitter*. https://bit.ly/3niLy6V
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.
Imhoff, R., Dieterle, L., & Lamberty, P. (2020). Resolving the puzzle of conspiracy worldview and political activism: belief in secret plots decreases normative but increases nonnormative political engagement. *Social Psychological and Personality Science*, *12*(1), 71–79. https://doi.org/10.1177/1948550619896491
Jiménez, Á. V, & Mesoudi, A. (2019). Prestige-biased social learning: current evidence and outstanding questions. *Palgrave Communications*, *5*(1), 20. https://doi.org/10.1057/s41599-019-0228-7
Jolley, D., & Douglas, K. M. (2014). The social consequences of conspiracism: exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint. *British Journal of Psychology*, *105*(1), 35–56. https://doi.org/https://doi.org/10.1111/bjop.12018
Jolley, D., Meleady, R., & Douglas, K. M. (2020). Exposure to intergroup conspiracy theories promotes prejudice which spreads across groups. *British Journal of Psychology*, *111*(1), 17–35. https://doi.org/https://doi.org/10.1111/bjop.12385
Kandler, A., & Powell, A. (2018). Generative inference for cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1743). https://doi.org/10.1098/rstb.2017.0056
Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy*, *96*(3), 109–126. https://doi.org/10.2307/2564659
Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: bridge-building between fields. *Trends in Cognitive Sciences*, *22*(7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003
Kofta, M., Soral, W., & Bilewicz, M. (2020). What breeds conspiracy antisemitism? The role of political uncontrollability and uncertainty in the belief in Jewish conspiracy. *Journal of Personality and Social Psychology*, *118*(5), 900–918. https://doi.org/10.1037/pspa0000183
Koumchatzky, N., & Andryeyev, A. (2017). *Using deep learning at scale in Twitter's timelines*. Twitter's Engineering Blog.
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
Matei, S. A., Russell, M. G., & Bertino, E. (2015). *Transparency in social media: tools, methods and algorithms for mediating online interactions* (S. A. Matei, M. G. Russell, & E. Bertino (eds.)). Springer. https://doi.org/10.1007/978-3-319-18552-1
Mesoudi, A. (2011). *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. University of Chicago Press. https://doi.org/10.7208/9780226520452

Min, S. J. (2021). Who believes in conspiracy theories? Network diversity, political discussion, and conservative conspiracy theories on social media. *American Politics Research*, *49*(5), 415–427. https://doi.org/10.1177/1532673X211013526
Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213
Olmsted, K. (2018). Conspiracy theories in US history. In J. E. Uscinski (Ed.), *Conspiracy Theories and the People Who Believe Them* (pp. 285–297). Oxford University Press. https://doi.org/10.1093/oso/9780190844073.001.0001
Osmundsen, M., Bor, A., Vahstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*(3), 999–1015. https://doi.org/DOI: 10.1017/S0003055421000290
Pagán, V. E. (2020). Conspiracy theories in the Roman Empire. In M. Butter & P. Knight (Eds.), *Routledge Handbook of Conspiracy Theories* (1st ed.). Routledge. https://doi.org/10.4324/9780429452734-5_1
Pigden, C. (1995). Popper revisited, or what Is wrong with conspiracy theories? *Philosophy of the Social Sciences*, *25*(1), 3–34. https://doi.org/10.1177/004839319502500101
Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery* , *9*(3), e1301. https://doi.org/https://doi.org/10.1002/widm.1301
Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118
Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, *35*(10), 1720–1728. https://doi.org/10.1093/bioinformatics/bty867
Romm, T., & Dwoskin, E. (2021, January). Twitter purged more than 70,000 affiliated with QAnon following Capitol riot. *Washington Post*. https://www.washingtonpost.com/technology/2021/01/11/trump-twitter-ban/
Roswell, M., Dushoff, J., & Winfree, R. (2021). A conceptual guide to measuring species diversity. *Oikos*, *130*(3), 321–338. https://doi.org/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
Salali, G. D., & Uysal, M. S. (2021). Effective incentives for increasing COVID-19 vaccine uptake. *Psychological Medicine*, 1–3. https://doi.org/10.1017/S0033291721004013
Sardarizadeh, S., & Lussenhop, J. (2021, January). The 65 days that led to chaos at the Capitol. *BBC News*. https://www.bbc.com/news/world-us-canada-55592332
Schöne, J. P., Parkinson, B., & Goldenberg, A. (2021). Negativity spreads more than positivity on Twitter after both positive and negative political situations. *Affective Science*. https://doi.org/10.1007/s42761-021-00057-7
Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A*, *376*(2128), 20170362. https://doi.org/10.1098/rsta.2017.0362
Skelley, G. (2021, May). Most Republicans still won't accept that Biden won. *FiveThirtyEight*. https://fivethirtyeight.com/features/most-republicans-still-wont-accept-that-biden-won/
Spring, M. (2020, November). "Stop the steal": The deep roots of Trump's "voter fraud" strategy. *BBC New*.
Stempel, C., Hargrove, T., & Stempel, G. H. (2007). Media use, social structure, and belief in 9/11 conspiracy theories. *Journalism & Mass Communication Quarterly*, *84*(2), 353–372. https://doi.org/10.1177/107769900708400210
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
Sunstein, C. R. (2014a). *Conspiracy theories and other dangerous ideas*. Simon & Schuster.
Sunstein, C. R. (2014b). *On rumors: how falsehoods spread, why we believe them, and what can be done*. Princeton University Press.
Tucker, E., & Bajak, F. (2020, November). Repudiating Trump, officials say election "most secure." *Associated Press*.
Uscinski, J. E., & Parent, J. M. (2014). *American Conspiracy Theories*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199351800.001.0001
van Prooijen, J.-W., Ligthart, J., Rosema, S., & Xu, Y. (2021). The entertainment value of conspiracy theories. *British Journal of Psychology*, *n/a*(n/a). https://doi.org/https://doi.org/10.1111/bjop.12522
West, H. G., & Sanders, T. (2003). *Transparency and Conspiracy: Ethnographies of Suspicion in the New World Order*. Duke University Press.
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
Zwierlein, C. (2020). Conspiracy theories in the middle ages and the early modern period. In M. Butter & P. Knight (Eds.), *Routledge Handbook of Conspiracy Theories* (1st ed., pp. 542–554). Routledge. https://doi.org/10.4324/9780429452734-5_2

Supplementary Information

	Minimum node size	MSE
c	61	3.18
d	150	1.10
a	39	0.055
g	47	2.15

Table S1. The optimal minimum node size for the random forests for each of the four dynamic parameters in the ABM.

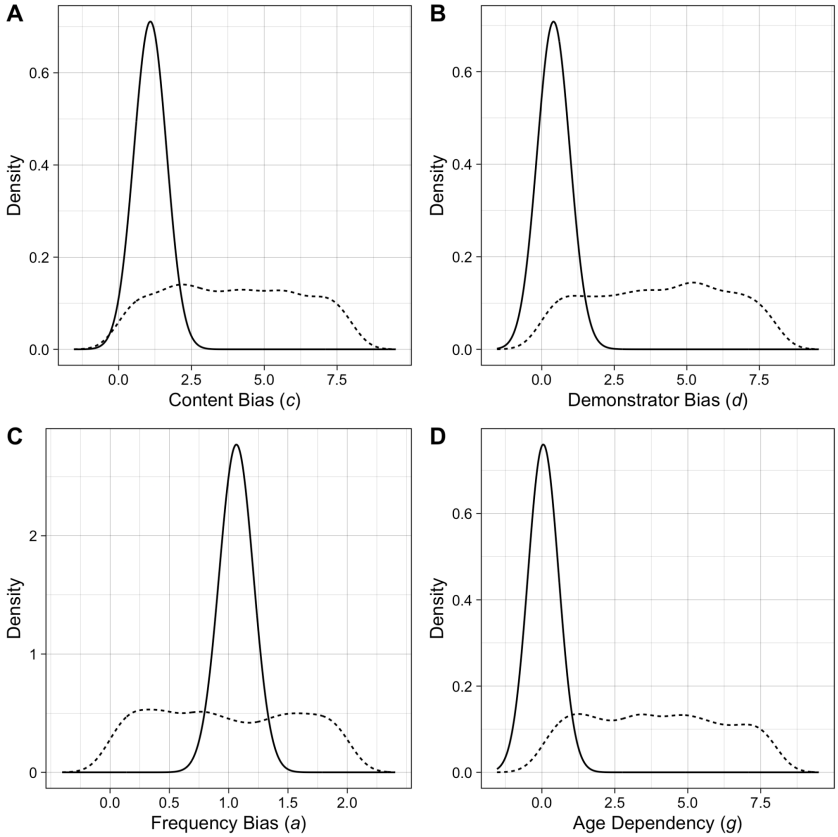

Figure S1. The prior (dotted lines) and posterior (solid lines) distributions from four additional rounds of the ABM. Each round was run for 1,000 iterations with only a single term from the probability function included. For example, panel A shows the posterior distribution for *c* when the ABM was run with uniform prior for *c* (from 0 to 8) with fixed neutral values for *d* (0), *a* (1), and *g* (0). In other words, panel A shows the estimated parameter values for *c* that best recreate the real data when all other parameters are ignored. Random forest ABC was conducted exactly as described in the main text.

	M	95% CI	NMAE
c	1.09	[0.64, 1.48]	0.029
d	0.43	[0.048, 0.85]	1.06
a	1.07	[0.95, 1.13]	0.0068
g	0.011	[0.011, 0.40]	2.44

Table S2. The median, 95% credible interval, and out-of-bag normalized mean absolute error of the posterior distributions from the four additional rounds of the ABM.

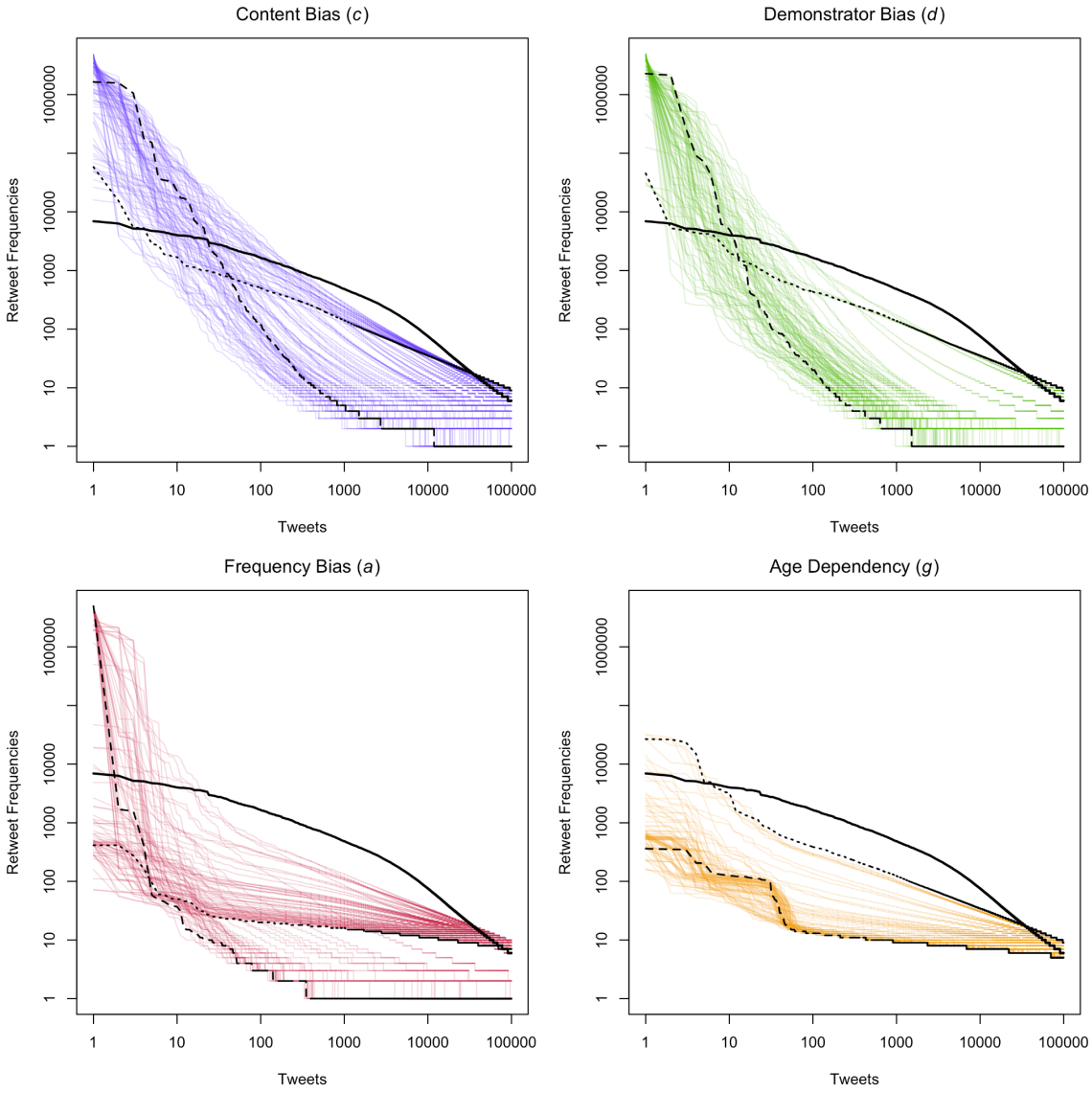

Figure S2. The retweet distributions from four additional rounds of the ABM (100 iterations each), alongside the observed retweet distribution (in solid black). The dotted black lines show the distribution resulting from the lowest value of that parameter sampled from the prior, whereas the dashed black lines show the distribution resulting from the highest value of that parameter sampled from the prior. Most importantly, frequency bias and age dependency (arguably the two processes most likely to confound one another) lead to different patterns in the retweet distributions.

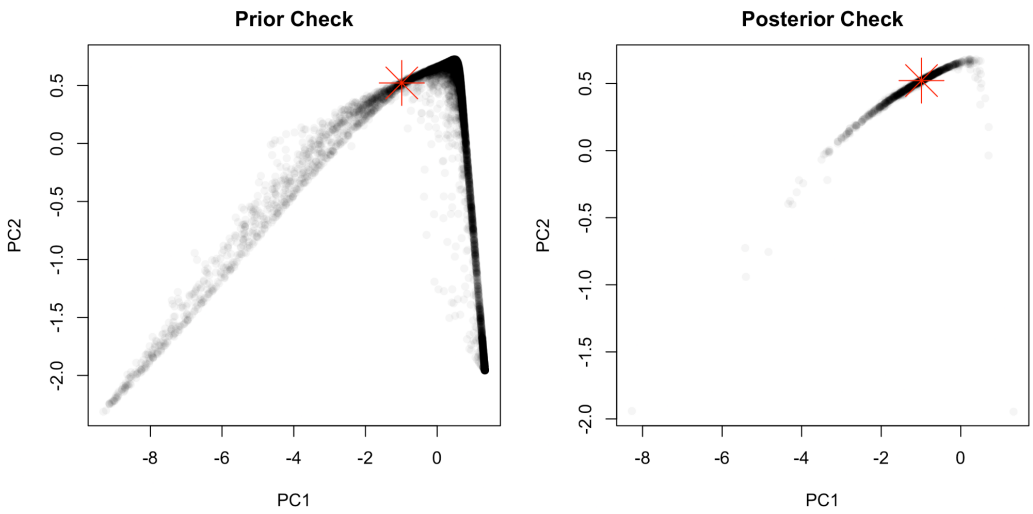

Figure S3. On the left, the first two principal components from 10,000 iterations of the model using the specified priors. We appear to be capturing more than enough of the parameter space to make inferences about our observed data, marked by the red star. On the right, the first two principal components from 1,000 iterations of the model with parameter values sampled from the posteriors. Importantly, the posterior distributions appear to have converged towards parameter values that do a much better job of recreating the observed data.

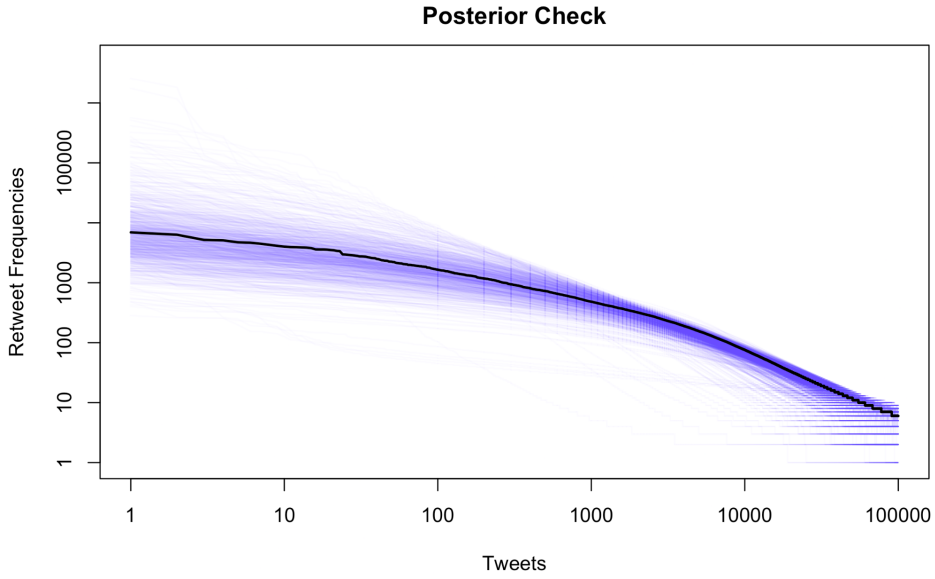

Figure S4. The retweet distributions resulting from 1,000 iterations of the model with parameter values drawn from the posteriors, alongside the observed retweet distribution (in black).

Model specification	df	AIC
retweets ~ (1 | user)	2	1121395
retweets ~ (1 | day)	2	3907893
retweets ~ (1 | hour)	2	3919423
retweets ~ scale(length) + (1 | user)	3	1080592
retweets ~ scale(followers) + scale(length) + (1 | user)	4	1074232
retweets ~ scale(verified) + scale(length) + (1 | user)	4	1080449
retweets ~ scale(compound) + scale(followers) + scale(length) + (1 | user)	5	1072285
retweets ~ scale(negative) + scale(followers) + scale(length) + (1 | user)	5	1073100
retweets ~ scale(positive) + scale(followers) + scale(length) + (1 | user)	5	1073560

Table S3. The model specification, degrees of freedom, and AIC for each candidate model for the primary GLMM (conducted with a random 10% of observations).

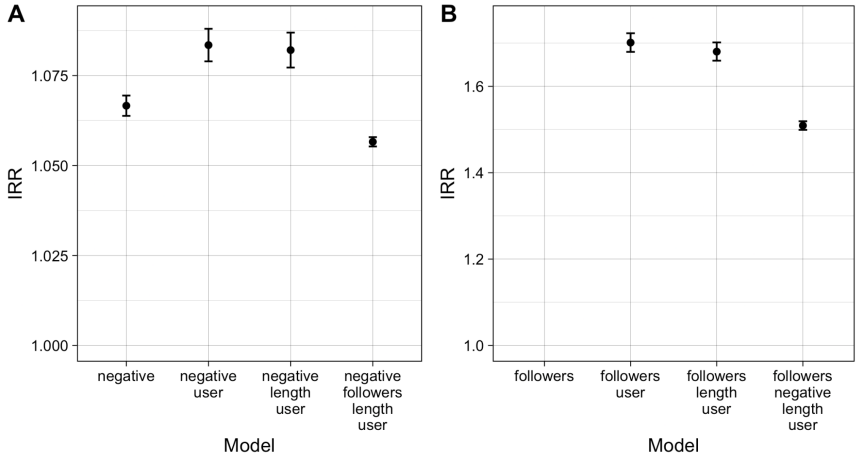

Figure S5. A partial specification curve analysis for the primary GLMM (conducted with a random 10% of observations). Panel A (on the left) shows the IRRs for the effect of the proportion of negative words on retweet probability under different combinations of predictors that appeared in the best fitting model. Panel B (on the right) shows the same for the effect of followers. Error bars represent Wald confidence intervals which were used due to high sample size. The IRR for followers from the simplest model in B is absent because the model did not converge. In all specifications, the proportion of negative words and the number of followers have significant effects in the same direction as in the best fitting model.

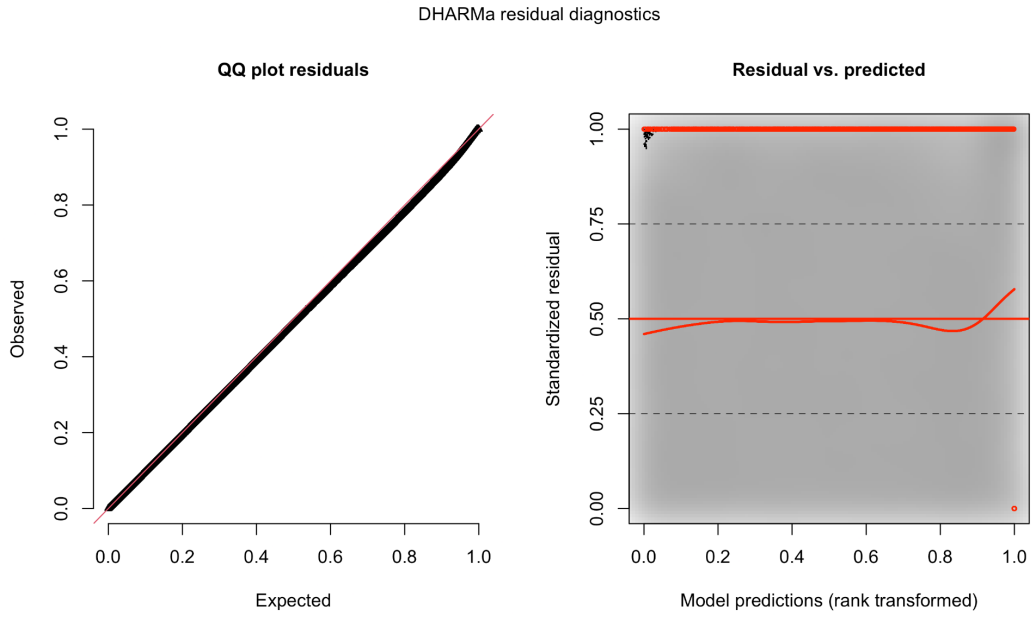

Figure S6. A Q-Q plot (left) and a standardized residual plot (right) for the main GLMM, constructed using the *DHARMa* package in R (Hartig, 2020). The model appears to be a good fit, and a dispersion test indicates that there is no significant dispersion in the data ($p = 0.73$). That being said, there is a low level of significant outliers among residuals (outliers = 0.97%, $p < 0.0001$), and a Kolmogorov-Smirnov test indicates that the residuals do significantly deviate from uniformity ($p < 0.0001$). Based on a visual inspection of the Q-Q plot, and the fact that the creator of the *DHARMa* package has suggested that slight departures from uniformity can be significant when sample sizes are extremely high[6], we conclude that the model is a good fit to the data. There is also a significant level of zero-inflation in the data ($p < 0.0001$), but the level of it is so low (2.2%) that we would prefer to use the Poisson family than to separately model zero and non-zero values.

[6] https://github.com/florianhartig/DHARMa/issues/181

Quote Tweet Analysis

After further subsetting the tweets for the quote tweet analysis, we ended up with 91,227 original tweets and 383,778 quote tweets from 102,227 users. The identity of each target tweet was included as a random effect (ICC = 0.15), and adding tweet length as a control variable significantly improved model fit (ΔAIC = 140971; LRT: $\chi^2 = 140972$, $p < 0.0001$). All five content measures further improved model fit (ΔAIC > 2), but the compound score improved fit significantly better than all of the content measures (ΔAIC > 2). All model specifications and AIC values for the quote GLMM are in Table S4. The best fitting model for the quote tweet analysis included the identity of each quoting event as a random effect, and tweet length and the compound score as predictor variables.

	OR	95% CI
Tweet length	0.176	[0.173, 0.178]
Compound score	1.865	[1.845, 1.884]

Table S4. The odds ratio (OR) and 95% confidence interval for each predictor in the best fitting model. OR, the exponentiated beta estimate, is interpreted as the rate at which the outcome variable is expected to change per unit increase in a predictor (one standard deviation for scaled and centered predictors). Wald confidence intervals were used due to the high sample size.

Tweet length and the compound score both significantly predict whether a tweet is original or a quote. The odds ratio (OR) for tweet length is 0.176, indicating that if a tweet is one SD longer then it is 82.4% more likely to be an original tweet. In other words, quote tweets tend to be shorter than original tweets. The compound score, on the other hand, has a strong positive effect on the probability that a tweet is a quote. The OR of 1.865 indicates that tweets with a compound score that is one SD higher are 86.5% more likely to be a quote. In other words, quote tweets tend to be more positive than their targets that they are quoting. The next best fitting model, which included the proportion of words that are negative, outcompeted the models with the absolute value of the compound score and the proportion of words that are neutral (ΔAIC > 2). This suggests that the shift in the tone of quote tweets is driven by a reduction in negative emotion rather than a general reduction of all emotion independently of valence. Pseudo R^2 values, calculated using the theoretical variances, indicate that the predictor variables alone account for about 47% of the variance in the data ($R^2 = 0.474$), whereas the predictor variables and random effects together account for about 61% of the variance in the data ($R^2 = 0.607$) (Nakagawa et al., 2017). A variance inflation factor test indicates that there are no significant issues with multicollinearity between predictors (VIFs < 2). Residual diagnostics for the best fitting model can be seen in Figure S7.

Model specification	df	AIC
retweets ~ (1 | event)	2	420624
retweets ~ scale(length) + (1 | event)	3	279654
retweets ~ scale(compound) + scale(length) + (1 | event)	4	264546
retweets ~ scale(abs(compound)) + scale(length) + (1 | event)	4	273125
retweets ~ scale(negative) + scale(length) + (1 | event)	4	269184
retweets ~ scale(positive) + scale(length) + (1 | event)	4	277239
retweets ~ scale(neutral) + scale(length) + (1 | event)	4	277363

Table S5. The model specification, degrees of freedom, and AIC for each candidate model for the quote GLMM (conducted with all observations).

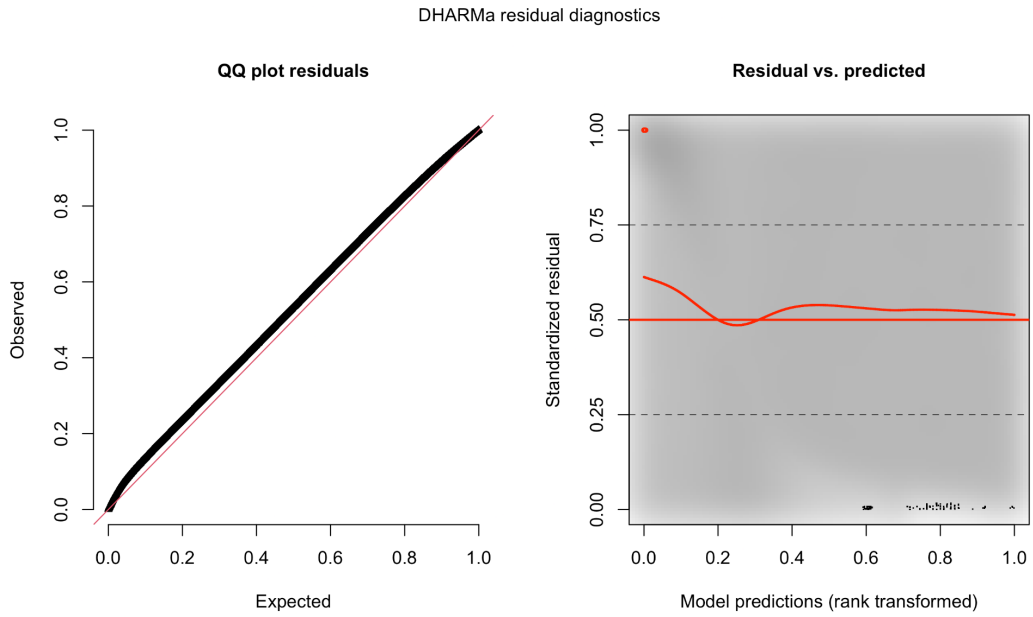

Figure S7. A Q-Q plot (left) and a standardized residual plot (right) for the quote analysis GLMM, constructed using the *DHARMa* package in R (Hartig, 2020). The model appears to be a good fit, but there is a low level of significant outliers among residuals (outliers = 0.77%, $p < 0.0001$), and a Kolmogorov-Smirnov test indicates that the residuals do significantly deviate from uniformity ($p < 0.0001$). Based on a visual inspection of the Q-Q plot, and the fact that the creator of the *DHARMa* package has suggested that slight departures from uniformity can be significant when sample sizes are extremely high[7], we conclude that the model is a good fit to the data. We did not conduct a dispersion test for the quote analysis GLMM, as they are unreliable for logistic models with binary outcomes[8].

[7] https://github.com/florianhartig/DHARMa/issues/181
[8] https://github.com/florianhartig/DHARMa/issues/79