The Effect of Misinformation Intervention: Evidence from Trump's Tweets and the 2020 Election

Zhuofang Li[1](✉), Jian Cao[2], Nicholas Adams-Cohen[3], and R. Michael Alvarez[1]

[1] California Institute of Technology, Pasadena, USA
zhuofang@caltech.edu
[2] Trinity College of Dublin, Dublin, Ireland
[3] Adobe, Inc., San Jose, USA

Abstract. In this study, we examine the effect of actions of misinformation mitigation. We use three datasets that contain a wide range of misinformation stories during the 2020 election, and we use synthetic controls to examine the causal effect of Twitter's restrictions on Trump's tweets in the 2020 presidential election on the spread of misinformation. We find a nuanced set of results. While it is not always the case that Twitter's actions reduced the subsequent flow of misinformation about the election, we find that in a number of instances content moderation reduced the flow of social media misinformation. We estimate that Twitter's actions, on the universe of tweets we study in our paper, reduced the flow of misinformation on Twitter by approximately 15%.

Keywords: Social media · misinformation · platform intervention · public opinion · Twitter · Election

1 Introduction

Research shows that people use social media platforms like Twitter, Facebook, and YouTube to spread misinformation and conspiracy theories about many different subjects [7]. Recognizing this problem, these platforms have engaged in different approaches to protect their users from misinformation and platform manipulation, for example Twitter's Platform Manipulation efforts.[1] However, recently some states like Florida and Texas have developed policies to block social media platforms from moderating conversations online, especially those that might involve constitutionally-protected political speech.

Much of the concern about the role of social media platforms in the rapid and viral spread of misinformation and conspiratorial ideas has roots in the 2016 American presidential election, with allegations of foreign interference on social media [22]. Other studies showed that the spread and consumption of fake news on social media was widespread among Americans in the 2016 election cycle

[1] See https://transparency.twitter.com/en/reports/platform-manipulation.html.

D. Ceolin et al. (Eds.): MISDOOM 2023, LNCS 14397, pp. 88–102, 2023.
https://doi.org/10.1007/978-3-031-47896-3_7

[6]. Social media platforms developed monitoring and intervention policies in the aftermath of the 2016 election, often with limited public transparency and unknown efficacy.

Detecting misinformation and other undesirable behavior on social media in real-time is difficult, in particular when well-resourced and strategic agents are conducting the behavior [20]. They engage in many strategies to avoid detection, and have strong incentives to hide their activities and identities. In response, social media platforms use many approaches to detect, mitigate, and prevent the spread of false and misleading information. However, research is mixed about whether the strategies used by social media platforms are effective at preventing the spread of misinformation [9,10,15,17,19,21,23].

In this paper, we use a unique set of natural experiments that occurred during the 2020 presidential election, employing three unique datasets described below in the Data section. In 2020 (as we discuss in the next section), Twitter used various tools to prevent the spread of information in a series of tweets that President Donald Trump posted. These tweets were deemed to violate Twitter's policies about spreading electoral misinformation. We use a synthetic control methodology to develop counterfactuals that allow us to test the efficacy of Twitter's actions on Trump's tweets, allowing us to make causal inferences from the real-world observational data from the 2020 election. Research demonstrates that the synthetic control methodology is a powerful tool for causal inference [1–3]. This is one of the important contributions of our work – showing how synthetic control can help researchers make causal inferences about interventions in social media.

Using this methodology we produce important causal estimates that allow us to study whether Twitter's content moderation actions in the 2020 presidential election were effective. Our results indicate that for the Trump tweets we studied, Twitter's actions can reduce their dissemination. This is not universally the case, as there are situations where misinformation continues to flow after Twitter's content moderation efforts — and where there seems to be little change (one way or the other) after the platform used restrictions or warnings to slow the spread of misinformation. Our results have implications for the current debates about social media platform content moderation which we consider in the paper's Discussion.

2 Twitter's Moderation of Trump's Tweets in 2020

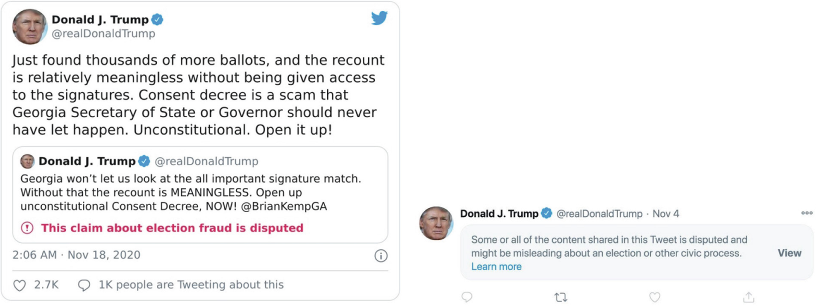
In October 2020, Twitter applied a "Civic Integrity Policy"[2] to prevent use of their platform for electoral or civic interference. Policy violations included misleading information about how to participate in the election, voter suppression or intimidation, and false details about electoral outcomes. Depending on the severity of the violation, Twitter could engage in several actions, including labeling the tweet as misinformation, deleting the message entirely, or locking or permanently suspending the offending account.

[2] https://help.twitter.com/en/rules-and-policies/election-integrity-policy.

One of the most prominent uses of the Civic Integrity Policy in 2020 was for Twitter to use warnings or restrictions on then-President Trump's tweets, as he was disputing the integrity of the election and disseminating misinformation about election fraud. During the period of time we focus on in this study, Twitter mainly applied two types of treatments to Trump's tweets that were determined as violations of the "Civic Integrity Policy":

- Disputed (restricted): Content could be hidden or deleted; the user's ability to reply, retweet, and like the tweet could be turned off; or a label/warning message could be applied to the tweet before it was shared or liked. This treatment was applied frequently between November 4th, 2020 to November 7th, 2020.
- Disputed (not restricted): Content was visible, and users can reply, retweet, or like the tweet; and a warning message was applied to the tweet. This treatment appears throughout the study period.

On January 8, 2021, Twitter suspended @realDonaldTrump, at which time the account had approximately 88.7 million followers.



Donald J. Trump @realDonaldTrump

Just found thousands of more ballots, and the recount is relatively meaningless without being given access to the signatures. Consent decree is a scam that Georgia Secretary of State or Governor should never have let happen. Unconstitutional. Open it up!

Donald J. Trump @realDonaldTrump

Georgia won't let us look at the all important signature match. Without that the recount is MEANINGLESS. Open up unconstitutional Consent Decree, NOW! @BrianKempGA

This claim about election fraud is disputed

2:06 AM · Nov 18, 2020

2.7K 1K people are Tweeting about this

Donald J. Trump @realDonaldTrump · Nov 4

Some or all of the content shared in this Tweet is disputed and might be misleading about an election or other civic process. Learn more

(a) (b)

Fig. 1. Examples of Twitter-labeled Trump tweets. Source https://factba.se/topic/deleted-tweets

In Fig. 1 we provide two examples of Trump's tweets, one for each measure. The example in Panel 1a regards allegations being made regarding the election tabulation and post-election auditing in Georgia from November 18, 2020. The example in Panel 1b shows an example of restricted tweet, which was posted on November 4th, 2020.

Twitter's decision to censor and label Trump's tweets in an attempt to prevent the spread of misinformation was highly controversial. Many people, particularly those within the Republican party, launched a backlash against Twitter following their decision to label Trump's tweets as misinformation. Crucially, it remains uncertain whether Twitter's actions worked as intended: did censoring and labeling these tweets in the 2020 election prevent the subsequent spread of election misinformation on Twitter?

We are interested in the effects of Twitter's actions. Labeling Trump's tweets could have two different consequences: it could have operated as (we assume) Twitter desired: suppressing the further spread of misleading information. Or, given the backlash towards Twitter's policy, it could have amplified the spread of misleading information.

In this paper, we use three novel datasets to study the question. The first one is a unique dataset of over 15 million tweets about the election, a real-time collection that started before the November 2020 general election and ended after Twitter suspended Trump's account. The second one is the $ElectionMisinfo$2020 dataset which consists of tweets directly linked to confirmed misinformation stories in the 2020 election. The third one is from Trump's tweet archive, which collected Trump's tweets and showed whether Twitter took action regarding each of those tweets.

Our main findings are nuanced. There is evidence that for some of Trump's tweets, Twitter's actions reduced misinformation. We find that this is in particular the case for a set of tweets that Twitter placed restrictions on early in election 2020. But we also show that the content moderation efforts generally worked in many cases, but did not work in others. In the set of social media conversations about election fraud in the 2020 election that form the basis of our study, we find that Twitter's actions reduced the subsequent flow of election misinformation by approximately 15%.

In the next section of the paper, we connect our research to the theory about how the public receives and processes information, and what happens when attempts are made to suppress the dissemination of political information. These theories guide and shape our hypotheses. Following this, we delve into our data sources, detailing both the collection and preprocessing of tweets, and then outline the methodologies employed to test our hypotheses. We then present our results and conclude by discussing the implications and limitations of our analysis.

3 Does Labelling and Limiting Misinformation Work?

We use public opinion and censorship theory to guide our research. Public opinion theory regards how the public receives, accepts, and processes political information. Assuming that the public acts in a rational manner, they will use information short-cuts to reduce information costs [11]. Rational citizens will not obtain and process all available information, as argued in the theory of public opinion [24], and applied to the reception and processing of social media information. We assume that citizens will follow and process incoming social media information following the "receive-accept-sample" (or RAS) model [4,5,24].

In the RAS model, the citizen receives information (usually from elites), accepts the information (usually filtering it ideologically or by partisanship), and then samples from recently received information when needed (say answering a survey or voting on a ballot measure). The RAS model provides a theoretical framework in which citizens will be selective about information; partisan citizens will receive and accept information from elites with whom they share partisan affiliations. Partisanship is an important heuristic or information shortcut used by citizens [13,14,16].

However, on social media platforms like Twitter, information is not necessarily passed from a partisan elite to a partisan citizen directly – the platform uses algorithms that can alter the flow of information. Furthermore, as was made clear with many of Trump's tweets concerning the 2020 election, the platform can intervene directly by blocking or impeding the ability of an elite to tweet, labeling the elite's messages as misinformation, or making it difficult or impossible for those who view the elite's post to redistribute the message. While Twitter, and similar social media platforms, are private companies, like governments they can control the flow of information on their platforms.

Next, we draw upon the theory of censorship [18]. That theory argues that three mechanisms can be used to censor information online: fear, friction, and flooding. Censoring information through fear means using tools like financial sanctions or the threat of imprisonment to coerce citizens and elites to not disseminate information. Friction regards efforts to slow or make more difficult the dissemination of information. Flooding involves disseminating large quantities of competing information, which serves to make it more difficult and costly to find the information that the government aims to censor.

As [18] points out, introducing friction works in situations where "the cost added by censorship to the information is enough to offset the benefits of consuming or disseminating information" (p. 72). Recall that the RAS model notes that citizens use heuristics like partisanship to determine which elites they follow and whether they receive information from those elites. In situations where Twitter imposes no friction on Trump's tweets, the RAS model should apply: Republicans should be more likely to receive and accept Trump's tweets, most likely in the form of additional conversation about the topics of Trump's tweets online.

This theoretical foundation allows us to formulate the following two hypotheses:

- Hypothesis 1: Trump's tweets steer the direction of conversation, resulting in a higher volume of tweets concerning the topics that Trump discusses.
- Hypothesis 2: Actions taken by Twitter (restrictions, warnings) lessen the influence of Trump's tweets. Intervening on Trump's tweets will reduce the subsequent discussions, mitigating the effect of our first hypothesis. Consequently, these measures decrease the number of election fraud tweets by Republicans relative to unrestricted tweets.

In the next section we describe our data and methods, as well as how we test these hypotheses.

4 Data

We used three datasets in this study: Trump's tweets obtained from the Trump Twitter Archive[3] and Factba.se[4]; 2020 general election tweets that we collected using the Twitter API; and election misinformation tweets dataset *ElectionMisinfo2020* [12].[5]

Since Twitter suspended Trump's account, we could not directly obtain his tweets from the Twitter API. Therefore, we used the Trump Twitter Archive to obtain all tweets posted by Trump from September 1, 2020, to December 15, 2020. This dataset contains tweet ids, times, retweet counts, and texts. Additionally, we used Factba.se to identify the tweets that were labeled. Since Factba.se does not differentiate between restricted and warned tweets (documenting both types as "flagged" tweets), we marked "flagged" tweets with zero retweet counts as restricted tweets, and the remaining "flagged" tweets as warned tweets.

We collected the 2020 general election tweets dataset using the Twitter API from June 2020 to January 2021. We utilized the long-term Twitter monitor developed by [8] and keywords related to election fraud, remote voting, polling places, and other election topics. We used this dataset to study how Twitter's restrictions influenced the retweeting of Trump's tweets.

The election misinformation tweets dataset [12] is at the core of this study. It comprises tweets identified in 456 distinct misinformation stories from September through December 2020. For each tweet, the dataset displays the misinformation story it is part of, its identification number[6], the identification numbers of the tweets it retweeted/quoted/replied to, and its partisan lean (left, right, unknown). We used this dataset to construct time series of misinformation counts and study how Trump's tweets and Twitter's labeling impacted these time series.

We find 576 tweets of Trump directly appear in the dataset. Among the 576 tweets, there are 10 restricted tweets, 108 warned tweets, and 458 unrestricted/unwarned tweets. Fifty-nine tweets are directly labeled as misinformation and the summary of the 59 tweets can be found in Table 1.

5 Effects of Trump's Tweets

The first question we are interested in is the effects of Trump's tweets on the spread of misinformation. Our hypothesis is that Trump's tweets would directly lead to an increasing spread of the corresponding misinformation. To investigate this, we take each of Trump's tweets that appear in the ElectionMisinfo2020 dataset, plot the volumes of the corresponding misinformation story around the posting time of Trump, and look at the direct effect of Trump's posting on the

[3] https://www.thetrumparchive.com.

[4] https://factba.se/topic/flagged-tweets.

[5] The data and code used in this paper is available at https://github.com/jian-frank-cao/Disinformation-Intervention.

[6] The tweet ID can uniquely identify a message on Twitter, including tweet, reply, quote, and retweet.

Table 1. Summary of Trump's Tweets in the ElecMisinfo2020 Dataset

Story Number	Description	Count	Hard	Soft	Unrestricted	Retweet
Story 1	ballot harvesting: Ilhan Omar Project Veritas Video	3	0	0	3	1
Story 2	tech: dominion	34	0	23	11	12
Story 3	Late:Extended Ballots	1	0	0	1	0
Story 4	dead voters: general ticket	5	0	5	0	1
Story 5	Digital dumps: Michigan 128000 votes	2	2	0	0	0
Story 6	partisan vcr: Nevada whistleblower	1	0	1	0	1
Story 7	Physical Mail Mistakes: Deceased and Inactive CA	1	0	0	1	1
Story 8	Physical Mail Mistakes:MI Misprints for Troops	2	0	0	2	2
Story 9	poll watchers: Philly no entry list	1	0	1	0	0
Story 10	Physical Mail Fraud: Democratic TX Mayor	1	0	0	1	0
Story 11	Other: Stop The Steal Pushed	1	0	0	1	0
Story 12	Other: Candidate Fraud Biden Fraud Quote	2	0	0	2	0
Story 13	protests:stop the steal rallies	1	0	1	0	1
Story 14	Physical Mail Fraud: PA Misprinted Corrections	2	0	0	2	0
Story 15	Statistics: Math Video	1	0	0	1	0
Story 16	Physical Mail Mistakes:NYPost Ballot Typo	1	0	0	1	0

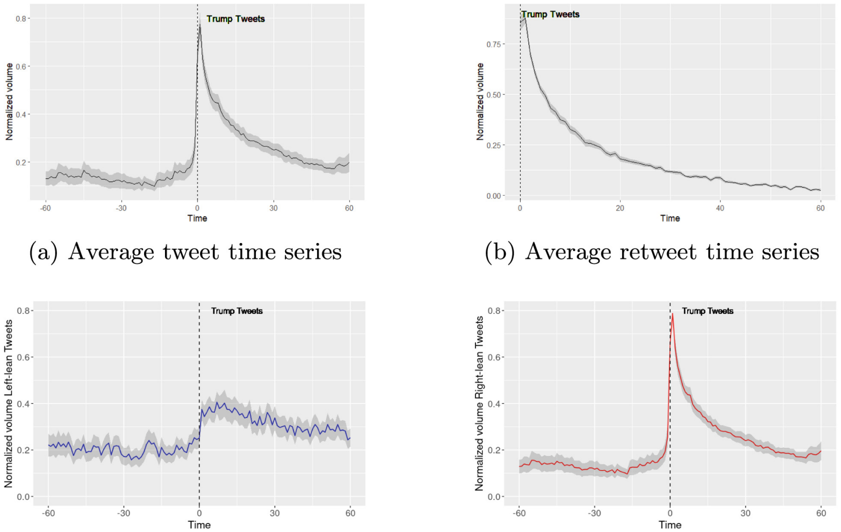
time series.[7] Note that we take all of Trump's tweets, regardless of Twitter's actions, which could bias the result downward. Therefore, the effect we discuss here might be a lower bound.

We find very similar patterns among almost all the 120-min windows around Trump's posting time: the tweet volume rises sharply, and then gradually decreases, eventually equilibrating at a stable volume that is higher than the level before the posting event. To estimate the average effect, we first normalize each 120-min window by applying the following transformation to each time window.

$$\hat{Y_i} = \frac{Y_i - min(Y)}{max(Y) - min(Y)}$$

Fig. 2a shows the average time series and the confidence interval among all the time windows. We also independently count retweets of Trump's tweets. The normalized average time series and the confidence interval among all windows are shown in Fig. 2b. Additional, we also plot the average normalized volume for left-lean and right-lean tweets separately in Fig. 2c and Fig. 2d. This shows that the overall effect on volume is much larger for right-leaning tweets, which is consistent with hypothesis 1.

[7] Out of all Trump's tweets, there are two that were posted close enough in time that their active periods overlap. In this specific instance, we study the combined effect of these tweets, using the timestamp of the first tweet as the reference point for our analysis.



(a) Average tweet time series

(b) Average retweet time series

(c) Average left-lean tweet time series

(d) Average right-lean tweet time series

Fig. 2. Average normalized time series among all 120-min time intervals affected by Trump's tweet

To quantify the effect of Trump's tweets and the heterogeneity across different topics, we perform a t-test for each story on the average column in three time periods. Period 1 spans 30 minutes before Trump's tweet ($T = -30$ to $T = -1$). Period 2 spans 30 minutes after Trump's tweet ($T = 0$ to $T = 29$). Lastly, Period 3 is 30 minutes to 60 minutes after Trump's tweet ($T = 30$ to $T = 59$).

We compare the data from Periods 1 and 2 to see the immediate effect of the tweet, and from Periods 1 and 3 for the longer-term impact. The volume per minute comparisons before and after each tweet, along with the t-test results, are displayed in Fig. 3. The graph indicates that Trump's tweet has a heterogeneous effect across different topics. We can observe that for most of the topics, there is an increase in volume either immediately or after 30 minutes. The volume does not immediately increase for some topics with Twitter's intervention like "Dead voters", "Nevada Whistleblower", and "Poll Watcher", which provides evidence in support of hypothesis 2.

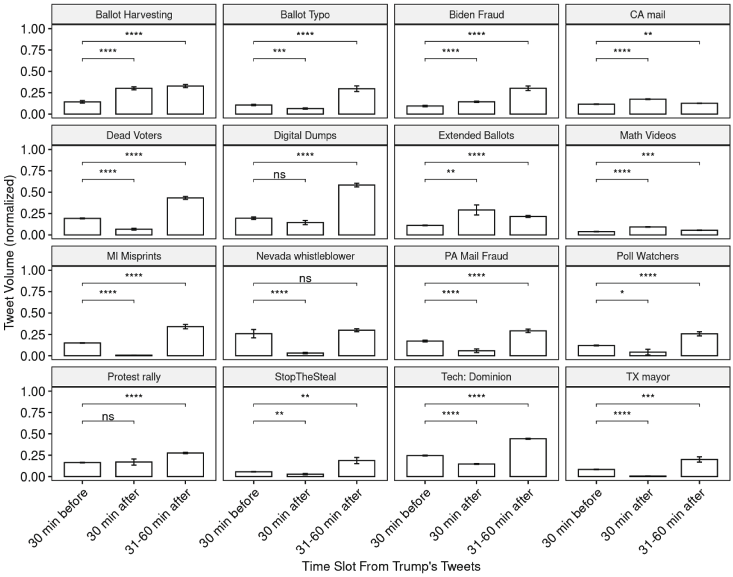


Fig. 3. Effect of Trump's tweets By Topic and Time. Note: ns: $p > 0.05$; *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$; ****: $p \leq 0.0001$

6 Effects of Twitter's Actions

To study the effects of labeling, we first estimate the time Twitter applied the label. We then derive time series of misinformation related to Trump's tweets. We compare the time series of messages where a label was applied against counterfactuals derived from messages that were not subject to any restriction. Using this data, we estimate the effects of labeling.

To study the effects of labeling, it is necessary to know when the labeling became effective, i.e., the treatment time. However, Twitter has not disclosed the exact timing of the labels, only stating that they applied labels between 5 to 30 minutes after Trump posted the tweets.[8] Fortunately, our 2020 general election data contains real-time retweets of 7 out of 26 of Trump's restricted tweets. Each retweet contains a retweet status object that points to Trump's original tweet and shows its latest retweet count by the time the retweet was collected by our Twitter monitor. The time series of cumulative retweets are shown in Fig. 4. We can see that the time series stopped around 20 to 240 minutes after Trump tweeted because Twitter restricted users' ability to retweet, and no more new retweets were collected. The stopping points (red) are our estimates of labeling time. Notice that labeling took around 1.5 to 4 h in September, while it only took around 30 minutes in November. Twitter expedited its labeling, likely because election misinformation was spreading fast and the potential damage to society was great. Since we cannot directly estimate the labeling time of warned tweets

[8] https://www.youtube.com/watch?v=ONYuLP7sHFQ&t=4701s.

as retweeting was not restricted, we assume it is similar to that of the restricted tweets.

Next, we derive the time series of misinformation tweets related to Trump's tweets. For each of Trump's tweets included in the misinformation data set, we find that tweet's corresponding story.[9] We then compute the number of tweets posted per minute, across the entire misinformation dataset, from the associated misinformation story. We focus on the time series from T to $T + 120$, where T is the timestamp of Trump's tweet. Thus, each of Trump's tweets produces a misinformation time series. When the Trump tweet that produces this time series is "labeled" by Twitter, we refer to this as a "labeled misinformation time series", and if the Trump tweet is "unlabeled", an "unlabeled misinformation time series."[10].

We use synthetic control to construct counterfactuals of the labeled time series. For each labeled misinformation time series, if there are more than five unlabeled misinformation time series in the same story, we use them to estimate the synthetic control. Otherwise, we disregard the stories and use all unlabeled time series. Based on the estimates of labeling time in Fig. 4, assuming most labeling was imposed after $T + 20$, we estimate the synthetic control using the T to $T + 19$ sub-series to ensure that it closely resembles the labeled time series in the first 20 minutes.

We show synthetic controls for all of Trump's restricted tweets in Fig. 5a and 15 out of 201 warned tweets in Fig. 5b. The synthetic control, i.e., the estimated tweets if there was no restriction, is shown in red, and the observed tweets are shown in blue. The area between the red and blue curves are the estimated effects of labeling. If the red curve is above the blue curve, then the effect of labeling is negative, which means labeling reduces the spreading of misinformation. For example, those where there is solid evidence from the synthetic control methodology that Twitter's content moderation reduced misinformation are Trump's that Twitter restricted on Nov 04 15:37:40, Nov 04 21:56:11, and Nov 05 16:22:46. Additionally the synthetic control methodology indicates that Twitter's content moderation reduced misinformation in the instances where they placed warnings on Trump's tweets about the election on Nov 04 21:56:10, Nov 09 00:23:26, and Nov 12 15:16:02. On the other hand, if the blue curve is above the red curve, this indicates positive treatment effects, in which Twitter's labeling stimulates more discussion about misinformation. For example, the synthetic control method indicates positive treatment effects when Twitter restricted Trump's tweets on

[9] In some cases, the Trump tweet is not directly in the misinformation dataset, but we do find the tweet's associated retweets, quotes, and replies. In all cases where we find more than ten examples of retweets, quotes, or replies with a story in the misinformation dataset, we define the Trump tweet's misinformation story as the most common across this set of retweets, quotes, and replies. If we find fewer than ten examples, we drop this Trump tweet from our analysis.

[10] If any Trump tweet did not lead to significant corresponding misinformation time series from T to $T + 120$, i.e. less than 100 tweets per minute on average, we dropped it from our analysis.

Nov 05 15:09:19 and when Twitter placed warning labels on his election-related tweets on Nov 19 17:34:26, Nov 30 00:34:38, and Dec 14 14:38:38.

Overall, the synthetic control results shown for the thirty examples in Fig. 5a and Fig. 5b provide a nuanced perspective on Twitter's attempts in 2020 to slow or stop the spread of election misinformation by restricting or placing warning labels on Trump's tweets. Among the restricted tweets (Fig. 5a) we see relatively clear evidence in 8 of the 15 instances for restriction reducing the subsequent spread of misinformation. Similarly, among the Trump tweets where warning labels were used, 6 of the 15 examples show that the subsequent spread of misinformation was slowed.

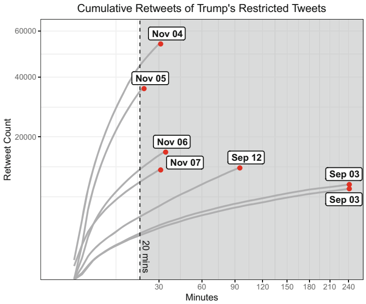


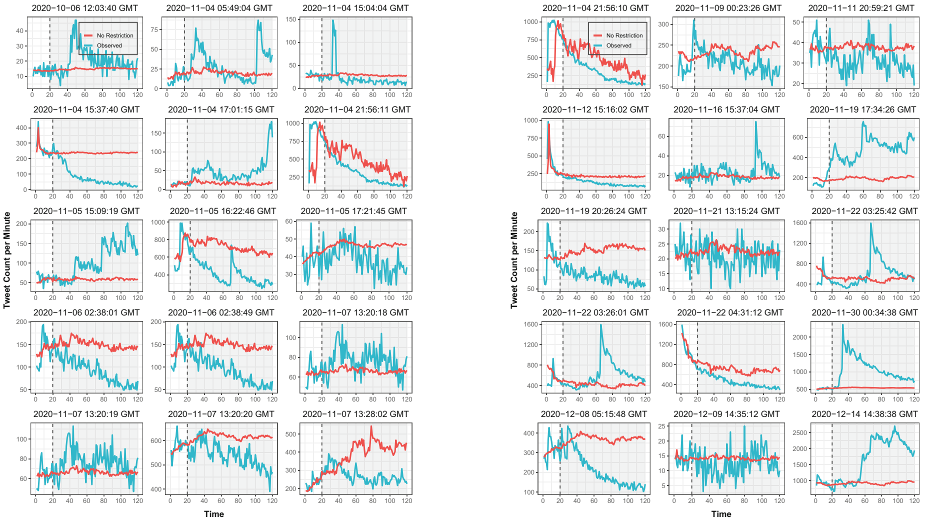
Fig. 4. How Fast Did Twitter Apply the Restriction?

With these synthetic controls, we quantify the labeling effect using ratios of average tweets in the second hour:

$$\phi_i = \frac{\frac{\sum_{t=61}^{120} (Observed)_{i,t}}{60}}{\frac{\sum_{t=61}^{120} (Estimated\ No\ Restriction)_{i,t}}{60}} \tag{1}$$

The ratio is less than one if the average observed tweets from $T + 61$ to $T + 120$ is smaller than the average estimated tweets if there was no restriction, i.e., the blue curve is above the red curve, and it is larger than one otherwise. Since this study is interested in analyzing how Twitter's labeling reduces the spreading of misinformation, we focus on Trump's tweets that are associated with a large number of misinformation tweets and exclude time series that have on average fewer than 100 tweets per minute.

The ratios are shown in Fig. 6 and Table 2. We use shapes to distinguish Trump's restricted and warned tweets and use colors to show stories. We see that the majority of the ratios are less than one, i.e., in the green area, as 55 of the tweets in this analysis are in the green area. Importantly we note that of the six tweets in this sample that were restricted, five of the restricted instances were ones where the subsequent flow of misinformation were reduced, and in only one of those instances was the subsequent flow of misinformation



(a) Estimated No Restriction Tweets vs. Observed Tweets (Restricted)

(b) Estimated No Restriction Tweets vs. Observed Tweets (Warned) (Sample)

Fig. 5. Effects of Restriction

not reduced by the restriction of Trump's tweets. It is also important to note that these six restricted tweets were in the immediate aftermath of the 2020 presidential election, at a time when mitigating the spread of misinformation might have been most influential. We also must note, however, that 30 of the tweets in this analysis (the vast majority of which were those with warning labels) show positive treatment effects, meaning that misinformation increased after the warning labels were used. Some of the tweets with warning labels have sizable increases in post-moderation spread, one of the tweets with a warning label in early December 2020 saw a 1600% increase in post-treatment misinformation spread.

Finally, for interested readers, we also show the distribution of ϕ_i in Table 2. A t-test of $log(\phi_i)$ yields a t-value of -2.4737 and a p-value of 0.0149, which means labeling significantly ($P < 0.05$) reduces the volume of misinformation tweets in the testing period $[T+61, T+120]$. The mean of the log effect $\overline{log(\phi_i)} = -0.1673$ indicates that, on average, Twitter's labeling reduces $1 - e^{-0.1673} = 15.41\%$ of misinformation tweets.

7 Discussion

Existing literature presents conflicting findings on the ability of social media platforms to mitigate the spread of misinformation effectively. Our study, however,

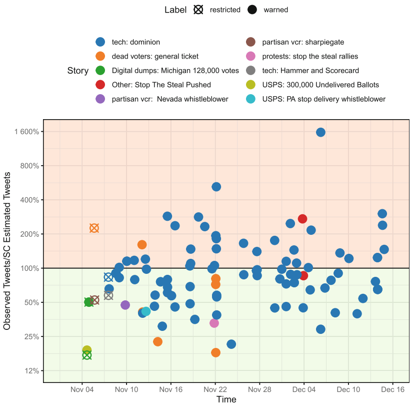


Fig. 6. Quantification of the Estimated Treatment Effect

Table 2. Tweets in Estimated Ratio Ranges

	Restricted	**Warned**
(12.5%, 25%]	1(16.67%)	4(3.92%)
(25%, 50%]	0(0%)	17(16.67%)
(50%, 100%]	4(66.67%)	45(44.12%)
(100%, 200%]	0(0%)	25(24.51%)
(200%, 400%]	1(16.67%)	9(8.82%)
(400%, 800%]	0(0%)	1(0.98%)
(800%, 1600%]	0(0%)	1(0.98%)
Total	6	102

takes a more targeted approach, examining a particular facet of platform moderation. We utilize a unique dataset and adopt a sophisticated causal inference methodology to increase the validity of our conclusions. Our findings suggest that actions taken by social media platforms can mitigate the subsequent spread of misinformation. We call for further research to better understand the conditions under which moderation is possible and which interventions are the most effective.

In particular, the next stage of research needs to tackle the conditions when content moderation has the desired treatment effect. Is restriction more effective than labeling (we see intriguing evidence that the answer may be yes in Fig. 6)? Does it matter when a platform applies restrictions or labels? Does the speed at which moderation is carried out affect its effectiveness? Is the wording of the warning label important for restricting subsequent spread? There are many additional questions that researchers and social media companies should tackle.

It is important to view our conclusions in through lens of the current moment, wherein some social media channels opt for less moderation, ostensibly to champion free speech. Discussions surrounding the policies being implemented by states such as Florida and Texas, in conjunction with legal debates about the moderation of certain social media dialogues, highlight potential restrictions on content moderation. While Constitutionally-protected political speech might be an area where content moderation is problematic, that should not imply that social media platforms should stop efforts to prevent the spread of child pornography, voting disenfranchisement, sexual and racial harassment, or the use of their platforms by terrorists organizations. The research community needs to step up our involvement in these debates, and provide research that can help social media platforms develop appropriate content moderation policies that protect rights while preventing illegal behavior and social harm.

References

1. Abadie, A., Diamond, A., Hainmueller, J.: Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. J. Am. Stat. Assoc. **105**(490), 493–505 (2010)
2. Abadie, A.: Using synthetic controls: feasibility, data requirements, and methodological aspects. J. Econ. Lit. **59**(2), 391–425 (2021)
3. Abadie, A., Gardeazabal, J.: The economic costs of conflict: a case study of the Basque country. Am. Econ. Rev. **93**, 113–32 (2003)
4. Adams-Cohen, N.: New Perspectives in Political Communication. Ph.D. thesis (2019), California Institute of Technology. https://doi.org/10.7907/7TDG-4R42
5. Adams-Cohen, N.: Policy change and public opinion: measuring shifting political sentiment with social media data. Am. Politics Res. **48**(5), 612–621 (2020)
6. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J. Econ. Perspect. **31**, 211–236 (2017). https://doi.org/10.1257/JEP.31.2.211
7. Allington, D., Duffy, B., Wessely, S., Dhavan, N., Rubin, J.: Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. Psychol. Med. **51**(10), 1763–1769 (2021). https://doi.org/10.1017/S003329172000224X
8. Cao, J., Adams-Cohen, N., Alvarez, R.M.: Reliable and efficient long-term social media monitoring. J. Comput. Commun. **09**(10), 97–109 (2021)
9. Carey, J.M., et al.: The ephemeral effects of fact-checks on COVID-19 misperceptions in the united states, great Britain and Canada. Nat. Human Behav. **6**, 236–243 (2022). https://doi.org/10.1038/s41562-021-01278-3
10. Clayton, K., et al.: Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Polit. Behav. **42**, 1073–1095 (2020). https://doi.org/10.1007/S11109-019-09533-0
11. Downs, A.: An economic theory of democracy (1957)
12. Kennedy, I., et al.: Repeat spreaders and election delegitimization: a comprehensive dataset of misinformation tweets from the 2020 US election. J. Quant. Descrip. Digital Med. **2**, 1–49 (2022)
13. Lupia, A.: Busy voters, agenda control, and the power of information. Am. Polit. Sci. Rev. **86**(2), 390–403 (1992)
14. Lupia, A.: Shortcuts versus encyclopedias: information and voting behavior in California insurance reform elections. Am. Polit. Sci. Rev. **88**(1), 63–76 (1994)

15. Pennycook, G., Rand, D.G.: Fighting misinformation on social media using crowdsourced judgments of news source quality. Proc. Natl. Acad. Sci. U.S.A. **116**, 2521–2526 (2019). https://doi.org/10.1073/PNAS.1806781116
16. Popkin, S.L.: The Reasoning Voter: Communication and Persuasion in Presidential Campaigns. University of Chicago Press (1991)
17. Porter, E., Wood, T.J.: The global effectiveness of fact-checking: evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. Proc. Natl. Acad. Sci. U.S.A. **118**, e2104235118 (2021). https://doi.org/10.1073/PNAS.2104235118/SUPPL_FILE/PNAS.2104235118.SAPP.PDF, https://www.pnas.org/doi/abs/10.1073/pnas.2104235118
18. Roberts, M.E.: Censored. Princeton University Press, Princeton (2018)
19. Sanderson, Z., Brown, M.A., Bonneau, R., Nagler, J., Tucker, J.A.: Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. Harvard Kennedy School Misinformation Review 2 (2021)
20. Srikanth, M., Liu, A., Adams-Cohen, N., Cao, J., Alvarez, R.M., Anandkumar, A.: Dynamic social media monitoring for fast-evolving online discussions. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 3576–3584. KDD 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3447548.3467171, https://doi.org/10.1145/3447548.3467171
21. Thèro, H., Vincent, E.M.: Investigating Facebook's interventions against accounts that repeatedly share misinformation. Inf. Process. Manage. **59**, 102804 (2022). https://doi.org/10.1016/j.ipm.2021.102804
22. U.S. Senate Select Committee on Intelligence: Russian active measures campaigns and interference in the the 2016 U.S. election, volume 2: Russia's use of social media. U.S. Senate Select Committee on Intelligence Report 116-XX (2020). https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf
23. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**, 1146–1151 (2018). https://doi.org/10.1126/SCIENCE.AAP9559
24. Zaller, J.R.: The Nature and Origins of Mass Opinion. Cambridge University Press, Cambridge (1992)