Measuring and mitigating misinformation at the scale of the social media ecosystem

Christopher K. Tokita[1,*], Kevin Aslett[2], William P. Godel[2,3], Zeve Sanderson[2], Joshua A. Tucker[2,3], Jonathan Nagler[2,3], Nathaniel Persily[5], and Richard Bonneau[2, 4]

[1]Department of Ecology and Evolutionary Biology, Princeton University
[2]Center for Social Media and Politics, New York University
[3]Department of Politics, New York University
[4]Prescient Design, a Genentech accelerator
[5]Stanford University Law School, Stanford University
[*]email: christopher.tokita@gmail.com

May 23, 2023

Abstract

While online misinformation has become a growing topic of scholarly inquiry and public concern, we are still developing the science to study its impact and evaluate potential remedies at the scale of the social media ecosystem. Here, we pair experimental survey data with observational Twitter data to create a more precise estimate of the impact of trending true and false news. We estimate the exposure of receptive users—that is, the users most likely to believe an article's content as predicted by their individual characteristics. Using this new approach, we find that millions of receptive users were potentially exposed to top-trending true and false news articles on social media. Importantly, we also find that the pattern of receptive user exposure differs between false news and true news: both true news and false news are seen by an ideologically diverse set of users, yet receptive users that are exposed to false news are far more concentrated on the conservative extreme of the ideological spectrum. Thus, efforts to infer the impact of misinformation by measuring total user exposure may not accurately capture the true impact of misinformation on social media. We extend this new method and conduct data-driven simulation to evaluate different interventions that social media platforms deploy to reduce the effect of misinformation. We find that interventions that are not deployed instantaneously are unlikely to reduce the exposure of receptive users to misinformation: most receptive users are exposed shortly after a URL is first shared and thus the effectiveness of interventions quickly dissipates with each hour delay in implementation. Moreover, we find that interventions that reduce the visibility of misinformation (e.g., down-ranking) are more effective than those that attempt to reduce the re-sharing of misinformation (e.g., fact-check labels). Our paper provides the first full-scale estimation of the exposure of receptive users to online misinformation and, in doing so, provides much needed evidence for informing potential remedies for online misinformation at the scale of social media platforms.

1 Introduction

Exposure to news on social media, whether intentional [1] or incidental [2], has the power to influence beliefs, which can in turn shape perceptions of reality [3, 4], influence political agendas [5], and start social or political movements [6, 7]. While social media has been associated with increasing factual political knowledge [8, 9], online misinformation has threatened both democracy[1] and public health,[2], leading some to deem the presence of an "infodemic" alongside the Covid pandemic [14]. In response to this online media environment, researchers from across disciplines [15, 16] are studying the complex

[1] On January 6th, 2020 misinformation fueled riots at the U.S. capitol building that threatened the transfer of power after an election [10, 11].

[2] Waves of misinformation about Covid-19 have increased distrust of public health officials, weakened responses to the pandemic, and increased skepticism of the Covid-19 vaccine [12, 13].

interplay between the exposure to [17], belief in [18, 19], and sharing of [20–22] both true and false information online.

However, to our knowledge, recent work has yet to unify the measurement of the diffusion of (i.e. sharing and exposure) and the belief in misinformation. On the one hand, literature on misinformation diffusion uses large-scale observational data from social networks to analyze how the velocity and scale of misinformation spread is affected by features such as veracity [20], novelty [23], and sentiment [24, 25]. However, this literature fails to incorporate a measure of user belief in misinformation, a critical missing component given that users on social media are not equally likely to adopt false beliefs after exposure to misinformation that may change an individual's behavior and thereby impact collective actions in society, including democratic participation [26]. On the other hand, current work estimating belief in misinformation typically uses survey instruments to measure the relationships between receptivity and individual characteristics such as ideological congruence [27–30], age [18], cognitive reflection [19, 31], and digital literacy [32]. This literature identifies key individual-level covariates associated with belief, but it has not been paired with observational social media data to understand how this plays out at scale on networked platforms. While recent innovations have paired digital trace data with longitudinal survey data [21, 33], these data also do not enable evaluations of potential belief across the entirety of a social media ecosystem. As a result, we are left without an estimate of the scale of belief in misinformation, which in turn limits our understanding of the impacts of misinformation on social media.

Likewise, without the ability to measure potential belief in misinformation at scale, we are not able to fully assess the effectiveness of interventions aimed at reducing the impact of misinformation on social media users. Recently, social media platforms have employed various strategies to limit the spread of misinformation [34], including labeling questionable articles with fact-check labels [35, 36], making them more difficult to share [22, 37], or simply reducing their visibility on users news feeds [38]. Despite the rapid rise of platform-level interventions [39] and the subsequent debate over whether they are worthwhile [40], we lack a way to measure how these interventions might ultimately change misinformation exposure among the users most likely to believe it (i.e., users who are receptive to a particular piece of misinformation). Recent work has provided insight into how interventions could reduce the likelihood that users share misinformation [41], but these insights stop short of understanding how these interventions alter exposure among receptive users. The studies that do attempt to measure the effect of interventions on belief in misinformation [22, 42] focus exclusively at the level of the individual and do not measure overall ecosystem effects, where a share can potentially expose thousands of other users. Therefore, measuring potential belief at scale is key first step for assessing the full effectiveness of interventions.

Aiming to bridge recent work on misinformation, in this study, we create a robust large-scale estimate of the exposure of receptive users to misinformation on social media. Focusing on 139 real highly popular U.S. news articles (True $n = 102$, False/Misleading $n = 37$), we combine (a) large-scale Twitter data tracking the spread of these articles and (b) real-time surveys of ordinary Americans measuring how likely users are to believe these articles. Because we cannot directly measure user belief—that is, we cannot causally say that a particular user saw an article and thus changed their beliefs—we frame our estimate as the exposure of receptive users: the number of users who see an article and are likely to believe it as predicted by their individual characteristics, such as ideology. Using this new approach, we show that the exposure of receptive users to misinformation is distinct from the pattern observed among true news: receptive users exposed to misinformation are concentrated on the conservative extreme of the political spectrum while receptive users exposed to true news are ideologically balanced. Importantly, general user exposure to misinformation often does not predict exposure among receptive users in the same way as true news: users across the ideological spectrum are exposed to both true news and misinformation, yet users that are exposed and receptive to misinformation skew very conservative. Crucially, those who see both true and false news earliest after publication are also the users most likely to be receptive to it, highlighting the fact that it only takes a few hours to quickly be believed millions of users. Additionally, we use these data to conduct data-driven simulations of misinformation interventions, finding that various attempts have minimal effect on how many receptive users see misinformation unless the intervention is introduced within a few hours after misinformation is first shared on social media.

2 Results

Using a set of 139 articles collected with a transparent selection process [43], we measured their spread on Twitter and estimated the total number of users exposed and the number of receptive users exposed to each article. Our measurement of exposed receptive users combined two data: survey data that estimates the probability that individuals with different characteristics (e.g., ideology) believe each news article, and Twitter data that identifies users who were exposed and their user characteristics.

Within 48 hours of the publication of each news article, we sent surveys to a representative sample of Americans to identify demographic covariates that predicted belief in that news article. We identified 139 articles that were the most popular daily articles from a mix of mainstream and fringe news sources published between November 2019 and February 2020. To establish the veracity of the articles in real time, each article was assessed by six professional fact-checkers within 48 hours of publication. Each fact-checker rated an article as "true" or "false/misleading", and we then used the modal rating from the fact-checkers to label an article as "true" or "false/misleading" in our dataset (for details about article collection and fact-checking, see Materials and Methods section). In the end, fact-checkers rated 102 articles as "true" and 37 articles as "false/misleading".[3] To measure user receptivity to these articles in real time, we asked approximately 90 US respondents to evaluate each article within the same 48 hours period: each respondent was asked to provide demographic information and whether they believed the article to be "true", "false/misleading", or "could not determine"[4] These survey responses allowed us to measure which individual-level characteristics were associated with believing each article in our study; echoing previous research,[28, 29] we found that a significant predictor of belief was congruence between a respondent's ideology and the the ideological perspective of the article [43].

We then measured the spread of each article by collecting all tweets and Twitter users that shared the article URL up to one week after publication. We also collected the friend and follower lists of each user who shared the article to identify all users potentially exposed to these articles on Twitter. [5] Because we identified that ideological congruence is a key predictor of belief in an article [43], we needed to generate ideology scores for these Twitter users so that we could later estimate the number of receptive users—those likely to believe the article—among all users exposed. To this end, we scored the ideology of all Twitter users in our dataset—article sharers, their followers, and their friends—using the method from [44]. If we did not have enough information to directly calculate the ideology of an article sharer, we calculated their ideology as the mean of their friends' ideology. Finally, to fill in missing ideology scores among a sharer's followers, we used the followers with known ideology scores to first fit a normal distribution and then draw missing scores from that distribution.

Finally, we used our exposure estimates and user ideology scores to estimate the number of receptive users among those exposed. We did this by multiplying our exposure estimates—which includes a breakdown of how many users of each ideology saw the article—by the corresponding belief rates of those same ideological groups[6] in the surveys [7] Our simulation approach therefore makes a simplifying assumption that receptivity is only conditioned on ideology, since ideology is shown to be a major predictor of belief in news content [27–30] and is readily measurable at scale on Twitter [44].

We have three primary motivations for bringing together the measurement of exposure and receptivity to misinformation. First, we examine whether the general pattern of user exposure mirrors the pattern of exposure among receptive users. In many studies of misinformation, researchers measure the impact of misinformation as exposure to it, without accounting for user characteristics that predict how likely a user is to believe it. As a result, it is important to understand whether researchers can safely assume that general user exposure accurately captures the impact of misinformation. Second, we measure how quickly the exposure of receptive users accumulates once an article is published and is shared on social media, as this rate affects approaches to mitigating belief in misinformation. Third, using simulations, we extend our methods to assess the efficacy of common social media platform interventions at preventing the exposure of receptive users.

[3] 16 articles were removed from the study because they either received a rating of "could not determine" or because there was no modal rating among the fact-checkers. See the Materials and Methods section for details.

[4] Note that respondents only evaluated three news articles, so most articles were not evaluated by the same respondents. More information about the study design can be found in the Materials and Methods section.

[5] We write *potentially* exposed because Twitter does not release data on which tweets that could have appeared in a user's timeline (i.e., they were tweeted or retweeted by an account followed by the user) actually were viewed by the user. As is the convention in the literature [33], we refer to these potential exposures simply as "exposures" for the remainder of the manuscript.

[6] The seven ideological groups each respondent could fit into were: Extremely Conservative, Conservative, Slightly Conservative, Moderate, Slightly Liberal, Liberal, and Extremely Liberal

[7] For example, if 100 liberal and 50 conservative users were exposed to an article, and the the corresponding belief rates were 50% and 20% respectively—that is, 50% of liberal and 20% of conservative survey respondents stated that they believed the article to be "true"—then we can estimate that 50 liberal and 10 conservative users were likely to be receptive to the article.

2.1 The patterns of exposure and belief differ for fake news

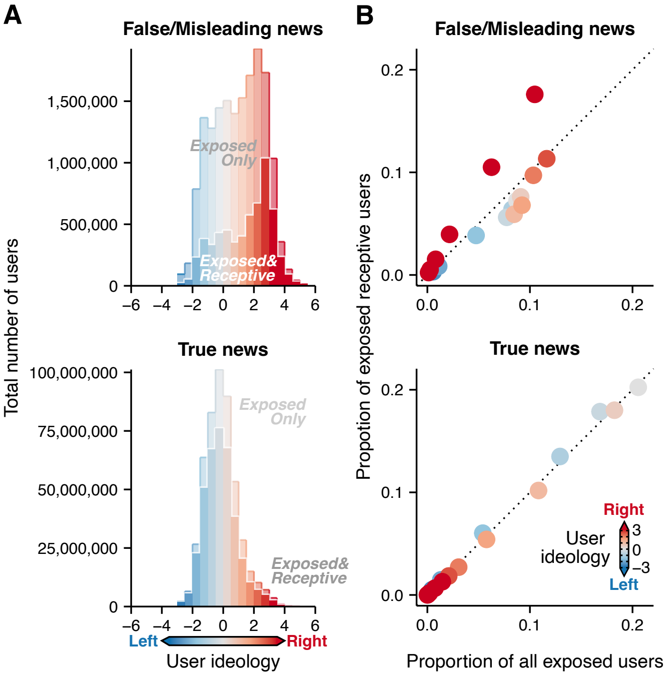

Figure 1: Estimated number of exposed and receptive to news on Twitter. (A) Estimates for the total number of users who could have potentially been exposed and receptive to the true and false/misleading news articles in our dataset. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology. (B) The relative abundance of a given ideology among all exposed users vs. all exposed receptive users. Points are the same user ideology bins found in Panel A. The dashed line represents a 1:1 ratio, such that points above the dotted line indicate that users of that ideology make up a disproportionate share of receptive users among those exposed to an article.

We find that tens of millions of unique Twitter users were potentially exposed to the articles in our data set (Figure 1). Of these articles, 37 news articles were rated as false/misleading by professional fact-checkers. These 37 articles, in turn, generated over 16.5 million potential instances of unique user exposure, resulting in over 5.8 million potential instances of exposure among receptive users. Concurrently, the 102 news articles that were rated as true by professional fact-checkers resulted in 492 million potential instances of unique user exposure and 375 million potential instances of exposure among receptive users. Thus, we estimate that the true news articles in our study were seen by nearly two orders of magnitude more receptive users on Twitter.

Using ideology estimates of Twitter users derived from their follower networks, we find striking differences between the pattern of exposure among all users and among receptive users. Relative to true news, false/misleading articles were shared by Twitter users with more extreme ideologies, in particular on the ideological right (Figure S1). Despite being shared by more extreme and right-leaning users, false/misleading articles had a cross-cutting pattern of exposure, wherein users from across the ideological spectrum likely saw the articles (Figure 1A, S2). However, unlike general user exposure, exposure among receptive users was not cross-cutting: most of the receptive users exposed to false/misleading articles were concentrated among a highly skewed and right-leaning subset of users (Figure 1A, S2). In fact, right-leaning users were over-represented among the population of receptive users exposed to false/misleading news (1B): right-leaning users made up 68.8% of all exposed users, but 75.7% of all exposed receptive users.

Unlike false/misleading news articles, true news articles generate a pattern of exposure among receptive users that mirrors the general pattern of user exposure. A more moderate user base shared true articles, which in turn exposed a more ideologically balanced audience. Importantly, users across the ideological spectrum were similarly receptive to true news articles, resulting in an ideologically balanced set of users that saw the articles and were receptive to their content.

Some of the mismatch between general user exposure and receptive user exposure misinformation may be explained by the partisan slant of the articles themselves. Among the false/misleading articles that we tracked, the articles with right-leaning slant gained the most traction and therefore drove most of the user exposure to misinformation. (Figure 2). Thus, while false/misleading news exposed a relatively cross-cutting audience, that audience was not equally receptive to it across the ideological spectrum: when we break out false/misleading articles by partisan slant, we see that right-leaning users were disproportionately likely to both see and be receptive to articles with right-leaning, while left-leaning users were most likely to see and be receptive to articles with left-leaning slant (Figure S5). This all stands in contrast to true articles: most user exposure were generated by articles that had a more neutral slant, and the pattern of exposure for receptive users generally matched the general pattern of user exposure across the ideological spectrum. For example, among true news articles, 42.5% of all exposed users were right-leaning and 40.4% of all exposed receptive users were right-leaning.

The structure of Twitter's social network may also play an important role in creating the mismatch between general user exposure and receptive user exposure to misinformation: center-right users have the most ideologically diverse followers (Figure S6) and consequently exposed the most diverse set of users to news articles (Figure S7A). Thus, center-right users generated a disproportionate amount of most cross-ideology exposure (Figure S7B), allowing moderate and left-leaning users to see articles that originated in right-leaning portions of the social network (see "Diffusion of news through social networks" in Supplemental Information). Conversely, left-leaning users have less diverse followers than center-right users, suggesting that misinformation originating in left-leaning portions of the social network has less potential to expose an ideologically diverse audience. Given that the most viral false/misleading articles that we tracked had right-leaning slant and that users tended to share articles that aligned with their ideology, center-right users might have played a key role in allowing the articles to spread out of right-leaning circles and expose a broader, albeit more skeptical, audience.

2.2 Exposure to misinformation among receptive users quickly accumulates over the first 48 hours

While we have characterized the ideological composition of the exposure of receptive users to true and false/misleading news, it is also important to understand how this exposure unfolds over time as an article is circulating through social media. Therefore, for each news article, we analyzed the exposure of receptive and non-receptive users over the first 48 hours after a news URL is first shared, given that this is the period in which most sharing occurs [20].

We find that the majority of news article exposure among receptive users happens within the six hours after an article URL is first shared on Twitter (Figure 3A). By hour six, the average true news article reaches 78.2% of its cumulative exposure among receptive user—that is, the total number of users who will see and be receptive to the article—with the average article crossing 50% within two hours. Similarly, the average false/misleading news articles reaches 60.8% of its cumulative exposure among receptive users within six hours, with the average article crossing 50% within three hours. (Figure 3B). When comparing the time it takes to reach 50% cumulative exposure among receptive users, we did not find a significant difference in the rate at which true and false/misleading news accumulate exposure of receptive users ($t(55.4) = -1.154, p = 0.25$. Overall, these patterns speak to the speed by which information—both legitimate and not—spreads and potentially impacts the beliefs of millions of users on social media.

We find that news articles have their highest rate of exposure among receptive users in the immediate hours after publication of an article. During the first hour after a news URL is first shared on Twitter, 76.4 out of every 100 users exposed to a the average true news article is also receptive to the article's contents, but 24 hours later, that drops to 65.8 out of every 100 newly exposed users (Figure 3C). While the average false/misleading news article starts out with a lower rate of exposure among receptive users at 42.1 out of every 100 users exposed also being receptive to it, we observe a similar decline 24 hours later at 36.6 per 100 newly exposed users (Figure 3C). It is important to note that we surveyed user belief in a given article only immediately after publication (see: Methods) and therefore we did not measure how baseline belief rates among users might shift as discussion of an article spreads. Instead, in our estimates, the decreasing frequency of receptive users over time is driven entirely by the shifting composition of newly exposed users: users with ideologies that are less inclined to believe an article tend to get exposed later.

In contrast to research finding that misinformation spreads faster than true information [20], we find that true news articles accumulate exposure among receptive users faster than false/misleading news. In the initial hours after an article URL is shared online, far more users see and believe true

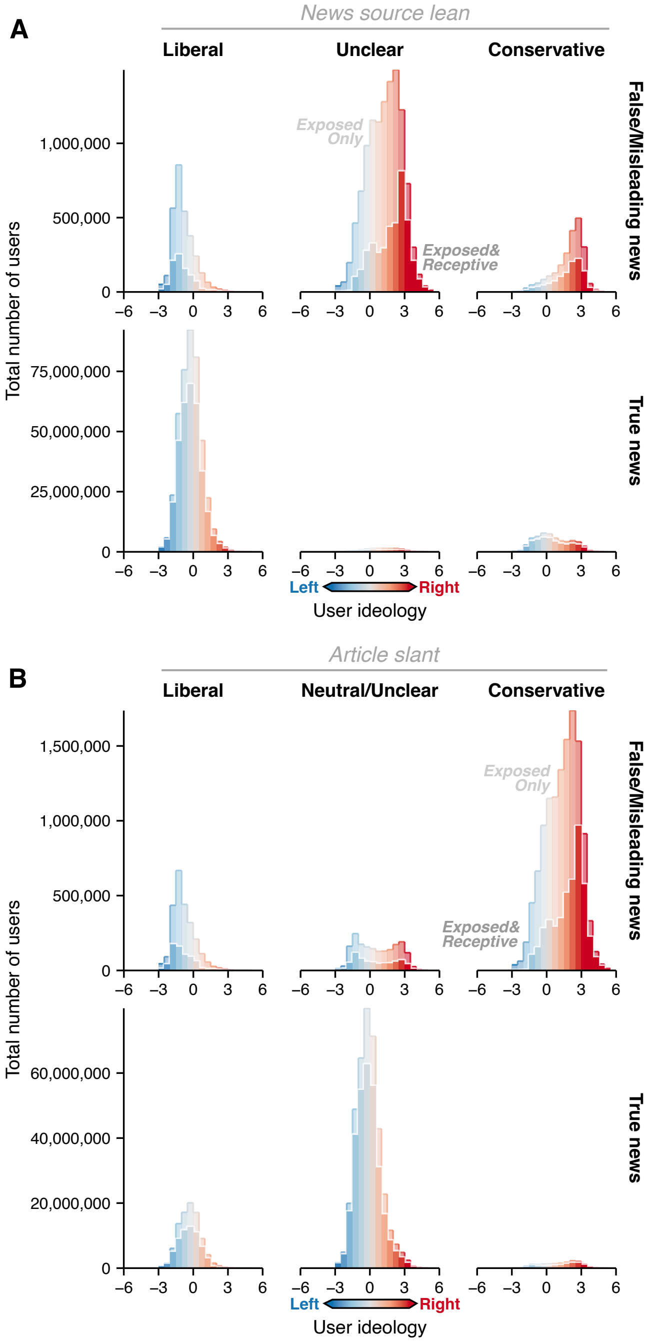

Figure 2: Total number of users exposed and receptive to articles, broken out by (A) news source lean and (B) article slant. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology.

news articles when compared against false/misleading news articles, which are initially seen by about an order of magnitude fewer receptive users (Figure 3A)—a difference that later builds to nearly two orders of magnitude more receptive users seeing true news articles over the course of a week (Figure 1A). This pattern may be partially driven by the news source itself: articles from fringe news sources are seen by fewer users (Figure S9) and are less likely to be believed by exposed users than articles published by mainstream sources (Figure S8). Thus, true news articles have a larger audience that is more likely to believe the article contents, allowing true news articles to generally accumulate exposure among receptive users faster than false/misleading articles. Overall, these patterns further highlight how the pattern of belief—as implied by our measurement of of exposure among receptive users—cannot be extrapolated from the pattern of general user exposure alone and must also account for individual and network dynamics, such as news source quality and exposed user ideology.

2.3 Common social-media platform interventions are largely ineffective at preventing receptive users from being exposed to misinformation

Social media platforms often deploy interventions in an attempt to stop the spread and impact of misinformation. To attempt to assess the impact of such interventions, we conducted data-driven simulations of common platform-level interventions to estimate how they might reduce misinformation exposure among receptive users at scale. We focused our simulations on the articles in our dataset that were labeled as false/misleading by professional fact-checkers, thus reflecting the platform policies that only act on articles that have been evaluated as false/misleading by external reviewers [45–48].

We examined three simple interventions commonly used by social media platforms (see Materials and Methods): sharing friction, fact-check labeling, and visibility reduction (e.g., downranking). We ran simulations in which we set individual-level effects of interventions and examined how it changed misinformation exposure among all users and among receptive users at the scale of the entire social media ecosystem. With the exception of visibility reduction, our estimates of the individual-level effects of interventions were based on results from other studies. For our simulations of sharing friction—adding extra steps to the retweet process for tweets sharing flagged material—we assumed that they made an individual 75% less likely to retweet a flagged article [22]. Similarly, for our simulations of fact-check

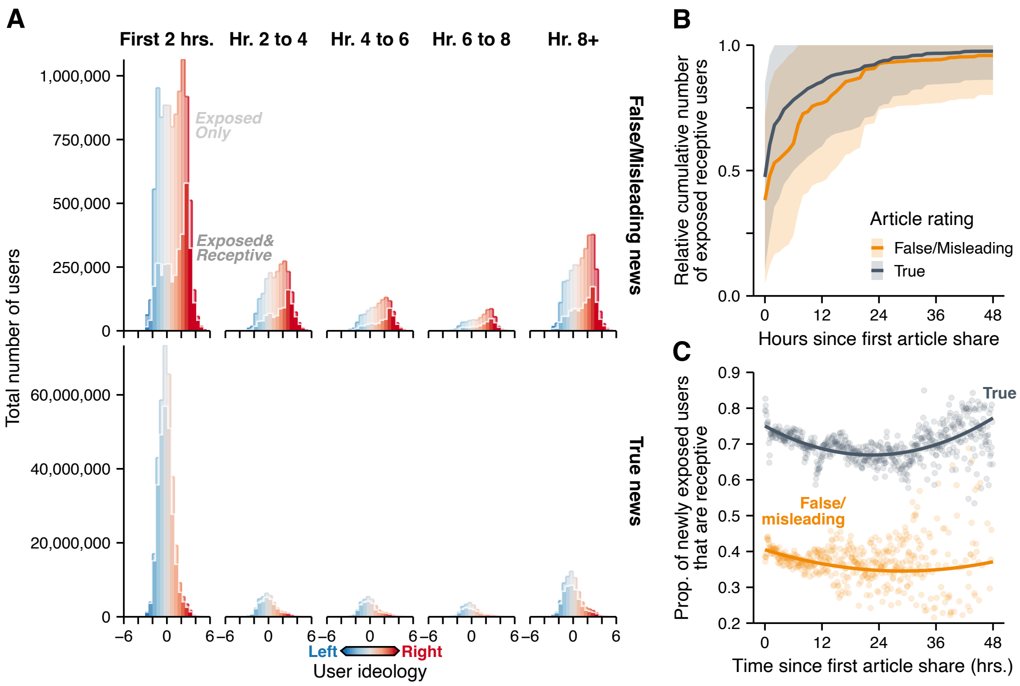

Figure 3: (A) The number of exposed and receptive users over the first 24 hours of article sharing. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology. (B) Cumulative number of receptive users (mean ± s.d.) exposed over time, normalized across articles. (C) Proportion of newly exposed users that are receptive. Points are the binned mean across all tweets within 5-minute intervals. The lines are the best-fit Bayesian regression.

labeling, we assumed that they made an individual 25% less likely to retweet [22] and 17% less likely to believe [42] a flagged article. Finally, for our simulations of visibility reduction—an intervention in which a tweet containing a flagged news article becomes less likely to appear in other user's timelines—we assumed a light and heavy version of the intervention: the interventions made misinformation tweets 25% (light) or 75% (heavy) less likely to appear in other user's timelines. The rates for visibility reduction interventions were not based on findings from studies, but were instead selected for ease of comparison with the effectiveness of the sharing-focused interventions.

For each intervention type, we also examined the importance of timing—specifically, how long it takes to verify an article before deploying an intervention. When considering using an intervention on a potential piece of misinformation, a social media platform will commonly use professional fact-checkers to first determine the veracity of the article [49]. While external fact-checkers helps ensure that interventions are only used on false or misleading content, it also comes with an opportunity cost: it takes time for fact-checkers to verify a news article, thereby allowing the piece of misinformation to circulate freely for some time. Therefore, our simulations estimated the effect of review time on intervention efficacy. We assumed that an intervention is deployed t_{int} hours after a piece of misinformation first appears on the platform, where t_{int} represents the review delay. We also simulated an ideal case where interventions can immediately be deployed on a piece of misinformation—i.e., $t_{int} = 0$—as might be possible with the use of artificial intelligence that can instantly flag questionable content.

Even when assuming that interventions can be deployed immediately, we find that simple interventions that decrease the sharing or visibility of false news articles have mixed success reducing exposure among all users (Figure S10) and exposure among receptive users (Figure 4). Attempts to slow the sharing of misinformation generally have a very modest effect on the overall number of receptive users exposed to false articles. Despite the assumption that adding sharing friction to misinformation tweets reduces the likelihood of individual retweeting by 75%, our simulations find that this intervention only reduces cumulative exposure among receptive users by an average of up to 16.5%. Similarly, we assumed that fact-checking labels reduce the likelihood of individual retweeting by 25% and the likelihood of individual belief by 17%, yet our simulations find that this intervention reduces cumulative exposure among receptive users by an average of up to 22.4%. In both cases, interventions that target individual re-sharing (i.e., retweeting) do not greatly reduce aggregate exposure among receptive users because article URLs are often first introduced by accounts with large follower counts. As a result, even if users are far less likely to retweet a piece of misinformation, many users still see the original tweets from these prominent accounts. On the other hand, compared to the previous sharing-focused interventions, we find that visibility-focused interventions—where Twitter makes flagged misinformation tweets less likely to appear in other users' timelines—are more effective at reducing the aggregate exposure of receptive users to misinformation. We assumed that visibility reduction decreased the likelihood that a user would see a misinformation tweet by 25% or 75%, and our simulations find that this reduced cumulative exposure among receptive users by an average of up to 21.7% and 68.2%, respectively. Unlike sharing-focused interventions that only affect retweets, visibility-interventions are comparatively more impactful because they decrease the likelihood that a user sees an article tweet, regardless of whether it is an original share of the article URL or a rewteet.

However, the efficacy of these fact-checker-backed interventions depends on how quickly they can be deployed. The longer it takes for professional fact-checkers to verify an article (or for a social media company to act after the fact check), the less an intervention can reduce misinformation exposure among receptive users. For example, if Twitter can get questionable articles fact-checked within one hour of their first appearance on Twitter, we estimate that a heavy visibility reduction intervention can reduce user exposure by an average of 60% or more—a number that could easily result in millions fewer users believing false or misleading news articles on social media. Conversely, if it takes over 10 hours to fact-check articles and deploy interventions, we estimate that heavy visibility reduction interventions can reduce exposure among receptive users by an average of 25% or less. Given the infrastructure and time needed to coordinate with and receive third-party fact-checkers, the timing of this intervention method is a significant consideration.

3 Discussion

Our study is the first to attempt to estimate ecosystem-level belief in top-trending news articles by tracking the exposure of users who are most receptive to an article's content. Our approach combines two common approaches to studying misinformation: analysis of social media data to assess the spread of an article at scale and survey-based studies to determine rates of individual-level belief. The method

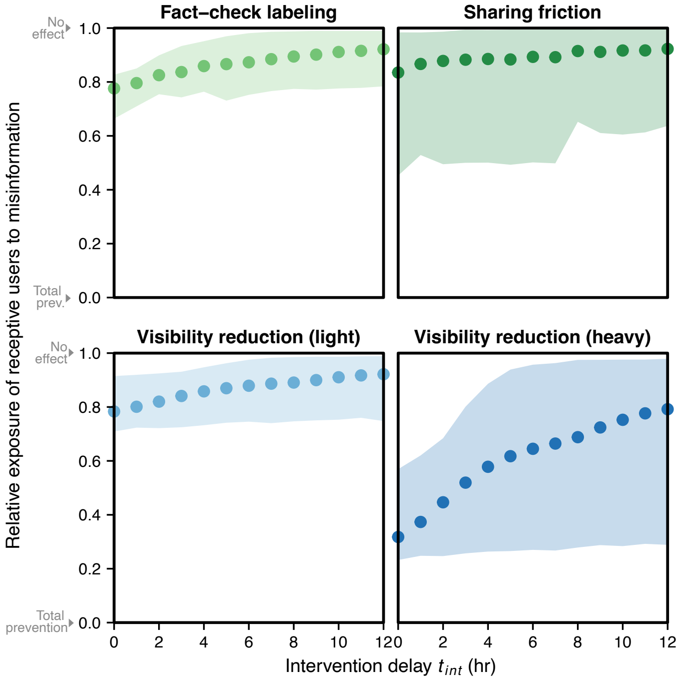

Figure 4: Simulating how intervention method and timing reduce misinformation exposure among receptive users. Intervention delay t_{int} is the number of hours it takes for an article to be fact-checked and an intervention to be deployed. Fact-check labeling decreases the probability of retweets by 25%, while sharing friction—adding extra steps to the retweet process for tweets sharing flagged material—decreases the probability of retweets by 75%. Visibility reduction is an intervention in which a tweet containing a flagged news article becomes 25% (light) or 75% (heavy) less likely to appear in other user's timelines. Points represent the mean of data-driven simulations across all false news articles in our data set and each ribbon covers 90% of all simulations. Each article was simulated 30 times for a given intervention type and delay.

that we have demonstrated here has promise to researchers and social media platforms alike as a method to measure the complex interplay of article exposure, receptivity, and sharing at scale. Importantly, our approach also allows for data-driven simulations of interventions so that researchers can assess the effect that various platform policies would have on article exposure among receptive users. Future work can build on our approach to create more precise methods for approximating exposure or add further important individual predictors of belief, such as cognitive reflection [50] or information literacy [51].

Beyond demonstrating a new method for inferring news article exposure among receptive users at scale, our study shows how the pattern of exposure among receptive users can look quite different from that among all users—a finding that is particularly relevant to researchers attempting to measure the impact of online misinformation. Current approaches focus on the patterns of sharing [20, 41, 52, 53] or general user exposure [53, 54], without considering individual predictors of belief among the exposed users. Our study, however, shows that belief in misinformation can look quite different when we account for user ideology and the ideological slant of individual articles. We found that misinformation can be shared by ideologically extreme users yet still expose a more cross-cutting audience; however, the political slant of misinformation articles greatly skewed who was receptive among those exposed because users were far more likely to believe misinformation if it aligned with their political leaning. Importantly, true news did not have this disconnect between exposure among all users and exposure among receptive users. Moreover, for both false/misleading and true news, the likelihood that a newly exposed user would be receptive to an article's content changed over time because the article's audience shifted as it spread: users who were most likely to believe an article tended to see it earlier than more skeptical users. Thus, our approach—one that accounts for individual user characteristics, article veracity, and article slant—provides a key insight into how exposure and receptivity can look quite different.

Of course, the magnitude of some of our findings, such as the amount of cross-ideology exposure, are subject to the assumptions of our simulations. of particular importance, we assumed that all followers of a given user are eventually exposed, which we know is not the case in reality given that, among other things, news feed ranking algorithms likely make users more likely to see content that aligns with their ideology (CITE). Still, even if news feed algorithms ensure that there was less cross-ideology exposure, our results would hold: the pattern of exposure among all users would not look like the pattern of exposure among receptive users because users are more likely to believe content that aligns with their ideology [43]. In fact, under that scenario, we would expect to see misinformation expose more receptive users on the ideological extremes relative to receptive users in the middle of the ideological spectrum.

In a sense, our findings stand in contrast to a prevailing notion that false news can spread faster and wider on social media than true news [20]. We instead estimated that true news accumulated far more exposure among receptive users over the initial hours after an article is first introduced on Twitter by virtue of the fact that true news is generally seen by more users and believed at a higher rate. In that sense, our findings are in line with research that has shown that the prevalence and impact of fake news on social media might be overstated [21, 53, 55]. However, since our study only focused on a relatively small number of top-trending news articles, further research will need to see if this pattern holds across a broader set of news articles. Still, like other research [53, 55, 56], we did find that misinformation sharing and receptivity was concentrated among very conservative users, which highlights concerns that, even when not widely believed, misinformation could reinforce political divisions [57] and further sort social networks along political lines [58].

Our results also highlight how the underlying social network structure of twitter is not symmetric, which may have created some of the asymmetries in how articles are seen and believed. Theoretical models predict that moderate users should bridge the ideological divide in social networks [58]. In one sense, our results matched this finding: we found that more moderate users did have more ideologically diverse follower networks. Yet, we also found center-right users had more diverse social connections than their center-left counterparts, allowing articles shared in right-leaning circles to gain a more diverse audience than those shared in left-leaning circles. Future work will need to further investigate the ideological organization of online social networks, but our findings suggest that this structural bias could drive major differences in how certain articles gain traction online.

Our simulations suggest that common interventions aimed at reducing the impact of misinformation are likely ineffective, particularly when it takes hours to identify, verify, and target a piece of misinformation. Due to the fact that exposure among receptive users quickly builds in the first few hours of circulation on Twitter, our simulations showed that interventions can only have a substantial effect if they are implemented within a few hours of the URL first being Tweeted. Research suggests that social media platforms could improve the effectiveness of interventions by combining multiple interventions [41] or instead using psychological nudges [19, 59]—for example, using cues in the user interface that prompt users to focus on the accuracy of the article—but these methods will still be highly time-sensitive and need to be deployed quickly to have an impact. Thus, the largest gain in effectiveness could come from methods that speed up the article verification process and offer faster turn around than professional fact-checkers—methods such as machine learning [60, 61] or crowd-sourcing [60, 62].

Besides highlighting how implementation speed is the key to effective misinformation interventions on social media, we also found that interventions that focus on decreasing the general visibility of misinformation (e.g., downranking) are more effective than other intervention methods. We assumed that visibility-focused interventions could reduce the likelihood of misinformation tweets showing up in users' news feeds by up to 75%, a rather conservative assumption given that aggressive downranking can, in effect, fully remove misinformation tweets. Despite this conservative estimate, we still found that downranking far out performs fact-check labeling and sharing friction. If we had run simulations of downranking with more aggressive assumptions (e.g., ¿90% visibility reduction), the gap in effectiveness between visibility-focused interventions and sharing-focused interventions would be even more stark.

Future work can expand on the modeling and data collection methods we attempted here. While we used a systematic method of data collection to ensure that the articles were top trending articles from a cross-section of news outlets, the five daily articles still represent a limited snapshot of trending news online. Moreover, we used simplistic assumptions—e.g., all followers are eventually exposed—as a first approach, but future work could implement more realistic assumptions grounded in data. For example, an extension of our method could use agent-based models to compare the spread of true and false/misleading news side-by-side under different intervention types. This would allow more robust estimations of the costs and benefits of certain interventions—e.g., one could quantify how much false/misleading news exposure among receptive users is reduced and compare it against the amount of true news that is inadvertently impacted by interventions.

4 Methods and Materials

In this section we present details about the data and methods used in this paper. In the first two subsections below we present how we collected survey data and separate Twitter data for the 139 news articles collected (Survey respondents are sampled separately from the Twitter users; the only overlap between the survey data and the Twitter data is the news articles themselves). Next, we explain how we constructed the retweet networks within the Twitter data, and how we estimated each Twitter user's political ideology, exposure, and receptivity to news articles. Finally, we explain the simulation of platform interventions designed to limit misinformation.

4.1 Survey Data

On 31 weekdays between November 18th, 2019 and February 6th, 2020, we sourced the most popular article published within the previous 24 hours from five news streams: liberal mainstream news domains; conservative mainstream news domains; liberal low-quality news domains; conservative low-quality news domains; and low-quality news domains without a clear political orientation. We created our two mainstream news streams by collecting the top 100 news sites by U.S. consumption.[8] To classify these news sources as liberal or conservative, we used scores of media partisanship from [63]. The top ten websites in each news stream (liberal or conservative) by consumption were chosen to construct a liberal mainstream and conservative mainstream stream. For our low quality news sources, we relied on the list of low-quality news sources from [64] that were still active in November 2019, which we then subsequently classified into three streams with a panel of three undergraduate research assistants (see: Supplemental Methods): liberal leaning sources, conservative leaning sources, and those without a clear partisan orientation. For the mainstream news feeds, we determined popularity in each news stream using CrowdTangle, a content discovery and social monitoring platform that tracks the popularity of URLs on Facebook pages. For the low-quality news feeds, we determined popularity using RSS feeds.[9] This transparent article selection process allowed us to source five daily top-trending news articles from across the ideological spectrum for 31 separate days.

To determine the veracity of the articles, we sent each day's five selected articles to professional fact-checkers. We hired six professional fact checkers from leading national media organizations to assess each article during the initial 24 hours after publication.[10] We used the modal response of the professional fact checkers ("true", "false/misleading", or "could not determine"). This yielded 37 false/misleading articles, 102 true articles, and 16 articles that the fact-checkers could not agree on, the latter of which were removed from analysis.

To determine how likely people were to believe the articles as they encountered them on social media, we sent each day's five selected articles to a panel of U.S. respondents [65]. Each daily survey was completed by between 140 and 160 American respondents that were recruited by the Qualtrics survey firm and that were balanced on age, gender, partisanship, and education. Every respondent evaluated three articles randomly selected from the day's five selected articles. Each article was therefore assessed by approximately 90 respondents who evaluated these articles within 48 hours of its publication, giving us a measure of real-time belief in these stories. Collecting evaluations of the most popular false news articles directly after publication is a key innovation that enabled us to measure belief in each article by ideological group in the period directly after publication that these news articles were spreading on social media [18]. Respondents evaluated each article using a variety of criteria, the most germane of which was a categorical evaluation question: "What is your assessment of the central claim in the article?" to which respondents could choose from three responses: (1) True (2) Misleading/False (3) Could Not Determine (See: Supplemental Information for full survey instrument). For the 37 false/misleading articles we collected 3,394 evaluations from 2,751 unique respondents. For the 102 true articles we collected 10,024 evaluations from 5,000 unique respondents.

[8] These were identified by Microsoft Research's Project Ratio between 2016-2019: https://www.microsoft.com/en-us/research/project/project-ratio/

[9] We used RSS feeds for the low quality sources because most low-quality sources were not tracked by CrowdTangle, as the publisher pages had been removed from the platform; for more on CrowdTangle see https://www.crowdtangle.com/.

[10] These professional fact-checkers were recruited from a diverse group of reputable publications (none of the publications that we ask individuals to fact-check to ensure no conflicts of interest) and paid $10.00 per article.

4.2 Twitter Data

Completely separate from the survey data, we also collected Twitter data for all of the articles. From the 139 true and false/misleading articles, we compiled a twitter data set of all tweets and users who shared each of the article URLs within up to one week after publishing. In total, our data set comprised 139,734 tweets and 92,514 unique users who shared at least one article link (hereafter, referred to as "tweeters"). The true articles comprised 94,422 tweets sent by 72,304 unique users, while false/misleading articles comprised 45,312 tweets sent by 27,430 unique users. We also collected each tweeter's friend and follower network in order to quantify how many other users may have been exposed to the article tweets. This totaled 128,453,928 unique followers and 21,871,687 unique friends of our tweeters (there is overlap between the set of users in the friend and follower lists).

4.3 Constructing retweet networks

On Twitter, a user can share an article by either (a) directly tweeting an article link, (b) retweeting another user who shared the article link, or (c) quoting another user's tweet of the article link and adding extra text. The latter two methods—retweets and quote tweets—are major features on Twitter that allow users to share another user's tweet and thereby allow information to spread beyond the followers of the original tweeter. Thus, information on Twitter can be introduced into new parts of the social network and quickly spread outward from that point of origin: after a user tweets a news article their followers (and their followers' followers) can retweet the article, expanding the audience of the original article tweet by many folds.

To capture this dynamic and visualize the spread of our tracked news articles on Twitter, we constructed retweet networks using established methods for time-inferred information diffusion on Twitter [20, 66]. Using data from the Twitter API, this method determines the flow of a retweet using the time and friend/follower networks of users. This is necessary because Twitter data does not directly include information about who retweeted which user and instead only includes information about which original tweet they retweeted, even if they retweeted a friend who had retweeted the original tweet.

To build the retweet network, we infer the path of a retweet (or quote tweet) by considering the time it was shared and friend-follower networks. If the user who retweeted the tweet (hereafter, "retweeter") follows the user indicated as the origin of the tweet in the Twitter API's data, then we consider that a direct retweet of the original tweeter. On the other hand, if the retweeter does not follow the original tweeter, we determine if any of her friends shared the same tweet in the time prior to the focal retweet and we assume the flow of the retweet is from the friend who most recently shared the same retweet. In the event the retweeter does not follow the original tweeter and also has no friends who shared the tweet prior to her own retweet, we consider it a direct retweet of the original tweeter. This may reflect instances where users see tweets on their timeline that are not from users they follow—a relatively common occurrence on Twitter[11]—which would create this pattern of information flow.

4.4 Estimating user ideology

To characterize the ideology of users in our study, we used an established method that infers a user's ideology from the news, political, and cultural accounts that they follow [44]. The method assumes that users are more likely to follow accounts that align with their personal ideology [44, 67]. Therefore, leveraging the known ideology of prominent news, political, and culture accounts, the method uses correspondence analysis to estimate a user's ideology, so long as they follow at least one of the prominent accounts with known ideology.

To determine the ideology of tweeters in our dataset, we cross referenced each tweeter's unique user ID in our study against the data set of scored ideologies. If we were unable to directly calculate the ideology of a user because they did not follow any prominent news, cultural, or political accounts, we first calculated the ideology of their friends and then used the mean of the ideological scores of their friends.

In our study, we are interested in measuring how many and what kind of users are exposed to news articles on Twitter. To do so, we needed to calculate or estimate the ideology scores for all followers who are exposed to article tweets since they sit "downstream" of tweeters in the directed social network of Twitter. To determine the ideology of the followers in our dataset, we again cross referenced each follower's unique user ID against the dataset of scored ideologies. For followers that we were unable

[11] This occurs for viral tweets about topics of interest to Twitter users or if a Tweet was liked by an account they follow

to find in the ideology data set, we were unable to estimate their ideology in the same method as tweeters—i.e., by averaging the ideology of their friends—since we lacked this necessary social network data. Instead, we estimated missing follower ideologies by assuming that the follower ideologies of a tweeter would follow a normal distribution [44, 67]. Thus, using the follower ideology scores we did have for a given tweeter, we could take a reasonable guess at what the value of the missing follower ideologies might be.

To infer the distribution of of follower ideologies for the tweeters—e.g., the mean and standard deviation of follower ideologies for tweeter i—we used a Bayesian approach that consists of two steps: (1) inferring the population distribution of user ideologies using the known ideology scores in our data set and (2) inferring the distribution of follower ideologies for each tweeter. The first step allowed us to create a baseline assumption for ideologies among Twitter users that could be used as a prior distribution when inferring the distribution of follower ideologies for each Tweeter.

To estimate the distribution of Twitter user ideologies, we sampled 500,000 ideologies without replacement from all scored users in our dataset, including ideologies from the tweeters and their friends and followers. We then used Bayesian Hamiltonian Monte Carlo to infer a normal distribution $N(\mu_{pop}, \sigma_{pop})$ that best described the population's ideology. Because ideology scores tend to be normally distributed around 0 [67], we assumed a normal distribution prior ($\mu = 0$, $\sigma = 2$) for our population mean, while we assumed an exponential distribution prior ($\lambda = 2$) for our population standard deviation.

Using the posterior of our population estimate as our prior distribution, we then inferred the distribution of follower ideology scores for each unique tweeter. For our each unique tweeter i, we used their followers with known ideology scores as samples and then used Bayesian Hamiltonian Monte Carlo to infer the mean μ_i and standard deviation σ_i for the normal distribution $N(\mu_i, \sigma_i)$ that fit the data. For the 2,678 tweeters (3.03% of all tweeters) that had no followers with known ideology scores and therefore no data we could use for inference, we used our prior distribution and assumed their followers had ideologies that matched that of the population. Users who did not have any followers with known ideology scores tended to have very few followers (mean follower count < 5) and therefore do not greatly affect the analyses in our study.

4.5 Estimating user exposure and receptivity to news articles

To understand who may see and be receptive to news on Twitter, we needed to establish when and how users may be exposed to a particular news articles. To do this, we needed to try to infer when a user might reasonably see an article shared by someone they follow, and we needed to know what kinds of users (i.e., ideologies) are being exposed, since the likelihood of belief is driven by individual-level characteristics. In this paper, we focused on ideology as the main predictor of article belief because it is one of the best predictors of belief in misinformation [18] and because there are established methods to estimate it from social media data [44]. Unfortunately, the other largest predictor from previous literature [43]—familiarity with the political narrative presented in the article—is much harder to infer from Twitter data, but future work could attempt to incorporate information in this regard as well.

We made the simplifying assumption that all followers of a user are potentially exposed to the user's tweet. We made this simplifying assumption because the Twitter API does not provide specific information on who actually saw a specific tweet, and therefore this was the most basic dynamic we could assume without knowledge of Twitter's news feed ranking algorithm. Thus, our measurement of exposure is quantified as potential exposure, which can be thought of as the upper limit of who potentially saw a tweet. We calculated exposure on a per-article basis, and we counted exposure as a one-time event, meaning that a user could only be "exposed" to an article once, even if multiple accounts that they follow shared the article over the span of a few hours.[12]

For each news article, we used a simple procedure to estimate when users were exposed to the story. First, for each tweet of that article, we assigned a specific "exposure time" to each follower of each tweeter. We assumed that users are exposed to a tweet with some delay rather than instantaneously seeing the tweet as it is shared. Thus, each follower's time of exposure was calculated by adding a time delay to the timestamp of the tweet: for each follower of a tweeter, we randomly drew a delay time (in hours) from a truncated normal distribution ($\mu = 1$, $\sigma = 2$, and minimum limit of 0) and added this to the timestamp of the tweet, yielding that follower's "time of exposure" for that specific article. [13]. Next,

[12] Future work could relax this simplifying assumption to account for the likely scenario that users may see an article multiple times if more than one of the people they follow shares it.

[13] While we explored this simplified assumption, future work could explore different assumptions for how exposure occurs over time.

if we estimated that a user was exposed to a tweet multiple times, we only used the earliest exposure time: we combined all the exposed follower lists of those that tweeted a specific article, sorted the list according to exposure time, and kept only the earliest exposure time for each unique follower. In the end, this process allows us to estimate how many unique users were potentially exposed to an article and when they were exposed.

Once we estimated who was exposed to each news article, we then estimated the ideologies among those exposed. The method mentioned in the previous paragraph results in a list of users exposed to each tweet, thus allowing us to say who a particular tweeter exposed. We then inferred the ideologies of these exposed users in a two-step process: (1) using the known ideology scores that we do have, and (2) drawing the remaining ideologies as samples from that tweeter's estimated distribution of follower ideologies. For example, imagine that we estimate that tweeter *@ExampleUser* exposed 1,000 users to a news article and that we know the ideology scores for 200 of those exposed users. Then to infer the total collection of ideologies exposed, we simply draw $1,000 - 200 = 800$ samples from the estimated distribution of *@ExampleUser*'s followers ideology scores.

To estimate receptive users among those exposed, we used the previously described survey data (see 4.1) that asked panels of random people to assess the veracity of an article. Because we also asked respondents for personal demographic data, we could use the survey to measure the frequency at which people of a given ideological category—”Very Liberal”, ”Liberal”, ”Somewhat Liberal”, ”Moderate”, ”Somewhat Conservative”, ”Conservative”, ”Very Conservative”—believed a given article to be true, i.e., $p(belief|ideology, article)$. However, since the users in our dataset have a numeric ideology score and the survey uses categorical ideology scores, we mapped the ideology of users in our data set according to the survey data using the following schema:

Category	Ideology scores
Very Conservative	$x > 2.5$
Conservative	$2.5 \leq x > 1.5$
Somewhat Conservative	$1.5 \leq x > 0.5$
Moderate	$0.5 \leq x > -0.5$
Somewhat Liberal	$-0.5 \leq x > -1.5$
Liberal	$-1.5 \leq x > -2.5$
Very Liberal	$x \leq -2.5$

Thus, for each tweet in our data set, we estimated the number of receptive users among those exposed by multiplying the number of newly exposed users in an ideology category by the corresponding probability of belief, $p(belief|ideology, article)$. For example, we might estimate that a given tweet of article X exposed 50 “very liberal” users, which we then multiply by the survey-provided 70% belief rate for this article among “very liberal” users, allowing us to estimate that 35 of the “very liberal” users were receptive to the article. Importantly, we describe our measurement of ”receptivity” because we cannot causally measure the belief among these social media users. Instead, we can concretely estimate who was both exposed and *highly likely* to believe the article.

Because our estimation process used random sampling—e.g., drawing exposure time or follower ideology from a distribution—we calculated total user exposure and receptive user exposure from scratch each time we need to estimate exposure under different scenarios, e.g., for each simulated intervention described in 4.6.

In one sense, our approach to estimating the exposure of receptive users is a conservative one: we only count strict affirmative belief (i.e., user says the article is true), while we are not counting cases where a user is unable to discern whether an article is true or false. On the other hand, our method provides a sense of an upper bound since we assume that all followers are eventually exposed—an assumption that will not hold up in reality. Thus, considering these two factors, our approach provides a conservative upper bound on the potential exposure of receptive users to misinformation, but further work can increase the accuracy of this estimate.

4.6 Simulating platform interventions to limit misinformation

Social media platforms, including Twitter, have begun attempting to limit the spread of misinformation. Individual-level experiments have informed us on how particular interventions may limit the chance that an individual shares a piece of misinformation, but we do not have an understanding of how these interventions work at scale and how effective they ultimately are at preventing misinformation exposure among receptive users.

We used our data set to conduct simulations that test how misinformation exposure among receptive users is affected by (a) the speed of fact-checking and (b) the method of intervention. Since an article must first be labeled as false/misleading by external professional fact-checkers before an intervention can be deployed, we focused our simulations on the articles in our data set that were rated false/misleading by fact-checkers. Due to the fact-checking process and other logistics of deploying an intervention, we assumed that an intervention does not target tweets sharing a flagged articled until time t_{int}, the number of hours after the first share of a targeted article's URL. We varied the intervention time t_{int} to see how the speed of this process may affect its effectiveness in decreasing misinformation exposure.

We simulated two methods that are commonly used to try to reduce the sharing of misinformation. First, we simulated *fact-check labeling*, in which tweets sharing misinformation are given a warning label stating that the tweet was fact-checked as false by independent reviewers. Using previous research on the effect of fact-check labels on individual behavior, we assumed that fact-check labeling makes a user 25% less likely to retweet [22] and 17% less likely to believe [42] a flagged tweet starting at time t_{int}. Second, we simulated *sharing friction*, in which extra steps (e.g., more clicks) are added to the retweet process in the hope of making users reconsider sharing a questionable tweet. Again using previous research on the effect of sharing friction on individual behavior, we assumed that sharing friction makes a user 75% less likely to retweet [22] a flagged tweet starting at time t_{int}. Thus, in a simulation run for a particular false/misleading news article, we took our set of real article tweets and removed a retweet after time point t_{int} with probability 0.25 (fact-check labeling) or 0.75 (sharing friction), thereby simulating the prevention of a retweet due to the intervention. Moreover, when a retweet was removed, we removed all subsequent retweets of that specific tweet, since users would be unable to see and retweet a tweet that did not occur. Additionally, in the case of fact-check labeling, we would also decrease the belief rates in the survey by 17% when calculating exposure among receptive users (see 4.5).

We also simulated *visibility reduction* methods. Twitter states in its terms of service that tweets sharing questionable content may be made ”less visible in other user's timelines.” Therefore, we simulated visibility reduction by making it 25% (light visibility reduction) or 75% (heavy visibility reduction) less likely that a user sees tweets sharing flagged articles. Unlike the simulations of sharing-focused interventions, these numbers were not based on the literature but are instead represented comparable values for what we might imagine a less aggressive and more aggressive visibility-focused intervention might look like. We simulated this intervention during the exposure calculation for a given false/misleading news article: taking the list of users that we estimated were exposed to the article, we would probabilistically remove users exposed after time t_{int}, prior to dropping duplicate exposures of the same user (see 4.5). Simulating the intervention in this manner allows us to realistically capture instances where a user may miss the first exposure chance due to the intervention but later see it if they are following multiple accounts sharing the article.

References

[1] Elisa Shearer. Social media outpaces print newspapers in the us as a news source. *Pew Research Center*, 2018.

[2] Jae Kook Lee and Eunyi Kim. Incidental exposure to news: Predictors in the social media setting and effects on information gain online. *Computers in Human Behavior*, 75:1008–1015, October 2017. Publisher: Pergamon.

[3] Meital Balmas. When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism. *Communication Research*, 41(3):430–454, July 2014. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[4] Leticia Bode. Political News in the News Feed: Learning Politics from Social Media. *Mass Communication and Society*, 19(1):24–48, January 2016. Publisher: Routledge.

[5] Jessica T. Feezell. Agenda Setting through Social Media: The Importance of Incidental News Exposure and Social Filtering in the Digital Era. *Political Research Quarterly*, 71(2):482–494, December 2018. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[6] Augusto Valeriani and Cristian Vaccari. Accidental exposure to politics on social media as online participation equalizer in Germany, Italy, and the United Kingdom. *New Media and Society*, 18(9):1857–1874, November 2016. Publisher: SAGE PublicationsSage UK: London, England.

[7] Sangwon Lee and Michael Xenos. Incidental news exposure via social media and political participation: Evidence of reciprocal effects. *New Media and Society*, 24(1):178–201, October 2022. Publisher: SAGE PublicationsSage UK: London, England.

[8] Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017.

[9] Nejla Asimovic, Jonathan Nagler, Richard Bonneau, and Joshua A Tucker. Testing the effects of facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences*, 118(25):e2022819118, 2021.

[10] Rachel Greenspan. The qanon conspiracy theory and a stew of misinformation fueled the insurrection at the capitol. *The Insider*, 2021.

[11] Bill Adair and Phillip Napoli. Misinformation fueled the capitol riots — a biden commission could chart a path forward. *The Hill*, 2021.

[12] Kirk Siegler. Misinformation spread by anti-science groups endangers covid-19 vaccination efforts. *NPR*, 2021.

[13] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

[14] John Zarocostas. How to fight an infodemic. *The Lancet*, 395(10225):676, February 2020. Publisher: Elsevier.

[15] Jay J Van Bavel, Elizabeth A Harris, Philip Pärnamets, Steve Rathje, Kimberly C Doell, and Joshua A Tucker. Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*, 15(1):84–113, 2021.

[16] Nathaniel Persily and Joshua A Tucker. Social media and democracy: The state of the field, prospects for reform. 2020. Publisher: Cambridge University Press.

[17] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378, January 2019. Publisher: American Association for the Advancement of Science.

[18] Kevin Aslett, William Godel, Zeve Sanderson, Nathaniel Persily, Jonathan Nagler, Richard Bonneau, and Joshua A Tucker. An externally valid method for assessing belief in popular fake news. *Unpublished Manuscript*, 2021.

[19] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, pages 1–6, March 2021. Publisher: Springer Science and Business Media LLC.

[20] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, March 2018. Publisher: American Association for the Advancement of Science.

[21] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586, 2019.

[22] Emeric Henry, Ekaterina Zhuravskaya, and Sergei Guriev. Checking and Sharing Alt-Facts. *American Economic Journal: Economic Policy*, 14(3):55–86, August 2022.

[23] Kelvin K King and Bin Wang. Diffusion of real versus misinformation during a crisis event: a big data-driven approach. *International Journal of Information Management*, page 102390, 2021.

[24] Babajide Osatuyi and Jerald Hughes. A tale of two internet news platforms-real vs. fake: An elaboration likelihood model perspective. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[25] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013.

[26] Joseph B Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T Bergstrom, Miguel A Centeno, Iain D Couzin, Jonathan F Donges, Mirta Galesic, Andrew S Gersick, Jennifer Jacquet, Albert B Kao, Rachel E Moran, Pawel Romanczuk, Daniel I Rubenstein, Kaia J Tombak, Jay J. van Bavel, and Elke U Weber. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 118(27):2025764118, 2021. Publisher: PNAS.

[27] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

[28] Dan M Kahan. Misconceptions, misinformation, and the logic of identity-protective cognition. *SSRN*, 2017.

[29] Jay J Van Bavel and Andrea Pereira. The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, 22(3):213–224, 2018.

[30] Patricia Moravec, Randall Minas, and Alan R Dennis. Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, 43, 2019.

[31] Gordon Pennycook, Jonathon McPhetres, Bence Bago, and David G. Rand. Beliefs About COVID-19 in Canada, the United Kingdom, and the United States: A Novel Test of Political Polarization and Motivated Reasoning. *Personality and Social Psychology Bulletin*, 48(5):750–765, May 2022. Publisher: SAGE Publications Inc.

[32] Nathaniel Sirlin, Ziv Epstein, Antonio A Arechar, and David G Rand. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review*, 2021.

[33] Gregory Eady, Tom Paskhalis, Jan Zilinsky, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Exposure to the russian internet research agency foreign influence campaign on twitter in the 2016 us election and its relationship to attitudes and voting behavior. *Nature Communications*, 14(62), 2023.

[34] Marianna Spring. Covid vaccine: Social media urged to remove 'disinfo dozen'. *BBC*, 2021.

[35] Adam Mosseri. Addressing hoaxes and fake news. *Facebook Newsroom*, 2016.

[36] Jeff Smith, Grace Jackson, and Seetha Raj. Designing against misinformation. *Medium*, 2017.

[37] Kate Conger. Twitter has labeled 38% of trump's tweets since tuesday. *New York Times*, 2020.

[38] Guy Rosen. Remove, reduce, inform: New steps to manage problematic content. *Facebook Newsroom*, 2019.

[39] Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School Misinformation Review*, October 2021. Publisher: Shorenstein Center for Media, Politics, and Public Policy.

[40] Christine Geeng, Tiona Francisco, Jevin West, and Franziska Roesner. Social Media COVID-19 Misinformation Interventions Viewed Positively, But Have Limited Impact. *arXiv preprint arXiv:2012.11055*, December 2020. arXiv: 2012.11055.

[41] Joseph B. Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Emma S. Spiro, Kate Starbird, and Jevin D. West. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, pages 1–9, June 2022. Publisher: Nature Publishing Group.

[42] Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior*, 42(4):1073–1095, December 2020. Publisher: Springer.

[43] Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, Richard Bonneaua, and Joshua A Tucker. An ecologically and externally valid approach to assessing belief in popular misinformation. Working paper, Unpublished.

[44] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, 2015.

[45] Meta. About fact-checking on facebook and instagram, 2023. `https://www.facebook.com/business/help/2593586717571940?id=673052479947730`, Last accessed on 2023-02-14.

[46] Twitter. How we address misinformation on twitter, 2023. `https://help.twitter.com/en/resources/addressing-misleading-info`, Last accessed on 2023-02-14.

[47] Youtube. How does youtube address misinformation?, 2023. `https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/`, Last accessed on 2023-02-14.

[48] TikTok. An update on our work to counter misinformation, 2022. `https://newsroom.tiktok.com/en-us/an-update-on-our-work-to-counter-misinformation`, Last accessed on 2023-02-14.

[49] Harrison Mantas. Twitter finally turns to the experts on fact-checking. *Poynter*, Aug 2021.

[50] Gordon Pennycook and David G Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019.

[51] S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. Does media literacy help identification of fake news? information literacy helps, but other literacies don't. *American Behavioral Scientist*, 65(2):371–388, 2021.

[52] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1):eaau4586, January 2019. Publisher: American Association for the Advancement of Science.

[53] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378, January 2019. Publisher: American Association for the Advancement of Science.

[54] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, April 2019. Publisher: SAGE Publications Ltd.

[55] Gregory Eady, Tom Paskhalis, Jan Zilinsky, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14(1):62, January 2023. Number: 1 Publisher: Nature Publishing Group.

[56] Andrew M. Guess, Brendan Nyhan, and Jason Reifler. Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 4(5):472–480, May 2020. Publisher: Nature Research.

[57] Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, 115(3):999–1015, August 2021. Publisher: Cambridge University Press.

[58] Christopher K. Tokita, Andrew M. Guess, and Corina E. Tarnita. Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*, 118(50):e2102147118, December 2021. Publisher: National Academy of Sciences ISBN: 2102147118.

[59] Valerio Capraro and Tatiana Celadin. "I Think This News Is Accurate": Endorsing Accuracy Decreases the Sharing of Fake News and Increases the Sharing of Real News. *Personality and Social Psychology Bulletin*, page 01461672221117691, August 2022. Publisher: SAGE Publications Inc.

[60] William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua Tucker. Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*, 1(1):1–36, October 2021. Publisher: Stanford Internet Observatory.

[61] Meta. Meta launches new content moderation tool as it takes chair of counter-terrorism ngo, 2022. https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/, Last accessed on 2023-02-14.

[62] Nicolas Pröllochs. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:794–805, May 2022.

[63] Gregory Eady, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler. News sharing on social media: Mapping the ideology of news media content, citizens, and politician. *OSF Preprints*, 2020.

[64] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, 2019.

[65] Kevin Aslett, William Godel, Zeve Sanderson, Nate Persily, Jonathan Nagler, Richard Bonneau, and Joshua Tucker. Measuring belief in fake news in real-time. *Charlottesville: OSF Preprints. Retrieved April*, 2021.

[66] Sharad Goel, Duncan J. Watts, and Daniel G. Goldstein. The Structure of Online Diffusion Networks. *Proceedings of the 13th ACM Conference on Electronic Commerce*, 1(212):623–638, 2012. arXiv: 1502.07526v1 Publisher: ACM Press Place: New York, New York, USA ISBN: 9781450314152.

[67] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, 26(10):1531–1542, October 2015. Publisher: SAGE Publications Inc.

5 Supplementary Information

5.1 Supplementary Text

5.1.1 Tweeting

Twitter users with more pronounced ideologies were more likely to share news articles relative to users with more moderate ideology scores (Figure S1A). In line with findings elsewhere in the literature, we find that the sharing of false/misleading news articles was more concentrated among the political right. Additionally, articles shares were more likely to come from a retweet rather than an original share of the article URL, and among retweets, most were direct retweets of the original tweeter rather than indirect retweets from a friend-of-a-friend (Figure S3). Overall, this suggests that information does spread readily from the original sharer, but typically not farther than one degree of separation.

5.1.2 Diffusion of news through social networks

By constructing retweet networks, we find that the diffusion of news on Twitter is a skewed structurally and politically (Figure S1D). For a given article, a few users garner most of the retweets, which likely reflects the hub-and-spoke (i.e., scale-free) nature of Twitter's social networks, where relatively few users have huge numbers of followers while most other users have a modest number of followers (Figure S4). In our retweet networks, the hubs tended to be the Twitter accounts of news outlets or prominent political/cultural figures. We also find that retweet networks are relatively politically homogeneous, such that a given article tends to only be tweeted by a largely left-leaning or right-leaning set of users. Interestingly, on a per-article basis, tweeters of false/misleading news articles tend to be less ideologically diverse than tweeters of true news ($BF_{10} \gg 100$, $p(\mu_1 \neq \mu_2) = 1$; $t(52.412) = -2.11$, $p = 0.0396$) (Figure S1B). This difference in ideological heterogeneity may reflect the homophily we find in social ties on Twitter (Figure 1C, S6): left-leaning users tend to have left-leaning followers, right-leaning users tend to have right-leaning followers, and moderate users tend to have the most ideological diverse set of followers (Figure S6). Thus, since highly ideological individuals are more likely to share news articles from fringe news sources—the main source of false/misleading news—and are more likely to have politically similar followers, false/misleading news articles may be more likely to be introduced to and circulate in politically uniform social networks.

5.2 Supplementary Methods

5.2.1 Coding the partisan slant of news sources and articles

We determined the partisan lean of the low-quality news sources through a panel of three undergraduate research assistants who served as independent coders. We asked the coders to classify each low-quality news source as either "liberal" or "conservative" by using information found in a news source's headlines, article content, website "about" page, and, if applicable, known partisan affiliations. If fewer than 50% of a news source's content appeared to have partisan content, coders were told to rate a news source as "unclear". In the event that the coders did not unanimously agree on the partisan slant of a news source, a fourth coder was brought in and the majority rating was used. Coders agreed unanimously on 75% of news sources. In the end, six low-quality sources were classified as "liberal", fifty were classified as "conservative", and forty three were classified as "unclear".

Similarly, we determined the partisan lean of an article's content through a panel of four undergraduate research assistants serving independent coders. For a given article, the coders were asked to view only the article's title and body text, and they were then asked to rate the article as "liberal", "conservative", "neutral", or "unclear". The final rating for an article's partisan lean was the modal rating of the four coders. In the event that there was a tie, a graduate student coder was used as a tiebreaker.

5.3 Supplementary figures

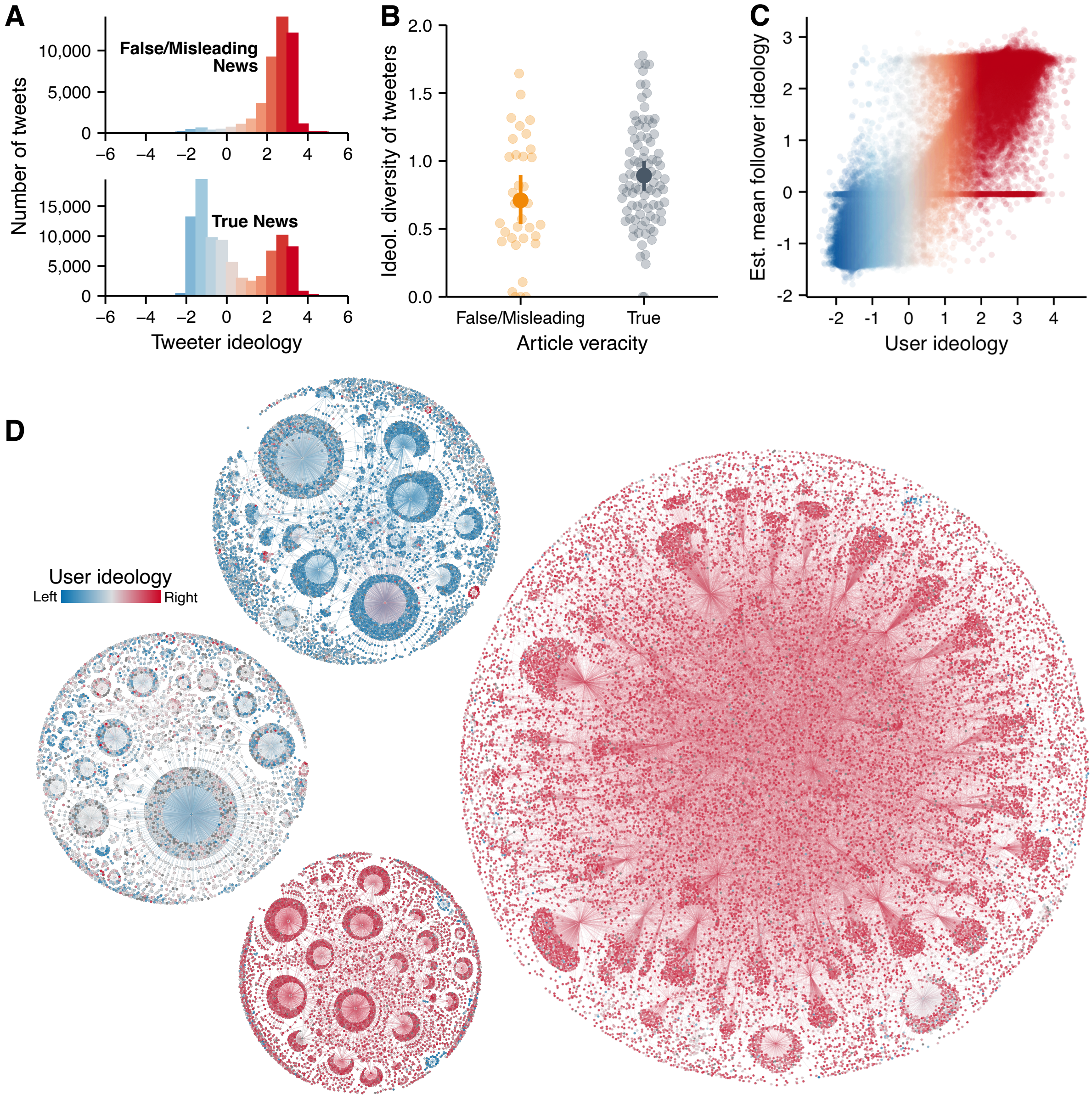

Figure S1: The pattern news articles sharing and diffusion on Twitter. (A) Histogram of ideology among sharers of real and fake news. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology. (B) The ideological diversity among tweeters of of real and fake news articles, as measured by the s.d. of ideology for all tweeters of a given article. Smaller points are the ideological diversity of individual articles, while the large points represent the mean (±99% credible interval) across all real/fake news articles. (C) User ideology and estimated mean follower ideology for all unique tweeters in our data set. A positive relationship would suggest homophily along political lines. (D) Example retweet networks for individual articles in our data set. Each point is a tweeter of the article, colored by that user's ideology, and arrows indicate the flow of retweets.

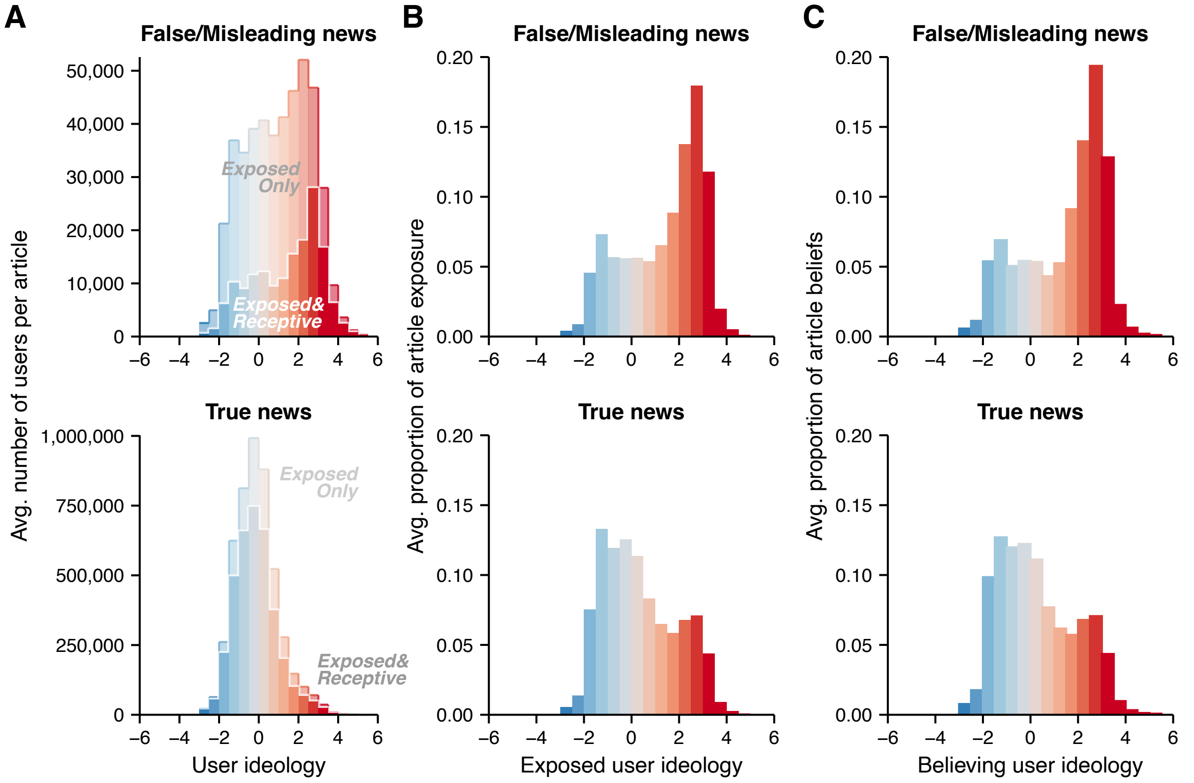

Figure S2: The average patterns of total user exposure and receptive user exposure per news article. (A) Estimates for the average number of users who could have potentially been exposed and receptive to the true and false/misleading news articles in our dataset. The average is calculated by first estimating the total number of exposed users and exposed receptive users, and then dividing by the number of articles in each dataset. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology. (B) The average distribution of all user ideologies exposed to a given false/misleading or true news article, as rated by professional fact-checkers. Ideology scores for each user are calculated using the method in [44], which infers a user's ideology from the political accounts that they follow. (C) The average distribution of all receptive user ideologies exposed to a given false/misleading or true news article.

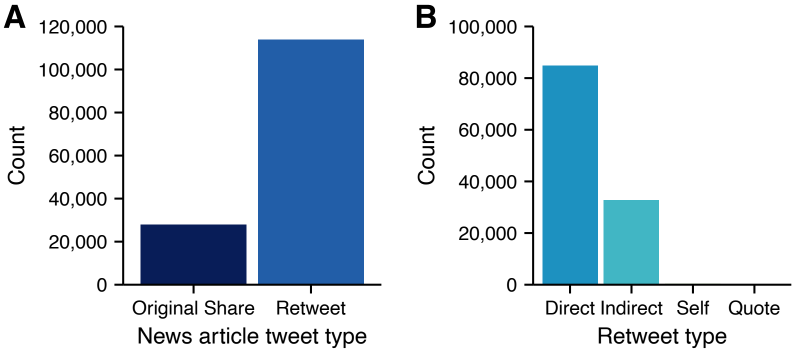

Figure S3: (A) Break down of the type of URL shares on Twitter in our dataset. (B) Breakdown of the type of retweets in our dataset. Direct retweets occur when a user retweets the original sharer of a link, indirect retweets occur when a user retweets someone who had already retweeted the original sharer, self retweets occur when someone retweets herself, and quote tweets occur when someone retweets another user but adds additional text to their retweet.

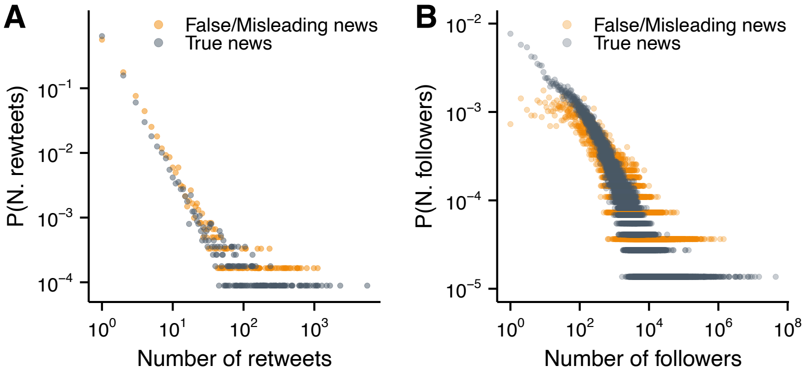

Figure S4: (A) Distribution of retweet frequencies for a given news articles. This plot describes how frequently an individual had N retweets among true and false/misleading news articles. (B) Degree distribution of users who shared false/misleading and true news articles. This plot describes how frequently a tweeter had K followers.

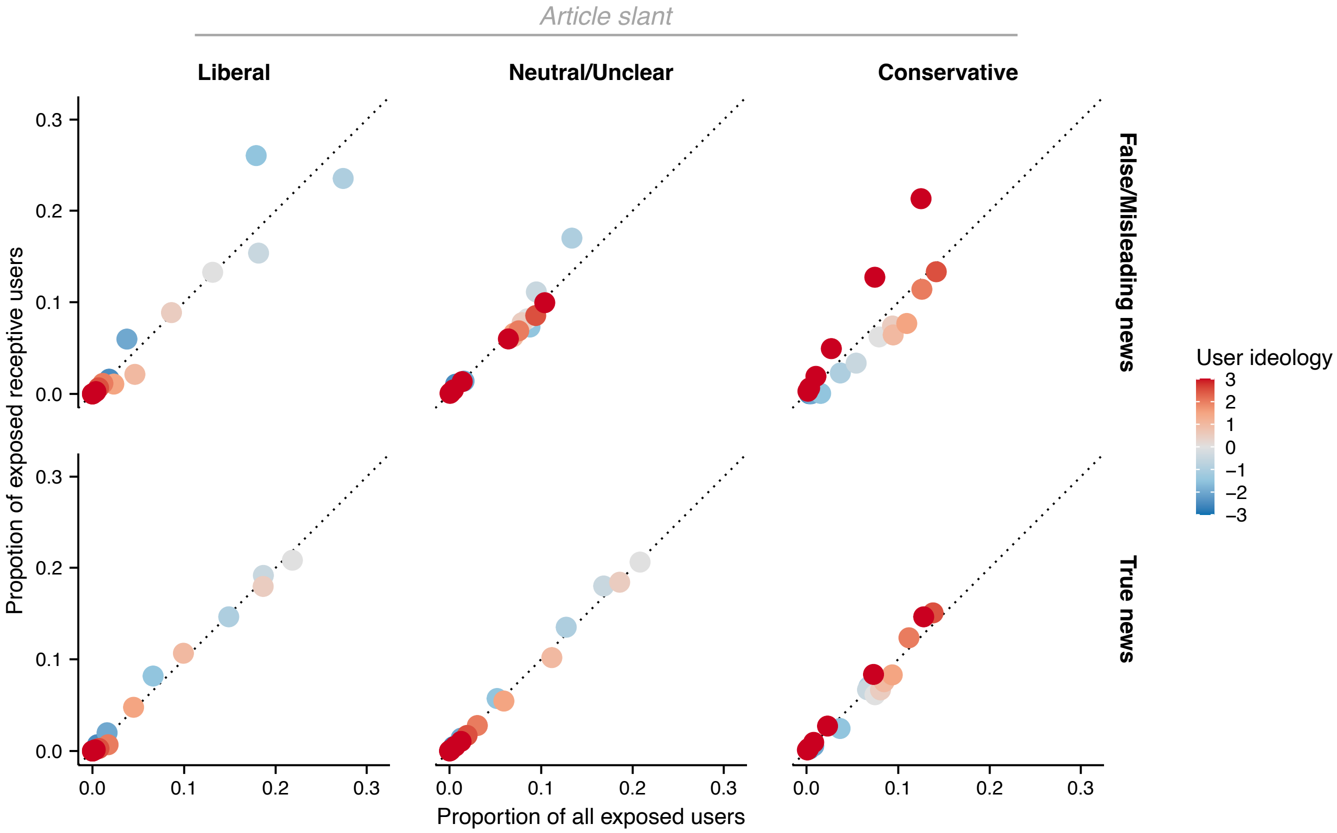

Figure S5: The relative abundance of a given ideology among all exposed users vs. all exposed receptive users, broken out by article content. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology. The dashed line represents a 1:1 ratio, indicating that users of that ideology represented a proportion of exposed receptive users that is expected given the proportion they made up of all exposed users. Points above the line indicate that users of that ideology make up a disproportionate share of receptive users among those exposed to an article.

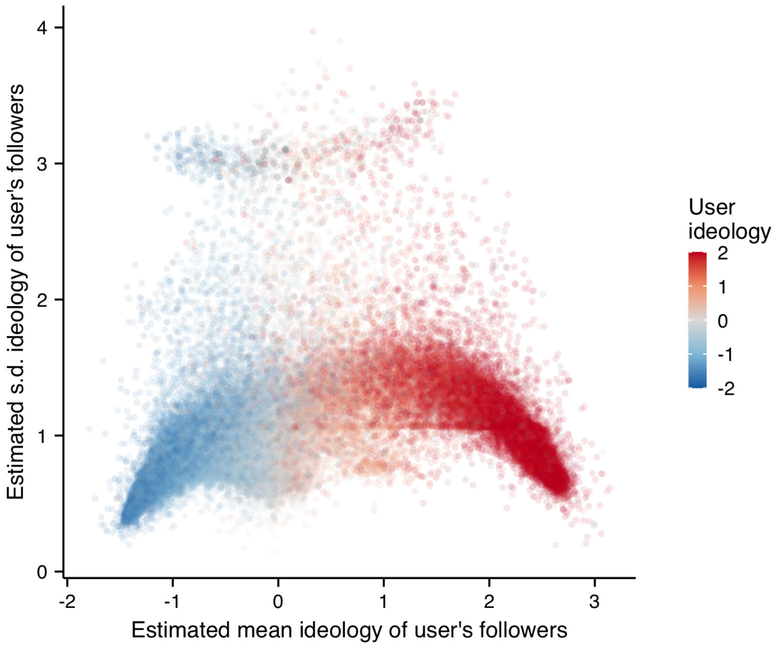

Figure S6: Estimated mean and standard deviation of follower ideology for all unique Tweeters in our dataset. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology. Higher values on the y-axis indicate that the user has more ideologically diverse followers, while values to the extremes of the x-axis indicate followers with strong political skew.

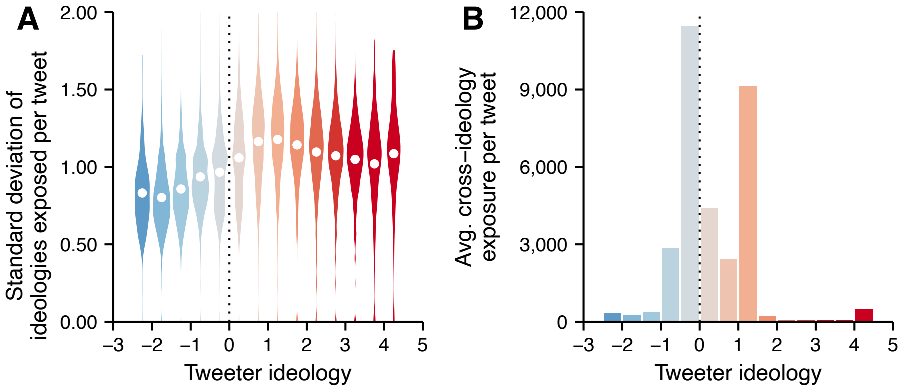

Figure S7: Center-right users exposed the most ideologically diverse audience on Twitter and created a disproportionate amount of cross-ideology exposure to news articles. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology. (A) Standard deviation of users exposed to tweets, broken out by tweeter ideology. White points represent the mean. (B) Average number of cross-ideology exposures per tweet, broken out by tweeter ideology. Cross-ideology exposure is defined as exposing a user who sits on the other side of the ideological spectrum, i.e., a right-leaning (ideology score > 0) user exposing a left-leaning (ideology score < 0) user.

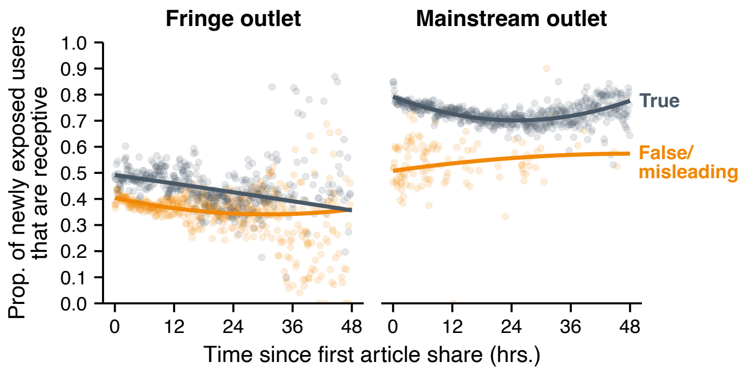

Figure S8: Proportion of newly exposed users that are receptive over time, broken out by article veracity and news source type. Points are the binned mean across all tweets within 5-minute intervals. Lines are the best-fit Bayesian regression.

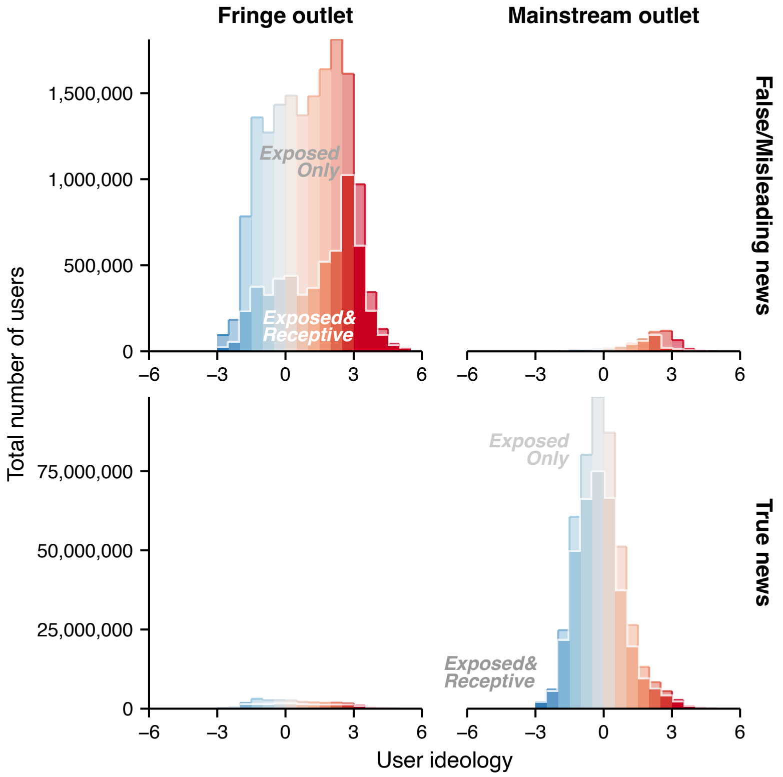

Figure S9: Total number of users exposed and receptive to articles, broken out by news source type. For user ideology, negative (blue) values indicate left-leaning ideology and positive values (red) indicate right-leaning ideology.

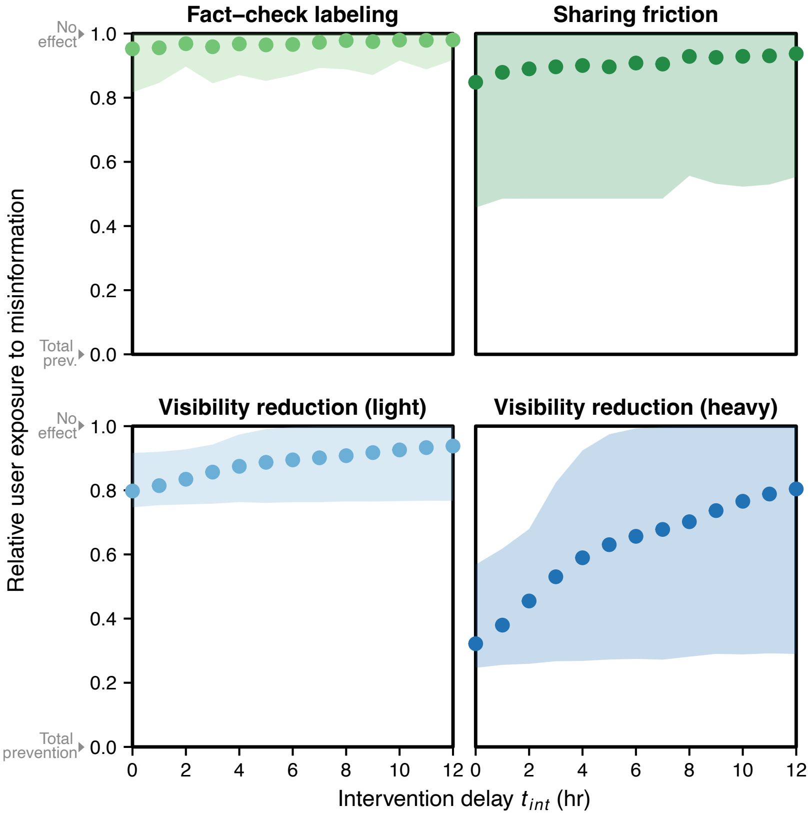

Figure S10: Simulating how intervention method and timing reduce total user exposure in misinformation. Intervention delay t_{int} is the number of hours it takes for an article to be fact-checked and an intervention to be deployed. Fact-check labeling decreases the probability of retweets by 25%, while sharing friction—adding extra steps to the retweet process for tweets sharing flagged material—decreases the probability of retweets by 75%. Visibility reduction is an intervention in which a tweet containing a flagged news article becomes 25% (light) or 75% (heavy) less likely to appear in other user's timelines. Points represent the mean of data-driven simulations across all fake news articles in our data set and each ribbon covers 90% of all simulations. Each article was simulated 30 times for a given intervention type and delay.