Identifying the Causal Effect of Twitter's Interventions on the Spread of Misinformation

Swapneel Mehta[ac1], **James Bisbee**[bc], **Zeve Sanderson**[ac], **Richard Bonneau**[ac], **Joshua A. Tucker**[ac], **and Jonathan Nagler**[ac]

[a]New York University; [b]Vanderbilt University; [c]NYU Center for Social Media and Politics

This manuscript was compiled on August 7, 2024

With misinformation on social media introducing new challenges for democratic politics, a growing body of literature aims to measure the efficacy of interventions to mitigate its harms. While observational data enable scholars to report the association between interventions and misinformation diffusion, estimating the causal effects of interventions has remained elusive, especially across multiple platforms. Here, we estimate the causal effect of interventions deployed by Twitter to limit the misinformation shared by former President Donald J. Trump via his Twitter account in the months following the United States Presidential election on November 3, 2020 leading up to the Capitol insurrection on January 6, 2021. Specifically, we adopt a trajectory balancing method (1) to compare the reach of tweets that received a warning label to those that did not.[a] Importantly, we exploit the fact that it took Twitter several minutes to apply a warning label to identify tweets that had similar content and were similarly viral, but then diverged in popularity after the intervention. Our outcomes of interest are the number of likes and retweets that intervened tweets received on Twitter, as well as the number of posts referencing or linking to these tweets on other social media platforms, including Reddit, Facebook, and Instagram. Our findings reveal that warning labels cause a "Streisand" effect, whereby the act of intervention increases attention to the offending content, although this effect is milder than previously estimated on Twitter.

misinformation | causal identification | social media policy

A growing body of literature across disciplines has provided evidence that the diffusion of online misinformation presents challenges for democratic politics (3–5). These challenges are especially acute during election periods when both the incentives for producing and the harmful effects of consuming political misinformation increase (6–8). Against this backdrop, there has been increased interest across academia (9–11), civil society(12, 13), and industry (14, 15) in the development of interventions aimed at mitigating misinformation, as well as the rigorous measurement of their efficacy.

However, this research has been hampered by difficulty in estimating the causal effect of platform interventions. On the one hand, experimental research provides causal evidence for the effects of a range of strategies, such as content labels (16–18), accuracy nudges (19, 20), inoculation (21, 22), credibility labels (23, 24), and fact-checks (25, 26). However, these studies cannot evaluate intervention efficacy on real-world behavior and at the scale of platforms. On the other hand, research that utilizes large scale observational data captures the diffusion of misinformation both within (27–30) and across platforms (31–33), but studies of platform interventions have struggled to isolate causal effects due to the dissimilarities between the moderated and unmoderated content.

In this paper, we apply sophisticated causal inference methods to a unique observational setting that allows us to recover causally identified estimates. Specifically, we measure the effect of two popular intervention strategies utilized by Twitter during the 2020 election: 1) a warning label attached to specific tweets (soft intervention); and 2) an overlayed warning coupled with the removal of the ability to like or retweet the tweet (hard intervention). Examples of both interventions are displayed in Figure 1. According to Twitter, soft interventions are designed to balance mitigating harmful election misinformation with promoting democratic discourse, while hard interventions are reserved for content that poses more severe risks, such as calls to violence or undermining election results (34, 35). While we recognize that all social media platforms evolve rapidly, with Twitter being a particularly salient recent example under the new leadership of Elon Musk, these two modalities of interventions represent widely used strategies across a variety of contexts (36, 37), broadening the generalizability of our findings.

We focus specifically on the impact of these interventions on former President Donald Trump's tweets posted between November 1, 2020 and January 6, 2021, when he was suspended from Twitter due to "due to the risk of further incitement of violence" related to the Capitol Insurrection.[b] This period is valuable both because it provides us with a discrete set of high-salience misinformation cues and examples of the intervention strategies designed to combat them, and because it is of sub-

[b]https://blog.twitter.com/en_us/topics/company/2020/suspension

Significance Statement

Misinformation has emerged as a key challenge for democratic politics, particularly during election periods. In response, social media platforms have developed and deployed interventions aimed at mitigating the diffusion and harmful effects of misinformation. However, there is limited causal evidence of intervention efficacy using real-world observational data, undermining our ability to identify the extent to which platform strategies produce their intended outcomes. Here, we measure the causal effect of Twitter's soft and hard interventions on the within- and cross-platform spread of Trump's misinformation tweets during the 2020 election. We find evidence of a Streisand effect on Twitter itself, with labeled tweets experiencing more engagement than similar unlabeled tweets. We identify the same messages across Facebook, Instagram, and Reddit, finding that Twitter's interventions have heterogeneous effects on the spread of the same messages across other platforms.

All of the authors designed the research. SM and JB carried out the statistical analyses. SM and JB wrote the first draft of the paper. All of the authors contributed to the revision of the manuscript.

The authors declare no conflicts of interest.

[1]To whom correspondence should be addressed. E-mail: swapneel.mehta@nyu.edu

Fig. 1. Examples of soft (top panel) and hard interventions (bottom panel).

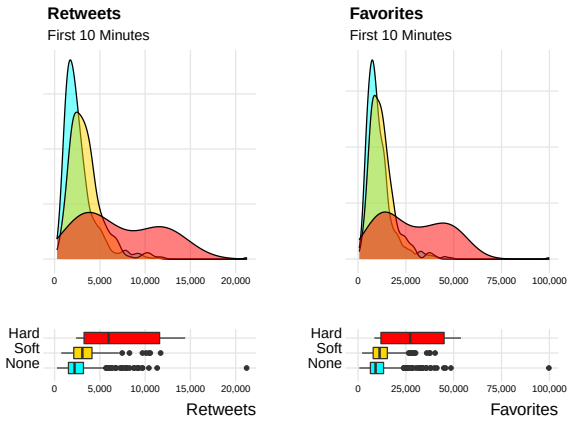

Fig. 2. Evidence of endogeneity. Tweets that were censored (soft interventions in gold, hard interventions in red) were significantly more popular in terms of both retweets (left panel) and favorites (right panel) in the first ten minutes after being posted. Under the assumption that this was prior to the intervention itself, these results suggest that a naive comparison of censored and uncensored tweet popularity will be biased toward a Streisand Effect.

stantive interest to evaluate Twitter's efficacy in combating misinformation that contributed to a decline in perceptions of democratic legitimacy (38). Previous studies have measured the differential spread of Trump's tweets with and without interventions through simple comparisons of diffusion and engagement (39–41). These studies provided descriptive evidence of a so-called "Streisand effect" for engagement on Twitter — a backfire effect in which attempts at censorship actually drive further engagement (42). (43) also provide evidence that messages blocked on Twitter spread further on other platforms, though the authors are not able to determine whether these patterns indicate a Streisand effect or a replacement effect, in which actors spread messages limited on one platform across others that have different content moderation policies (44). However, as we demonstrate in Figure 2, the tweets which were censored received significantly more attention (in the form of engagement on Twitter) than those which were not censored, even in the first 10 minutes after being posted, a period during which (we argue) Twitter could not yet have applied the interventions. Based on this descriptive plot, a naive comparison by intervention type will conflate the virality of intervened tweets with the intervention itself, producing a spurious association biased toward a 'Streisand Effect.

To overcome this challenge, we utilize recent innovations in generalized difference-in-differences methods to create synthetic control groups against which we compare treated groups to achieve causally identified estimates: the treatment groups are composed of tweets that received either a hard or soft intervention, and the control groups are composed of tweets that were otherwise similar but did not receive an intervention.

Existing literature provides competing expectations as to whether these intervention strategies might limit the spread of misinformation. While past experimental studies have found labels reduce the spread of misinformation (45–47), others have reported null (48), mixed (49), or unintended effects (50). And while previous studies using platform data have largely found soft interventions to be associated with increased spread (39, 40), little is known about the causal effects of warning labels on user behavior due to challenges for causal inference in studies utilizing observational data.

By removing the ability to engage with tweets, hard interventions will necessarily limit the tweets' spread on Twitter itself, but users exist in a multi-platform information ecosystem where messages blocked on one platform can migrate to other platforms with divergent moderation practices. Previous work on more severe interventions, such as removals of subreddits and Twitter accounts, have shown that users associated with those moderated communities improve their behavior on that platform (51, 52), but there is evidence that moderation practices on one platform can lead to worsening behavior on other platforms (44). These studies provide evidence of a substitution effect at the user level, whereas here we aim to understand the multi-platform diffusion of specific messages that had been moderated on one platform.

Building on this literature, we are substantively interested in two research questions. First, what was the causal effect of soft interventions on the tweet's popularity on Twitter itself? We compare the numbers of retweets and favorites ("likes") of treated tweets (i.e., those receiving the soft intervention) and control tweets (i.e., those that were otherwise similar but did not receive a soft intervention) over the first 24 hours following the soft intervention, finding consistent evidence of a modest but statistically significant Streisand Effect.

Second, what was the effect of both soft and hard interventions on the tweet's popularity on other social media platforms? Here we identify posts on Reddit, Instagram, and Facebook that reference Donald Trump's tweets and compare the number of posts referencing treated tweets versus control tweets. We find that, as a result of soft interventions, the number of public posts referencing the treated tweets on Facebook and Instagram decreased, while the number of posts that reference these tweets increased on Reddit. Conversely, we find no evidence of a statistically significant effect of Twitter's hard interventions on these other platforms, although this may partly be due to greater measurement error in the data (See the Limitations section in the SI for a more detailed discussion of the generalizability of our results across other platforms).

Taken together, our analysis provides causally identified

estimates of the real world impact of two widely used interventions by social media platforms in an attempt to combat misinformation. In contrast with previous research on this topic, we find milder evidence of a Streisand Effect on Twitter and mixed evidence on other platforms. At best, Twitter's interventions had no effect on the spread of Donald Trump's tweets about election fraud. At worst, their efforts contributed to the increased visibility of the very tweets that they had been hoping to limit via interventions. Importantly, in none of our analyses do we find a consistent, statistically significant causally identified decline in the popularity of the intervened-upon tweets. Our results highlight the difficulty in content moderation policies achieving their desired effect of reducing the spread of socially harmful content.

Measuring Engagement with Donald Trump's Tweets

We identified every tweet written by President Donald Trump between November 1st, 2020 and January 1st, 2021, and recorded whether it was "treated" with Twitter's soft intervention, with Twitter's hard intervention, or was part of the untreated "control" group. Over this period, Trump posted 1,306 total tweets (39), of which 303 were flagged with Twitter's soft intervention and 16 were flagged with the hard intervention.

Our outcomes of interest are 1) the tweet's popularity on Twitter itself, measured as either favorites or retweets of the original tweet, and 2) the tweet's spread on other social media platforms, measured as the number of posts containing links to, or the text of, the original tweet. To measure the first set of outcomes on Twitter, we leverage Twitter's "Decahose", a service that provides researchers access to a streaming 10% random sample of every tweet written each day. Any time someone retweets one of Trump's tweets, we extract the number of favorites and retweets his tweet had *at the moment it was retweeted*. While only 10% of all tweets are available, Trump's popularity ensures that we have a detailed set of snapshots of each of his tweets, especially in the first few hours after they are posted when even a 10% sample of Twitter returns second-by-second estimates of a tweet's popularity.[c] We standardize these snapshots into 6-20 second bins, interpolating where necessary to recover a rich time-series characterization of the trajectory of a tweet's popularity on Twitter.

Our approach to measuring a tweet's popularity on other platforms is similar in that we look for references to the set of Trump tweets that are cited on Facebook, Reddit, and Instagram. Unlike the Twitter outcomes of favorites and retweets, here we are only interested in whether someone posted a link to the Trump tweet. As such, we search for either the original URL to the tweet, or copied text from the original tweet, among posts on Facebook, Instagram (retrieved via the Crowdtangle (53)) and Reddit (retrieved via the Pushshift API (54)). As with the Twitter data, we standardize these snapshots into 10-minute bins and interpolate to create a smooth time-series measure of a tweet's popularity across these other social media platforms.

As illustrated above in Figure 2, a naive comparison between control and intervened tweets would conflate the causal effect of the intervention with the selection effect of more inflammatory tweets being both more likely to receive an intervention and being more viral. The granularity of these time-series vectors allow us to implement a method for causal identification in a generalized difference-in-differences setting in which we re-weight the non-intervened tweets such that their pre-intervention outcome vectors match those of the intervened tweets (55). By conditioning on the pre-intervention outcomes, we theoretically condition on all qualities that determine a tweet's virality aside from the intervention itself, meaning that the re-weighted controls are a valid counterfactual for what the intervened tweets would have looked like without the intervention. Insofar as there may remain some differences in the topic(s) of discussion, the number of reshares from users with large followings, and the toxicity of the content of the tweet (measured through Google's Perspective API [d]) even after re-weighting the control tweets, we ensure further balance on these additional covariates of interest.[e]

On Twitter, suppression efforts backfired

We begin by summarizing the effectiveness of Twitter's attempts to suppress the spread of misinformation on Twitter itself. Our results compare the observed popularity of the tweets which received a soft intervention with a weighted combination of "control" tweets which were not flagged, but are otherwise similar to the treated tweets in terms of semantic content and initial virality in the first ten minutes after the tweet was posted. Figure 3 plots the Average Treatment Effect on the Treated (ATT) estimates for retweets (top panel) and favorites (bottom panel), revealing a modest but clear increase in popularity that we attribute to the intervention itself.[f] As illustrated in the top row of plots, within a day of posting, Twitter's soft interventions caused an additional 2,343 retweets, compared to the weighted counterfactual, an increase of 4.7%. Importantly, this is about five times smaller an effect than what we would conclude were we to naively compare the raw retweet counts of untreated and treated tweets (12,049 additional retweets or a 24.4% increase). Similar conclusions are found for favorites, as displayed in the bottom row of Figure 3. Here we document an increase of 3,428 (1.6%) for the weighted controls, compared to an increase of 23,818 (11.3%) for the raw data, although the ATT estimates are not statistically significant at the 95% threshold.

Importantly, by magnifying our data to the first 10 minutes, we confirm that our matching solution achieves almost perfect balance over this period, as illustrated by the indistinguishable overlap between the treated and weighted control trajectories. Yet despite this superior balance and the substantially attenuated treatment effects, we underscore that our findings are nevertheless consistent with a Streisand effect in

[c] See SI Section XX for a more detailed description of how we collected these data.

[d] https://perspectiveapi.com

[e] An additional challenge to implementing the trajectory balancing method is that we don't observe the precise moment when the soft intervention was applied. For hard interventions, it is clear when the intervention took place since there is no change in engagement with the tweet after the hard intervention removes the ability of users to interact with the tweet. However, for soft interventions it is not as clear. In order to ensure our results are robust to the time of intervention, we conducted a sensitivity analysis assuming generous periods starting at 2 minutes and ranging up to 15 hours that the soft intervention may have taken place within, creating an interval before which we ensure balance in the groups, and after which we estimate the average treatment effect on the treated (ATT). We choose to present results at a relatively conservative soft intervention timing of 10 minutes and share the sensitivity checks in the SI.

[f] Twitter has not provided the exact timing of each intervention but former CEO Jack Dorsey stated in congressional testimony that they applied interventions generally between 5 and 30 mins. of tweets being posted. Therefore, in our analysis we run a sensitivity check of potential intervention timings starting from a minute since a tweet was posted up to an hour and a half of a tweet being posted to ensure our results are robust to the actual choice of intervention timing. We expect that allowing for this wide a window for interventions to have been applied accounts for a large number of actual interventions by Twitter.

which Twitter's attempts to soft intervene on Trump's tweets containing misinformation backfired. Both metrics exhibit the same pattern, importantly moving in the opposite direction of the presumed goal of the intervention: to reduce the spread of misinformation.

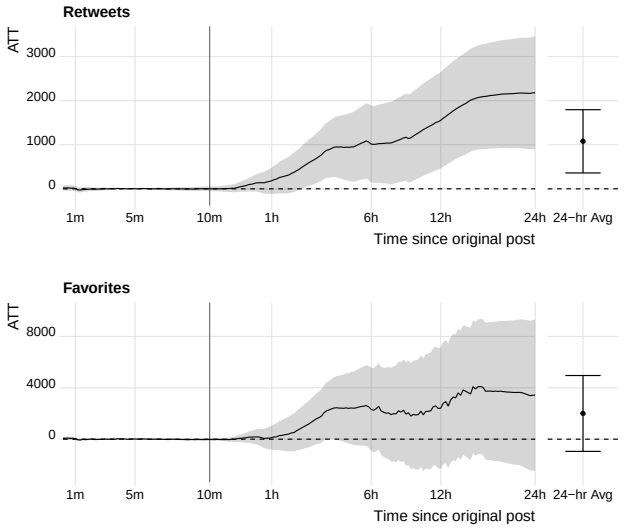

Fig. 3. ATT plots of the effect of soft interventions on retweets (top) and favorites (bottom). 95% confidence intervals estimated shaded in gray. X-axes are not to scale to highlight the pre-intervention balance results.

On other platforms, the results are mixed

Do Twitter's efforts similarly backfire on other platforms? On the one hand, the generic intuition of a Streisand Effect is that any effort to suppress information will draw more attention to it. On the other hand, Twitter's policy decisions may be less salient or relevant for users of other social media platforms, meaning that the backfire dynamic may not materialize in a similar fashion. Instead, the limited access to intervened tweets – particularly in the case of the hard interventions – might actually work as intended, reducing the spread of misinformation on other platforms where the attention-generating act of censorship is irrelevant. To investigate these expectations, we re-run our analyses on Facebook, Instagram, and Reddit, replacing the outcome measures described above (Twitter-specific engagement metrics consisting of likes and retweets) with a count of the number of posts containing a link to one of Trump's tweets was shared to the platform.

Soft interventions. We start with an analysis of the spread of tweets that received soft interventions on these platforms, again restricting attention to the first 24 hours after the tweet was written. We compare the flagged tweets' popularity to a set of unflagged tweets that were otherwise similarly popular on these alternative platforms within the first ten minutes of being posted. We summarize the overall average ATT for the 24-hour period in Figure 4, finding no statistically significant evidence of a treatment effect on any of the platforms. This conclusion is particularly noteworthy for Facebook and Instagram, where the estimates are not only null but also negative, suggesting that concerns about a Streisand Effect on the two most widely used social media platforms in the United States are overblown.[g]

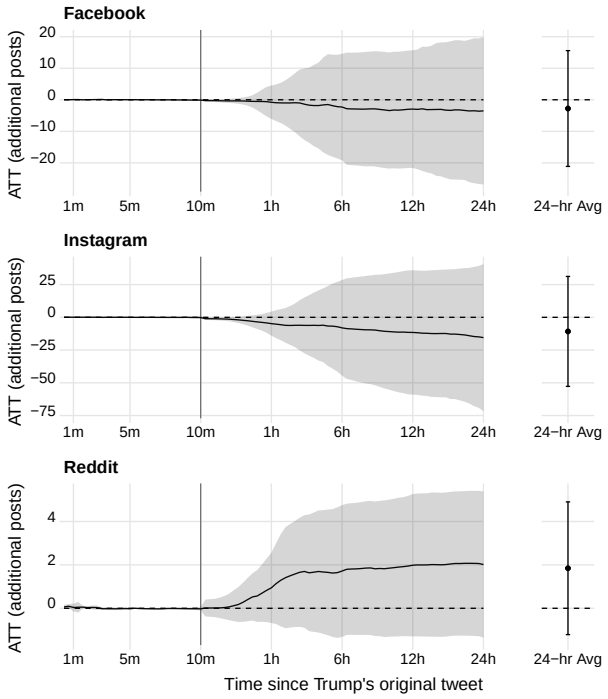

Fig. 4. ATT plots of the effect of soft interventions on Instagram (top, overall average posts for any Trump tweet = 11.3) Facebook (middle, average posts for any Trump tweet = 34.34) and Reddit (bottom, average posts for any Trump tweet = 9.5). 95% confidence intervals estimated shaded in gray. X-axes are not to scale to highlight the pre-intervention balance results.

Hard interventions. Hard interventions blocked all engagement with Trump's tweets on Twitter, rendering an empirical analysis of their effect on that platform pointless.[h] However, hard interventions do not limit the spread of these tweets on other platforms. This allows us to measure the trajectory of the spread of Trump's tweets outside Twitter as the cumulative number of posts mentioning each tweet on Instagram, Reddit, and Facebook respectively. Additionally, the timing of the hard intervention is known to us since we can directly observe the (binned) moment at which the tweet stopped receiving further engagement on Twitter.

With only 16 total hard interventions in our data (only 3 of which were shared on Facebook, 5 of which were shared on Instagram, and 8 of which were shared on Reddit), we are unable to calculate standard errors directly. To capture our confidence, we implement the trajectory balancing method using an exact test in which we compare the observed ATT for hard intervened tweets to 200 placebo tests where we pretend that a control tweet is actually intervened upon. Specifically, we calculate the observed ATT using the trajectory balancing method described above, and compare it to 200 placebo ATTs

[g] In the case of Reddit, we are also able to observe some of the posts that link to Trump's tweets in their entirety. We present results from analyzing the normalized sentiment of the Reddit posts in a corresponding SI Section describing a sensitivity check of soft and hard interventions outside of Twitter, finding that those Reddit posts which do link to Trump's censored tweets are less positive than similar posts linking to Trump's uncensored tweets.

[h] Every hard intervened tweet can no longer accumulate retweets or favorites after the intervention, meaning that these outcomes are no longer valid proxies for the tweet's popularity on Twitter.

calculated by selecting one of the control tweets at random and re-calculating the ATT for it, as though it was treated. These 200 placebo ATTs constitute the null distribution, against which we compare the observed ATT for each hard intervened tweet, expressing confidence in a positive or negative effect in terms of the proportion of placebo ATTs that are less than or greater than the hard intervened tweet. In Figure 5, we combined all treated tweets and balance on the pre-intervention vector for the earliest value of t_0.[i] The null distributions are depicted in gray and the observed ATT in red. We plot the pre-intervention and post-intervention results separately to facilitate visual inspection of the pre-intervention balance, highlighting the much poorer performance on these other platforms compared to Twitter, especially for Reddit and Instagram.

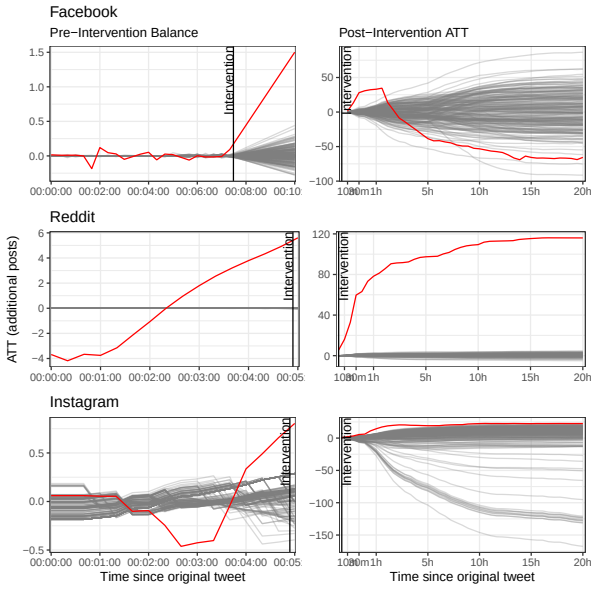

Fig. 5. Observed ATT trajectories for hard intervened tweets on Facebook (top), Reddit (middle), and Instagram (bottom), compared against a null distribution of 200 simulations in which 30 control tweets are defined as "treated" and the resulting ATT trajectory is recorded. Plots are separated by pre-intervention balance (left column, used to check identifying assumptions) and post-intervention ATT estimates (right column). All treated tweets are combined, and pre-intervention balance is calculated on earliest known intervention. Only 3 of 16 hard intervened tweets were found on Facebook, 8 on Reddit, and 4 on Instagram. Tweet-by-tweet results presented in the SI, section XX.

Overall, the results are mixed. While there is striking evidence of a statistically and substantively significant Streisand Effect on Reddit (the 8 hard intervened tweets were posts roughly 100 more times than similar non-intervened tweets), we find the worst pre-intervention balance on this platform. Furthermore, the imbalance is in an anti-conservative direction, suggesting that the trajectory balancing algorithm failed to recover a suitable set of counterfactual tweets on this platform and raising the possibility that the estimated "effect" is confounded by the fact that these 8 tweets were also more popular for reasons unrelated to the intervention. Imbalance is also visible on Facebook and Instagram, although is of smaller magnitude and of varying signs over the pre-intervention period.

[i] We include the tweet-by-tweet estimates in the Supporting Information, Section on Sensitivity Checks.

The post-intervention ATT estimates on these platforms are similarly mixed, with Facebook finding evidence of a short-run Streisand Effect in the first hour, followed by a suppressive effect after the fifth hour. Instagram exhibits a small but positive effect consistent with the Streisand Effect.

Discussion

In particular, our analysis of soft interventions on Twitter indicates that the bulk of the "Streisand" effect occurs within the first few hours of applying the misinformation warning. Importantly, we fail to document statistically significant evidence of a similar backfire dynamic for soft interventions on other social media platforms, although we emphasize that the estimates are positive and substantively significant for Reddit and Instagram. Our conclusions about hard interventions – which are only measurable on these other platforms – are similarly mixed. Yet importantly, none of our analyses suggests that the interventions actually worked as intended by reducing engagement with content that violated Twitter's policies on misinformation.

The mixed evidence on other platforms underscores the importance of greater data availability for researchers. Our analysis of Twitter was possible thanks to access to detailed metrics via the Decahose and Twitter's API, both of which have subsequently been taken down. Yet on Facebook, Instagram, and Reddit we were limited to only measuring engagement as the number of total posts that linked to Trump's misinformation tweets. Unrestricted access to richer data would allow us to more carefully test engagement in the form of comments, likes, and reactions, potentially revealing stronger evidence of a backfire effect by reducing noise.

This work quantifies the impact of certain types of platform policies at a time when major social platforms are laying off teams working on reducing the spread of misleading information on the platform, and when lawmakers are inviting congressional testimony to review the impact of censorship of online discourse during elections. Our finding that soft interventions on Twitter contribute to an increase in popularity of the tweet underscores the need to better plan and deploy interventions if the intention is to limit its popularity. Yet the information required to improve on the design and deployment of these methods for combating misinformation – the data necessary to optimize research-informed policies – is precisely the information whose access is being restricted. Donald Trump's lies about the 2020 presidential election are the sum of all fears about how social media-enabled misinformation can damage democracy, and yet this paper's research is itself no longer possible three years after this watershed moment.

Materials and Methods

Data creation. We identified every tweet written by President Donald Trump between November 1st, 2020 and January 1st, 2021, and recorded whether it was "treated" with either Twitter's soft interventions, hard interventions, or was in the control group. We then assembled a rich time series dataset of each tweet's popularity on the platform by leveraging Twitter's "Decahose", a service that provides researchers access to a streaming 10% sample of every tweet written each day. The random nature of this sample ensures that we are able to observe tweets written by other users who retweet one of

Trump's tweets. Each retweet includes a measure of how many total retweets and favorites Trump's original tweet had gathered at the time of the creation of the retweet, providing a snapshot of Trump's tweet's popularity that we interpolate between to create a trajectory of the tweet's popularity.

Platform	Hard Interventions	Soft Interventions	No Interventions
Twitter	16	303	614
Facebook	3	136	263
Reddit	8	140	321
Instagram	4	14	64

Table 1. Number of observations by intervention type (columns) and platform (rows).

Constructing valid counterfactuals. Using data collected about intervention labels, we are able to group tweets into treatment and control groups. However, the fact that Twitter does not randomly select tweets for interventions makes a simple comparison of the future spread of intervened and non-intervened tweets unlikely to return a valid causally identified estimate of intervention due to selection bias and reverse causality.

For instance, differences in the distribution of early (up to ten minutes from tweet creation)[j] retweets and favorites indicate that tweets which were eventually intervened upon by Twitter were significantly more viral in the first minutes after being written than those tweets that were not subjected to interventions. Unintervened tweets amassed an average of 2,800 retweets and 11,383 favorites in their first 10 minutes after posting, while soft-intervened tweets had an average of 3,495 retweets and 12,726 favorites, and hard intervened tweets had an average of 7,321 retweets and 28,939 favorites. Kolmogorov–Smirnov tests of these distributions confirm that these differences are highly significant, bolstering the concern that a naive comparison would inflate the magnitude of a Streisand effect of the intervention.

To overcome this source of bias, we adopt a matching solution in which we pair intervened tweets with tweets written by Donald Trump in the same time period that are otherwise identical in terms of the topics they discuss, their semantic content, and – most importantly – in terms of their pre-intervention popularity. Borrowing notation from (1), consider a number of tweets $i \in N$ whose popularity Y_{it} is observed at time period t. Some of these tweets are intervened upon by Twitter at time T_0 and fall into group $G_i = 1$, while others never are, and are denoted with $G_i = 0$. Finally, let $D_{it} \in [0, 1]$ be an indicator for an intervention as follows:

$$D_{it} = \left\{ \begin{array}{ll} 1 & \text{when } G_i = 1 \text{ and } t > T_0 \\ 0 & \text{otw} \end{array} \right\}$$

In the potential outcomes framework, Y^1_{it} represents the popularity of tweet i at time t when the tweet is intervened upon ($D_{it} = 1$), and Y^0_{it} represents the same tweet's popularity at the same time when it is not intervened upon ($D_{it} = 0$). Our theoretical quantity of interest is the causal effect of interventions, $\tau_{it} = Y^1_{it} - Y^0_{it}$, which we operationalize with the Average Treatment Effect on the Treated (ATT): $ATT_t = \mathbb{E}[\tau_{it}|G_i = 1]$ for $t > T_0$.

Since the fundamental problem of causal inference means we can never observe both Y^1_{it} and Y^0_{it}, we must appeal to an assumption of conditional independence, which we express in terms of all three theorized quantities described above. Formally,

$$Y^0_{it} G_i | \mathbf{X}_i, \mathbf{Y}_{i,t<T_0}, T \qquad \forall t > T_0 \qquad [1]$$

where $\mathbf{X}_i$ is a vector of time-invariant covariates (topic model loadings of the tweet, NLP classifier probabilities for toxicity, follower counts from users resharing the tweet), $\mathbf{Y}_{i,t<T_0}$ is a vector of pre-intervention measures of popularity (favorites and retweets on Twitter; shares on other social media platforms), and T is a constraint on the period in which we compare intervened and unintervened tweets.

This expression states that conditioning on pre-intervention outcomes and time-invariant characteristics implies that treatment assignment is as-if random, allowing us to consider the conditioned observed values of $\hat{\mathbf{Y}}^0_t$ as valid counterfactuals for the observed $\mathbf{Y}^1_t$ in periods $t > T_0$. Substantively, this means we are focusing the comparison between the popularity of intervened tweets and the popularity of unintervened tweets that were written around the same time (in our setting, we require T to either be tweets written by Trump after November 1st, 2020 or those written by Trump within a week of the treated tweets), are about the same topics and written with similar language ($\mathbf{X}_i$), and had the same trajectories of popularity between when they were written and when the intervention occurred ($\mathbf{Y}_{i,t<T_0}$). Technically, we recover these valid counterfactuals via a method referred to as trajectory balancing (1) which calculates a set of weights w that satisfy at minimum

$$\frac{1}{N_{tr}} \sum_{G_i=1} \mathbf{Y}_{i,t<T_0} = \sum_{G_i=0} w_i \mathbf{Y}_{i,t<T_0} \qquad [2]$$

where N_{tr} is the number of tweets in the treated group $G_i = 1$. Armed with these weights, we can recover the estimand of interest via:

$$\widehat{ATT}_t = \frac{1}{N_{tr}} \sum_{G_i=1} Y_{it} - \sum_{G_i=0} w_i Y_{it} \qquad [3]$$

Recognizing that exact balancing weights may not exist, particularly where there are few control units and/or many pre-intervention periods, (1) propose an approximate balance solution that balances on the first P principal components of the pre-intervention outcome matrix

$$\mathbf{Y}_{t<T_0} = \begin{bmatrix} Y_{1,t_1} & \dots & Y_{1,T_0} \\ \vdots & \ddots & \vdots \\ Y_{N,t_1} & \dots & Y_{N,T_0} \end{bmatrix}$$

with P chosen to minimize bias. Importantly, this framing clarifies how higher order moments of the pre-intervention outcome vectors can be accommodated in the calculation of the weights via kernel expansion, denoted $\phi(\cdot)$. As the authors clarify, the choice of kernel requires only that the post-intervention potential outcomes Y^0 are linear in $\phi(\mathbf{Y}^0_{i,t<T_0})$, but in applied settings propose the Gaussian kernel: $k(Y_i, Y_j) = exp(-||Y_i - Y_j||^2/h$ where h is the bandwidth

[j] We assume that Twitter's interventions couldn't have occurred faster than ten minutes of the original tweet being written based on conversations with Twitter and the empirical evidence of hard interventions which clearly reveal the fastest Twitter was able to intervene. None of the hard interventions in our data occurred faster than 10 minutes after the original tweet was written, with a median lag of approximately 30 minutes.

and Y_i and Y_j are now two vectors of pre-intervention outcomes for tweets i and j. Functionally, each kernel K_i is combined into a kernel matrix $\mathbf{K}$ which replaces the outcome matrix $\mathbf{Y}_{t<T_0}$ described above. Substantively, this feature expansion step lends credibility to the argument that we are estimating weights that balance not only on the first moment of these vectors (the period-by-period mean) but also higher order moments (variance, skewness, and so on subject to P) – in other words, the pre-intervention trajectories of the outcome of interest. A much more detailed explanation of these methods can be found in (1) and in (56).

Author Affiliations. Swapneel Mehta is a Postdoctoral associate at Boston University and Massachusetts Institute of Technology, formerly a Data Science Ph.D. candidate at New York University when pursuing this research.

James Bisbee is Assistant Professor of Political Science at Vanderbilt University and a Faculty Research Affiliate of the NYU Center for Social Media and Politics.

Zeve Sanderson is...

Richard Bonneau is...

Joshua A. Tucker is Professor of Politics at New York University and Co-Director of the NYU Center for Social Media and Politics.

Jonathan Nagler is...

ACKNOWLEDGMENTS. We acknowledge financial support from... We thank...

1. Hazlett C, Xu Y (2018) Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data. *Available at SSRN 3214231*.
2. Liu L, Wang Y, Xu Y (2022) A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*.
3. Watts DJ, Rothschild DM, Mobius M (2021) Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences* 118(15).
4. Lewandowsky S, Ecker UK, Cook J (2017) Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition* 6(4):353–369.
5. Persily N, Tucker JA (2020) *Introduction*, SSRC Anxieties of Democracy, eds. Persily N, Tucker JA. (Cambridge University Press), p. 1–9.
6. Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2):211–36.
7. McKay S, Tenove C (2020) Disinformation as a threat to deliberative democracy. *Political Research Quarterly* 74(3):703–717.
8. Cantarella M, Fraccaroli N, Volpe R (2023) Does fake news affect voting behaviour? *Research Policy* 52(1):104628.
9. Wittenberg C, Berinsky AJ (2020) *Misinformation and Its Correction*, SSRC Anxieties of Democracy, eds. Persily N, Tucker JA. (Cambridge University Press), p. 163–198.
10. van der Linden S (2022) Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine* 28(3):460–467.
11. Lewandowsky S, Cook J, Lombardi D (2020) Debunking handbook 2020.
12. Green Y, et al. (2023) Evidence-based misinformation interventions: Challenges and ...
13. Studdart A (year?) Building civil society capacity to mitigate and counter disinformation.
14. Meta (2021) Combating misinformation.
15. Google (year?) Info interventions.
16. Grady RH, Ditto PH, Loftus EF (2021) Nevertheless, partisanship persisted: Fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications* 6(1).
17. Ecker UK, Lewandowsky S, Tang DT (2010) Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory Cognition* 38(8):1087–1100.
18. Peacock C, Masullo GM, Stroud NJ (2020) The effect of news labels on perceived credibility. *Journalism* 23(2):301–319.
19. Pennycook G, Rand DG (2021) Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*.
20. Pennycook G, et al. (2021) Shifting attention to accuracy can reduce misinformation online. *Nature* 592(7855):590–595.
21. Roozenbeek J, van der Linden S, Goldberg B, Rathje S, Lewandowsky S (2022) Psychological inoculation improves resilience against misinformation on social media. *Science Advances* 8(34).
22. Traberg CS, Roozenbeek J, van der Linden S (2022) Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science* 700(1):136–151.
23. Pennycook G, Rand DG (2019) Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116(7):2521–2526.
24. Aslett K, Guess AM, Bonneau R, Nagler J, Tucker JA (2022) News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances* 8(18).
25. Carey JM, et al. (2022) The ephemeral effects of fact-checks on covid-19 misperceptions in the united states, great britain and canada. *Nature Human Behaviour* 6(2):236–243.
26. Ecker UK, O'Reilly Z, Reid JS, Chang EP (2019) The effectiveness of short-format refutational fact-checks. *British Journal of Psychology* 111(1):36–54.
27. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science* 359(6380):1146–1151.
28. Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on twitter during the 2016 u.s. presidential election. *Science* 363(6425):374–378.
29. Del Vicario M, et al. (2016) The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113(3):554–559.
30. Guess A, Nagler J, Tucker J (2019) Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances* 5(1).
31. Ng LH, Cruickshank IJ, Carley KM (2022) Cross-platform information spread during the january 6th capitol riots. *Social Network Analysis and Mining* 12(1).
32. Aiyappa R, et al. (2023) A multi-platform collection of social media posts about the 2022 u.s. midterm elections.
33. Golovchenko Y, Buntain C, Eady G, Brown MA, Tucker JA (2020) Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 u.s. presidential election. *The International Journal of Press/Politics* 25(3):357–389.
34. Twitter (2020) An update on our work around the 2020 us elections.
35. Twitter (2020) Additional steps we're taking ahead of the 2020 us election.
36. Rosen G, Lyons T (2020) Remove, reduce, inform: New steps to manage problematic content.
37. TikTok (2019) An update on our work to counter misinformation.
38. Levendusky M, Pasek J, Jamieson KH (2023) "stop the steal": Effects of the assaults on electoral and democratic legitimacy. *Democracy amid Crises* p. 301–334.
39. Sanderson Z, Brown MA, Bonneau R, Nagler J, Tucker JA (2021) Twitter flagged donald trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review.*
40. Zannettou S (2021) "i won the election!": An empirical analysis of soft moderation interventions on twitter.
41. Papakyriakopoulos O, Goodman E (2022) The impact of twitter labels on misinformation spread and user engagement: Lessons from trump's election tweets. *Proceedings of the ACM Web Conference 2022*.
42. Jansen S, Martin B (2015) The streisand effect and censorship backfire. *International Journal of Communication* 9(0).
43. Sanderson Z, Brown MA, Bonneau R, Nagler J, Tucker JA (2021) Twitter flagged donald trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review.*
44. Mitts T, Pisharody N, Shapiro J (2022) Removal of anti-vaccine content impacts social media discourse. *14th ACM Web Science Conference 2022*.
45. Epstein Z, et al. (2021) Developing an accuracy-prompt toolkit to reduce covid-19 misinformation online. *Harvard Kennedy School Misinformation Review*.
46. Moravec P, Kim A, Dennis AR (2018) Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *SSRN Electronic Journal*.
47. Yaqub W, Kakhidze O, Brockman ML, Memon N, Patil S (2020) Effects of credibility indicators on social media news sharing intent. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
48. Johnson HM, Seifert CM (1994) Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(6):1420–1436.
49. Grady RH, Ditto PH, Loftus EF (2021) Nevertheless, partisanship persisted: Fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications* 6(1).
50. Pennycook G, Rand DG (2017) The implied truth effect: Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy. *SSRN Electronic Journal*.
51. Trujillo A, Cresci S (2022) Make reddit great again: Assessing community effects of moderation interventions on r / the_donald. *Proceedings of the ACM on Human-Computer Interaction* 6(CSCW2):1–28.
52. Jhaver S, Boylston C, Yang D, Bruckman A (2021) Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–30.
53. Team C (2020) Crowdtangle. *Menlo Park, CA: Facebook*.
54. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J (2020) The pushshift reddit dataset in *Proceedings of the international AAAI conference on web and social media*. Vol. 14, pp. 830–839.
55. Hazlett C, Xu Y (2018) Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data. *SSRN Electronic Journal*.
56. Hazlett C (2020) Kernel balancing. *Statistica Sinica* 30(3):1155–1189.