When Partisans Fly: Twitter Community Notes and the Political Economy of Social Media Disinformation

John Kim

April 13, 2023

Contents

List of Figures

List of Tables

Abstract

As concerns over political disinformation grow increasingly prevalent on social media platforms, crowdsourced content moderation schemes have emerged as a possible method of addressing these concerns. Twitter's Community Notes is the first social media platform to implement such a scheme, and in this study, I examine the effects of Community Notes on engagement with political Tweets. Using a novel dataset collected via public data releases and the Twitter API as well as a model of Bayesian media consumption, I find evidence that Community Notes decreases engagement at a significant level at-large ($p < 0.01$), but conditioning on Tweets by politicians, the effect disappears ($p > 0.1$). In addition, consistent with strategic models of media consumption, I find that Democratic agents tend to be sensitive to the partisan bias of the source cited in a Community Note ($p < 0.05$), while Republican agents tend to reject Community Notes overall ($p < 0.01$). This suggests that from a social planning perspective, Community Notes and similar crowdsourced content moderation interventions may be successful in moderating speech at low cost and with less controversy than other stronger forms of content moderation.

Introduction

The Internet has long been regarded as a boon to free speech and political discourse in 21st century American politics. Yet, over time, the Internet's potential as a democratic forum has been undermined by its potential to spread hate speech and political disinformation, and such content runs amok on well-known social media platforms like Facebook and Twitter. In turn, social media companies and policymakers alike seek a careful balancing act in the tradeoffs between free speech rights, algorithmic transparency, and moderation of political disinformation.

Indeed, these concerns are not unfounded. With increasing evidence of the influential role of social media in US politics, social media effects have entered the forefront of political

economy literature: studies range from the effects of social media on the 2016 and 2020 presidential elections (Allcott and Gentzkow 2017; Fujiwara et al. 2021) to the integration of social media in the existing literature on media effects, echo chambers, and political dysfunction (Haidt and Bail Ongoing). In turn, strategic models of media consumption have grown in popularity, alongside experimental and computational methods of analyzing social media, politics, and communications.

Most relevant to the social planner's problem, then, are advances in political economy of social media that assist in evaluating the aforementioned tradeoffs. In turn, political scientists and economists alike have dedicated their efforts to the research of content moderation. The study of content moderation methods is extraordinarily nuanced: what methods are effective in preventing the spread of disinformation and hate speech? How do they reach their goals without interfering with free speech? How do agents strategically respond to these methods? And do they have real-world political consequences, such as in polarization or electoral competition?

In evaluating such content moderation methods, one innovation, crowdsourced content moderation, has emerged as a promising alternative. Rather than leaving content moderation as a unilateral action by social media platforms, crowdsourced content moderation schemes instead allow users to submit fact checks or other "notes" onto social media communications in response to misleading content. This presents a clear agency problem: to what degree do we expect agents to act in good faith to produce quality content moderation interventions, and in response, to what extent can we expect agents to positively respond to such interventions?

Only one source of data has emerged to evaluate such a program: in October 2022, Twitter became the first and only major social media platform to deploy a crowdsourced content moderation scheme, called Community Notes. In Community Notes, Twitter invites select users to contribute notes on content the user deems to be misleading or potentially harmful,

often combined with a citation of some external source. Finally, the note is voted on by users; when notes receive a sufficient number of supposedly bipartisan votes, the note is displayed.

Several relevant issues to the political economy of social media emerge from Community Notes. First, from a social planner's perspective, is such an intervention actually effective in preventing engagement with disinformation with social media, and with what tradeoffs? Second, once consumers of information are added as strategic agents in content moderation, how do different agents respond to such a program? Finally, how does such a program translate to real-world political outcomes?

Indeed, Twitter has committed to data transparency in the Community Notes program in the form of data releases, but with such novel and complex data, definitive answers to the above questions remain far-off. Nevertheless, in one of the first evaluations of such a program, I present a strategic model of Bayesian media consumption with an uncertain external signal (via informational cascades), and use this model to guide a direct analysis of the effects of Community Notes on social media engagement. I also perform a secondary analysis on strategic media consumption, in the form of selective engagement with Community Notes based on partisan priors and cited partisan sources.

Through a difference-in-difference design on data retrieved from both Community Notes data releases and the Twitter API, I find that though Community Notes causes a significant decrease in Twitter engagement on Tweets at-large ($p < 0.01$), the effect disappears conditional on Tweets by politicians ($p > 0.1$). This is also mediated by evidence that Democrats tend to have longer-lasting notes on their Tweets relative to Republicans ($p < 0.01$) and that Democrats in particular tend to reject notes that cite the partisan source Fox News ($p < 0.05$).

This leads to several conclusions. In particular, since agents who follow politicians are most likely to be agents with the strongest partisan priors, the Community Notes program

effectively has a significant causal effect on the agents most susceptible to misinformation: those who have weak partisan priors. Furthermore, the evidence that Community Notes has significant effects on political engagement for individuals with weak partisan priors implies that Community Notes may indeed enact real-world political effects, given that these same individuals are more likely to change their beliefs and electoral choices (relative to agents with strong priors).

In turn, the results here have many implications, not limited solely to the evaluation of Community Notes. For example, I provide real-world evidence supporting many contentious theories in the existing literature on social media and politics, such as selective exposure based on partisan priors, consumer rejection of soft content moderation schemes, and strategic methods of disseminating disinformation. Likewise, I conjecture on the real-world political effects of social media disinformation and outline further avenues of research for this developing program. Finally, I provide a perspective on Community Notes relevant to policy and social planners, given increased policymaker attention to the issue of content moderation as well as Community Notes's own open-source availability.

The paper proceeds as follows. First, I introduce the literature on the political economy of social media consumption and content moderation, followed by a detailed description of the the Twitter Community Notes program. Second, I introduce a formal model for the study of Community Notes, using a simple model applying the informational cascades model presented by Bikhchandani et al. (1992); the formal model in turn guides my hypotheses. Third, I introduce the data, including method of data collection, matching of units, and analysis, and I present the results in the section following. Fourth, I discuss my results in-depth, both in the context of Community Notes and within the context of existing theories of the political economy of social media consumption. Fifth, I discuss limitations in my study of Community Notes, as well as possible extensions of my study and the study of crowdsourced content moderation as a whole. I conclude with a discussion of Community Notes relevant to policymaking and social planning.

Background

Media & Politics

Though the study of politics is inextricably linked with media, the exact nature of media effects on politics remains highly contentious, much less social media. Notwithstanding the diverse channels and methods through which media attempts to influence agents, the study of media is extraordinarily difficult: randomized controlled trials (RCTs) that hold ecological validity in the actual media consumption environment are few and far between, and the presence of many confounders hinders the ability to make causal claims in observational studies. This has produced a rich yet uncertain literature on the effects of media on politics, and in particular, scholars and policymakers alike still debate the effects of social media on politics and political dysfunction (Haidt and Bail Ongoing).

Three theories of media effects dominate the literature: the hypodermic needle model, the minimal effects model, and contingent effects model (Arceneaux and Johnson 2013). The hypodermic needle model posits that media has a substantive and direct effect on political beliefs and opinions (Arceneaux and Johnson 2013). In contrast, the minimal effects model hypothesizes that media has very little direct effect on politics and policy, instead acting as a mediating or reinforcing signal (Arceneaux and Johnson 2013). Similarly, the contingent effects model theorizes that media effects are based on ongoing political factors such as elite actors, with component theories like agenda setting, framing, and priming (Arceneaux and Johnson 2013).

These theories reflect the scholarly debate over not only *whether* media affects political discourse, but to its *extent* and *nature*. For example, the hypodermic needle model was widely accepted throughout the early 1900s with the impetus of World War I propaganda and the infamous War of the Worlds broadcast (Lasswell 1971). The model was in turn "disproven" by the landmark Columbia Studies, which instead proposed the dominant role

of opinion leaders, as in the minimal effects model (Lazarsfeld, Berelson, and Gaudet 1988).

Yet, even with growing adoption of the minimal effects and contingent effects models, both the direct effects and real-world impacts of media, particularly social media, are undeniable. Nearly every Congressperson and government body in the US has adopted social media[1] as a method of reaching out to constituents and citizens, and recommendation systems on social media platforms themselves ensure certain messages can spread far and wide among users. In turn, the real-world effects of such systems are increasingly studied by researchers; for example, Fujiwara et al. (2022) find that Twitter usage may have directly affected electoral outcomes in the 2016 and 2020 US Presidential Elections in favor of Democratic candidates.

Haidt and Bail (Ongoing) attempt to summarize the vast, complicated literature on social media and politics in their collaborative literature review, investigating both the mediums through which social media might operate (e.g. echo chambers, amplification) and their real-world effects (e.g. promoting violence, strengthening populist movements). Relevant to Community Notes is the question of the spread of disinformation, and so I examine more closely one of their component research questions: "Does social media amplify posts that are more emotional, inflammatory, or false?"

Spread of Disinformation

Though the literature on disinformation on social media is by no means conclusive, Allcott and Gentzkow (2017) provide a good starting point. In their investigation of social media and disinformation in the 2016 US presidential election, the authors first find that though political news consumption is significant on social media, social media is by no means the largest source of political news. They then note that while only a slight majority of Americans believe fake news when it is presented to them, they are more likely to believe it when it agrees with their priors. So, Allcott and Gentzkow (2017) reveal the complicated nature of

[1] see "Politicians tracked by Politwoops," ProPublica Data Store (2019), https://www.propublica.org/datastore/dataset/politicians-tracked-by-politwoops.

social media and disinformation: it is ever-present, limited but significant, and ever nuanced as to its effects. Most importantly, Allcott and Gentzkow (2017) describe social media disinformation as an *agency problem*; that is, a self-interested principal might strategically spread disinformation according to their biases and reputation, and an agent believes such information according to their priors and personal costs in turn.

Understanding individual behaviors, then, is crucial to understanding the spread of political disinformation on social media. For example, by nature of its production, fake news is naturally more emotionally charged than other posts, and Twitter users respond in turn. An observational study by Vosoughi et al. (2018) reveals that falsehoods were disseminated far more widely than true news, hinting that the emotional calls to action sparked by fake news lead to further dissemination.

This is notwithstanding further strategic elements and concerns regarding the spread of disinformation. For example, Freelon et al.'s (2020) study of disinformation on digital media notes the difference in the dissemination of information (not necessarily misleading) between left-wing and right-wing actors. In particular, information (true or false) disseminates quickly and among non-elites through left-wing actors through the use of hashtags and trends (what they call *hashtag activism*), while right-wing actors tend to appear more strategic in disseminating misinformation and in concert with offline media outlets or actors.

In turn, my study, and similar research on crowdsourced content moderation, may point to further understanding of how disinformation spreads on platforms like Twitter. In particular, assuming Freelon et al.'s (2020) hypothesis to be true, I examine the different attitudes and strategies adopted between left-wing vs. right wing actors. Likewise, as in Allcott and Gentzkow (2017), I examine the role of priors and biases to determine whether certain citations in news sources and other measures of partisan bias results in further dissemination of or engagement with disinformation.

Content Moderation

Content moderation is a hotly contested topic in both academic and policymaking circles. In particular, debates on content moderation revolve around the effectiveness of different methods of countering hate speech or misinformation, the free speech implications of such methods, and whether such methods have real-world political effects. Given the expansive and contentious nature of this field, I concern myself primarily with the first question on the effectiveness of content moderation.

In accordance with its complicated nature, methods of content moderation adopted by online platforms have become increasingly diverse. Traditional methods such as deleting certain content or blocking certain users have remained popular, though the use of such methods is extremely controversial[2]. Meanwhile, platforms have turned to "softer" methods that do not outright restrict speech, including labeling of content/users and including links to external sources (see Figure 1).

Figure 1a: Labeled Content on Tweet

[2]see "Permanent suspension of @realDonaldTrump," Twitter (2021).

Figure 1b: Labeled User on YouTube Content

Undoubtedly, programs like Community Notes were born out of the growing popularity of such content moderation methods, and the effects of different content moderation methods have been scrutinized in the literature. For example, Aslett et al. (2022) conduct a study on the effects of source credibility labels on media consumption. The authors find that by and large, consumption of low-quality media sources did not change, but among the largest consumers of misinformation, the source labels caused a substantive increase in quality of media consumed. Barrera et al. (2019) instead find that consistent with models of Bayesian consumers, exposure to fact-checking sources do indeed change beliefs/priors, subject to quality of the source (official sources vs. politicians). However, this is also accompanied by a lack of a real-world effect: Barrera et al. (2019) observe that despite their updated beliefs, voters were unlikely to change their electoral choice.

Specific to Community Notes, Allen et al. (2021) describe the potential impacts of "soft" content moderation via crowdsourced content using an experimental sample. Their trial experiment showed that if politically unaligned individuals contributed to a fact-check to-

gether, the accuracy of the crowdsourced fact-check was in-line with the work of professional fact-checkers. Still, the authors note these effects are conditional: in particular, alignment with the Democratic Party and greater political knowledge are substantially correlated with greater trust in the fact-checks overall.

Thus, the research I present here contributes several key insights to the literature on content moderation. First, I provide a direct evaluation of how users interact with automated and crowdsourced content moderation systems: as online platforms face increasing difficulty in moderating vast social networks, such methods of content moderation have faced increased interest and scrutiny by technologists and policymakers online. Second, I help describe the exact nature of content consumption and trust in media: since Community Notes fact-checks are provided by users, and by extension, the media sources they cite, I can test the views and impacts of various sources and how they are presented. Finally, as data on content moderation becomes more widely available, as with the Twitter API and Community Notes data releases, I outline and expand on new approaches integrating computational social science and econometrics to suggest further development in this rich literature.

Overview of Twitter Community Notes

Twitter's Community Notes program (originally called Birdwatch) was announced in January 2021 as part of Twitter's efforts to counter the spread of misleading information on Twitter (Coleman 2021). The program allows Twitter contributors to write notes to provide additional context or information for Tweets they believe contain misleading information (Coleman 2021). Notably, to sign up for the program, Community Note authors must fulfill and maintain certain requirements beyond those required to join Twitter as a general user, including an account age restriction, continued submission of generally helpful notes, and contribution to ratings of other notes. The notes themselves are intended to identify missing context behind the Tweet or misinformation, often followed by a reference to an external article or news source (for example, cnn.com or foxnews.com) (see Figure 2[3]).

[3]Twitter post by @WhiteHouse, May 12, 2022, 5:45PM, https://twitter.com/WhiteHouse/status/1524868269148192779.

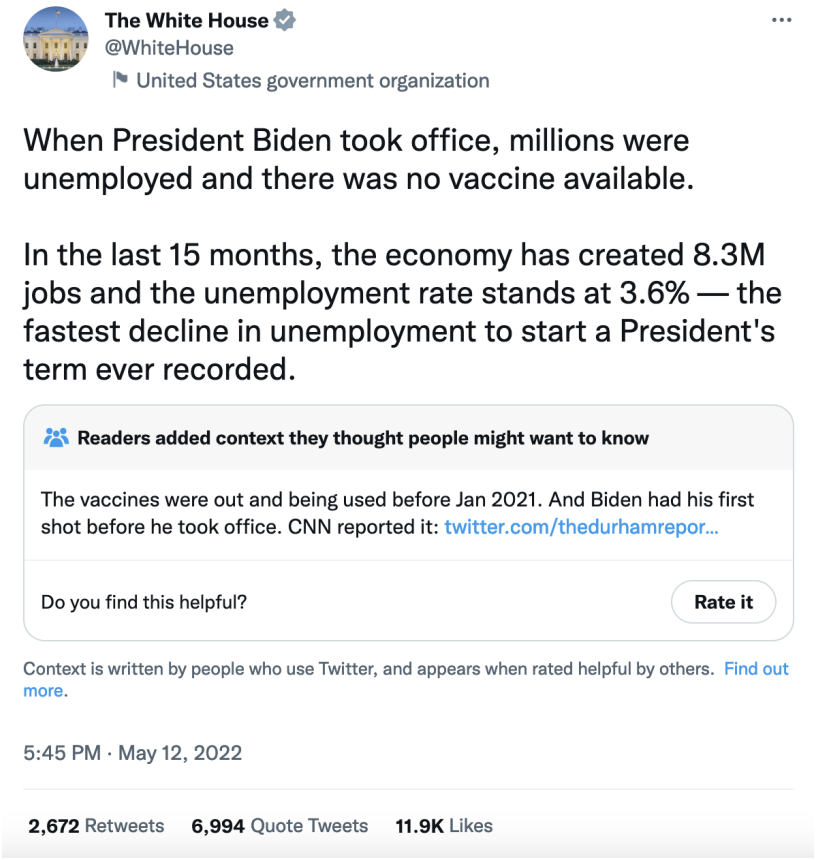

Figure 2: Example Community Note Added to White House Tweet

Then, notes are voted on by Twitter users and Birdwatch contributors as a whole; if the note is rated as helpful by users of both supporting and opposing views, the note will be displayed on the Tweet ("Community Notes Guide"). Votes are taken on an ongoing basis: once a note is displayed on a Tweet, if additional users deem the note unhelpful, the note may subsequently be removed. Likewise, if multiple notes are submitted and rated helpful on the same Tweet, Twitter will automatically cycle through displaying each of the notes at random ("Community Notes Guide").

Community Notes was made public in the US in October 2022[4]. With its release, Twitter revealed the results of its internal tests of the program through 2021 and 2022. This research, as outlined in Wojcik et al. (2022), provide an optimistic view of the program: surveyed

[4]see "Helpful Birdwatch notes are now visible to everyone on Twitter in the US," Twitter Blog (2022).

populations who saw an attached Birdwatch note were upwards of 26% less likely to agree with a Tweet containing misleading information and up to 34% less likely to retweet a Tweet containing misleading information. These results paint a rosy picture for the potential of Community Notes; nevertheless, with an extremely limited sample size relative to the rest of Twitter users and a much smaller experimental setting, the results found by Wojcik et al. (2022) have limited external validity and may fail to replicate in real-world conditions.

Though Community Notes is relatively new, however, research on the program is burgeoning. In one of the few external sources of research on the program, Allen et al. (2022) investigate the role of partisanship in the effectiveness of the Community Notes program. The authors find that shared partisanship between a note author and a Tweet results in a higher likelihood of the note being kept, indicating that Community Notes may have trouble challenging viewers' partisan priors. Nevertheless, research on the success of Community Notes remain largely inconclusive, yet such studies, including my own, are essential to understanding the potential for automated content moderation systems in facilitating political discourse.

Model & Hypotheses

To evaluate the effects of Community Notes, I consider a model of informational cascades. A formal model is useful here for two reasons: first, it guides my hypotheses to examine both why a certain result may occur and in what circumstances we might expect a counterfactual result. Second, given the wealth of data offered by the Community Notes public release, it helps narrow down our variables of interest to those that may be relevant.

In turn, informational cascades are a useful model to consider here, since they model sequential decision-making under uncertainty with simplicity. As prior research has shown, informational cascades have proven very effective in modeling network effects and social interactions (Easley and Kleinberg 2010), and network cascade models have been used in similar research regarding the dissemination of information on social media (Friggeri et al. 2014). Likewise, the model incorporates Bayesian consumers of media, allowing agents to have individual priors and update their beliefs in response to the Community Note signal to form a posterior.

The Model

The model outlined here is a simple model of informational cascades, as presented by Bikhchandani et al. (1992). Consider a sequence of n individuals, making a sequential decision to believe or reject some information. Believing the information has some cost C, and the payoff from believing the information is $V \in \{0, 1\}$ (assume for now $C = \frac{1}{2}$). Each individual $i \in \{1, 2, 3, ..., n\}$ observes a private signal $x_i \in \{H, L\}$, such that each individual has prior $P(x_i = H \mid V = 1) = p$ (for simplicity, assume p is identical across all individuals; also assume for now $p > \frac{1}{2}$). As a tiebreaking convention, if any agent i is indifferent between believing and rejecting the information, they will pick either with equal probability.

The well-known result of this informational cascade model is that agents will base their de-

cisions off of relatively little information and ignore their private information. In particular, if any two individuals consecutively pick the same belief, following agents will ignore their private signals and follow the prior decision to believe or reject some information (see Appendix 1). Under the conditions above, Bikhchandani et al. (1992) solve that the probability of an H cascade after two individuals given $V = 1$ is:

$$\mu \equiv P(\text{H cascade} \,|\, \text{V=1}) = p^2 + \frac{1}{2}p(1-p)$$

which is increasing in p (see Appendix 1).

In their original paper, Bikhchandani et al. (1992) argue that informational cascades in this context are extremely fragile. They note that agents are aware that in an informational cascade, they ignore their private signals based on the actions of only two individuals, and so revealing any conflicting public information (i.e., any signal H or L that opposes the cascade) will be sufficient to reconsider their private signal. Thus, even if a cascade appears to have the strength of numerous agents, it may break quite easily, not unlike a turtle egg.

In the context of Twitter Community Notes, this model has several desirable traits. First, it imitates the sequential decision-making of Twitter users well: once a user likes or retweets a Tweet, the Tweet is (either algorithmically or by design) pushed to other users, who then view the Tweet and indicate their own belief (like/retweet, an action with cost C) or rejection (inaction, with no cost). Likewise, it leaves open the possibility of incorporating an external signal, as with a Community Note.

Consider, then, the intent behind a Community Note. Incorrect information is disseminated to individuals, such that $V = 0$, but an H cascade nonetheless begins. An agent j realizes that their decision to follow an H cascade is based off of the signal of only the first two agents who chose *believe*, regardless of the current length of the cascade. Thus, if agent j receives signal $x_j = L$ and is also provided an exogenous signal (a Community Note) $y = L$,

given probability $P(y = H \mid V = 1) = q$, agent j instead chooses *reject* if:

$$P(V = 1 \mid x_j = L, y = L, H \text{ cascade}) < C$$

So, we have (see Appendix 2):

$$\frac{(1-p)(1-q)\mu}{(1-p)(1-q)\mu + pq(1-\mu)} < C \ \forall p > C \tag{1}$$

Even if agent j is provided the same signal as the cascade $x_j = H$, we have:

$$\frac{p(1-q)\mu}{p(1-q)\mu + q(1-p)(1-\mu)} < C \ \forall p > C \tag{2}$$

Note that $p > C$ is a necessary assumption to ensure agents have positive probability of participating in a cascade.

Observations

The first result is the same as the one observed in Bikhchandani et al. (1992):

Observation 1

To induce an agent to reconsider their private information during a cascade, an external signal need not be strong or certain; i.e., an agent may overturn a cascade even if $q < p$ and even if the agent's signal does not match the external signal.

This is relevant because Community Notes are crowdsourced, rather than verified facts. Thus, a Community Note need not be entirely convincing to overturn an informational cascade; rather, their mere presence may cause Twitter users to reconsider their actions before deciding to like or retweet a post.

The second result also follows fairly simply from Equation 1:

Observation 2

As the cost of belief, C, decreases, the strength of an external signal, q, must increase proportionally to overturn a cascade.

To verify this, verify that $\frac{\partial}{\partial q}P(V = 1 \,|\, x_j = L, y = L, H \text{ cascade}) < 0$ for $q \in (0, 1)$ (see Appendix 3). In turn, Observation 2 identifies two points of contention for research on Community Notes.

First, the perception of the external signal's prior, q, matters: in particular, since q is a function of the content of the Community Note, we can examine whether certain features of the note (for example, whether the note cites a source, cites a certain partisan source, and challenges a partisan view) affects the note's effectiveness.

Second, the cost of belief C matters. Any action on Twitter is rather costless, so the degree to which a note is effective is limited by the simplicity of the action *believe*. In particular, since we measure multiple outcomes (likes, retweets, comments), it is easy to imagine how the effects of Community Notes are conditional on different C; that is, it may well be the case $C_{like} < C_{retweet} < C_{comment}$. This may also explain why holding q and p constant, a user may choose to, for example, like but not retweet a statement.

Finally, we see that if $p = \frac{1}{2}$, Equations 1 and 2 are equivalent. So, we have our final observation:

Observation 3

If an agent has weak priors regarding given information $p \to \frac{1}{2}$, their choice to either believe or reject information increasingly depends solely on q and C.

This observation leads to another potential point about Twitter users in Community Notes: the more complicated or foreign an issue $p \to \frac{1}{2}$, the more uninformative their own be-

liefs, and so the more agents will base their decisions solely on whether they trust the content/citations/partisanship of a Community Note and the relative cost of their action.

Hypotheses

From the model above, as well as the remaining literature of media and politics, we can form several hypotheses to analyze in the data.

Hypothesis 1

A noted Tweet will receive substantively fewer engagements than non-noted Tweets. In particular, retweets may decrease more relative to likes.

Hypothesis 1 is derived from Observations 1 and 2. If the data indeed show that signals from Community Notes are sufficiently strong, we may expect Tweet engagement to decrease at a higher rate (see Figure 3).

Likewise, if we make the assumption that retweets are costlier actions than likes ($C_{like} < C_{retweet}$), we can expect an agent to more strongly reconsider their action of retweeting relative to the action of liking a Tweet.

Hypothesis 2

The effects of a Community Note on user engagements are stronger if the Tweet and the cited source have similar partisan biases.

Hypothesis 2 is derived from Observation 1. Consider that a Twitter user's (or in this case, a US politician's) followers are by and large of the same partisan group. Then, if a Community Note cites a partisan source consistent with that bias, an agent's perception of that external signal increases accordingly ($q \uparrow$), and its ability to overturn the cascade increases in turn.

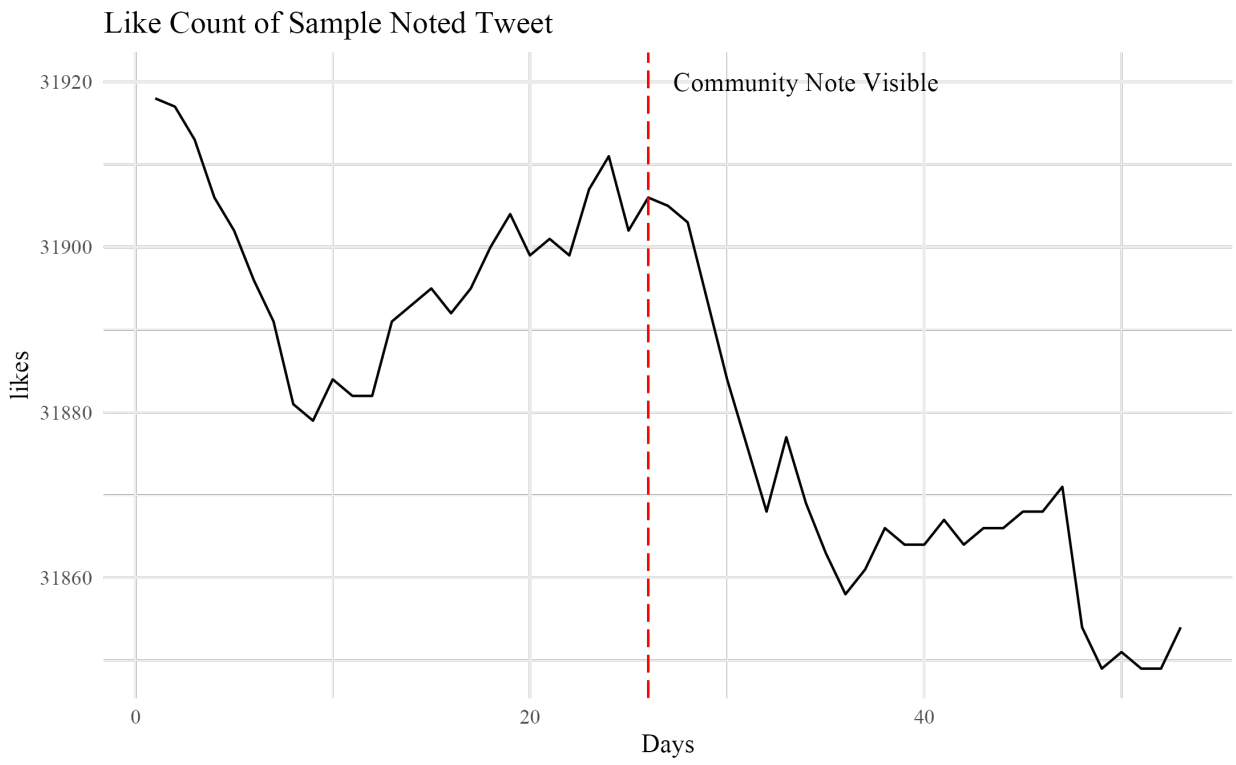

Figure 3a: Likes over Time for 1 Sample Tweet from Dataset

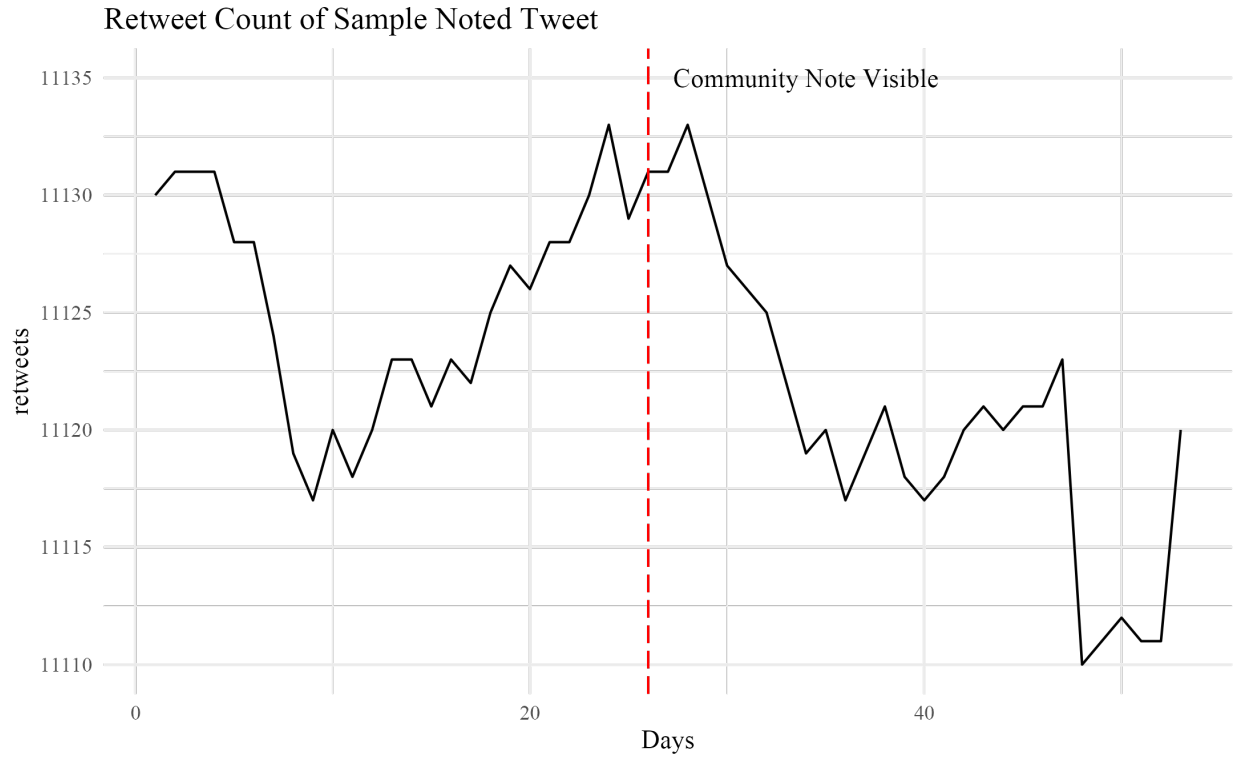

Figure 3b: Retweets over Time for 1 Sample Tweet from Dataset

For a hypothetical example, consider if Democratic Senate Majority Leader Charles Schumer's Twitter account posted a Tweet containing misinformation to his majority-Democratic Twitter followers. If a Community Note was posted to the Tweet, then by Hypothesis 2, followers would more strongly believe a note that cites CNN, a traditionally left-leaning source, as opposed to Fox News, a traditionally right-leaning source.

Hypothesis 3

Both the direct effects of a Community Note (Hypothesis 1) and conditional effects via partisan bias (Hypothesis 2) will be weaker for more salient political issues.

Hypothesis 3 is derived from Observation 3. That is, the overall effects of Community Notes are also conditional on the agent having weak priors; as an agent assumes they have greater knowledge on a topic ($p \to 1$), the ability of a Community Note to alter an agent's beliefs becomes much more difficult. In practice, I consider this a measure of political knowledge and polarization: that is, the greater preexisting knowledge an agent has on an issue, coupled with greater confidence in that knowledge, the less likely they are to defer to a Community Note. Notably, this is also in stark contrast to the result obtained in observational evidence by Allen et al. (2021) for a similar intervention, where the authors find those with greater political knowledge seem more likely to believe the content of a crowdsourced fact-check.

Data & Analysis

Data Collection

Data were collected using the public data releases of the Twitter Community Notes program[5], along with the Twitter API v2[6]. Data collection began October 27, 2022 and ended December 21, 2022. On each day of data collection, updated datasets were downloaded from the Twitter Community Notes public data release. Then, using a column for Tweet IDs that were linked to submitted notes, a program was used to send requests to the Twitter API to collect data on Tweets corresponding to individual notes (Kim 2023). This program was run at approximately 10am EST each day of data collection, with few exceptions. Notably, the public data releases of the Community Notes were released on a time-lag: for example, on data collection day November 5th, the program collected data on Tweets as of 10am on November 5th, using data on notes released on November 2nd. To account for this lag, dates for all collected datasets were adjusted accordingly, such that outcome counts and treatment intervention times correspond with each other.

The Twitter Community Notes public data release includes four datasets updated daily: Notes, Ratings, Notes Status History, and User Enrollment[7]. Functionality for User Enrollment was added midway-through the data collection period, so data from that dataset were collected whenever available but otherwise ignored. Between the other three datasets, variables of interest include (from the Notes dataset) note text, corresponding Tweet ID, (from the Ratings dataset) number of times rated helpful/unhelpful, number of ratings submitted from Twitter users who agree/disagree with the Tweet, and (from the Notes Status History dataset) whether the note was rated as helpful/displayed at a given time.

[5]see "Community Notes: Download data," Twitter (2022) https://twitter.com/i/birdwatch/download-data.

[6]see "Twitter API Documentation," *Development Platform*, https://developer.twitter.com/en/docs/twitter-api.

[7]see "Community Notes: Download data," Twitter (2022) https://twitter.com/i/birdwatch/download-data.

From the Twitter API, using Tweet IDs that correspond to noted Tweets from the public data releases, general metrics at the time the program was run were collected. These include number of likes, retweets, replies, and quotes, as well as the Tweet author's User ID and text of the Tweet. Altogether, data was collected on 26,904 Tweets, on which 44,410 Community Notes were submitted during the study period.

To test my hypotheses, I take two different approaches in analyzing these data. First, I use the full dataset, mentioned previously, to conduct a general evaluation of Community Notes on engagement measures. Second, I use a subset of the dataset to limit the scope of analysis to political Tweets. To achieve this subset, I filter the dataset to include only Tweets posted by current politicians. This filter was applied using the Twitter User IDs of politicians, as sourced from the ProPublica Politwoops database[8]. These data were joined with the public data releases, along with a variable for time of collection, to create the subsetted dataset for analysis, with 2,630 unique Tweets and 4,256 corresponding Notes.

[8]see "Politicians tracked by Politwoops," ProPublica Data Store (2019), https://www.propublica.org/datastore/dataset/politicians-tracked-by-politwoops.

Month-Year	Rank 1	Rank 2	Rank 3
Jan 2021	trump	2020	cruz
Feb 2021	trump	covid	2020
Mar 2021	trump	covid	election
Apr 2021	covid	trump	election
May 2021	trump	election	jan
Jun 2021	trump	election	earthquakes
Jul 2021	covid	trump	earthquakes
Aug 2021	covid	earthquakes	afghanistan
Sept 2021	covid	2020	election
Oct 2021	covid	earthquakes	vaccines
Nov 2021	covid	rittenhouse	earthquakes
Dec 2021	covid	earthquakes	vaccines
Jan 2022	covid	earthquakes	account
Feb 2022	ukraine	trump	earthquakes
Mar 2022	ukraine	russian	books
Apr 2022	account	musk	earthquakes
May 2022	abortion	bill	politics
Jun 2022	covid	abortion	court
Jul 2022	covid	abortion	trump
Aug 2022	trump	inflation	covid
Sept 2022	trump	migrants	covid
Oct 2022	covid	trump	musk
Nov 2022	election	elon	musk
Dec 2022	musk	covid	trump
Jan 2023	covid	tax	bill
Feb 2023	earthquakes	covid	trump

Table 1: Most Commonly Mentioned Unique Terms, Jan 2021 - Feb 2023

Table 1 shows the most commonly mentioned unique terms in user-submitted Community Notes. There are several patterns of note: first, Community Notes appears to be remarkably consistent with both political and current issues: topics like COVID-19 and vaccines appear consistently during the ongoing pandemic, and issues like the January 6th riots ("Trump", "Cruz"), the withdrawal of troops from Afghanistan ("Afghanistan"), the Ukraine invasion ("Ukraine", "Russia"), the Dobbs decision ("abortion"), and more are consistent with the timeline of real-world events and high-visibility topics for misinformation[9]. Thus, the data

[9]"earthquakes" refers to misinformation regarding prediction of earthquakes, a common general source of misinformation on social media.

reveal that even with such a recent release, the study period appears to effectively capture data that is both relevant to current events and related to common sources of misinformation.

Treatment & Matching

The treatment variable of interest is whether a Community Note was shown on a Tweet, to be used in an event-study difference-in-difference design. In this context, a "treated Tweet" is a Tweet that had such a corresponding note, and vice-versa. Since multiple notes can be active on a single Tweet at a given time, I pick a single "representative note" for each Tweet to indicate at what time t treatment has occurred. For each treated Tweet, the representative note is chosen as the note that was picked to be displayed earliest for each Tweet in the given study period. Weaknesses of this approach are discussed in the Limitations section.

Because our treatment of interest, whether a given Tweet received a visible note, is determined entirely by user ratings on notes, we can employ matching methods on the user ratings to estimate whether Tweets have similar probabilities of treatments (that is, the propensity score, P(treatment), is solely a function of the ratings data, of which we have exact counts). I use the Ratings public data release to get data on the number of reviews left on a note, the number of Helpful/Somewhat Helpful/Not Helpful ratings, and the number of ratings left by individuals who Agree/Disagree with the note. Using these variables, I match representative notes to control notes (notes that were submitted but never shown on untreated Tweets) to identify notes (and associated Tweets) that were substantially similar on the margin for treatment/non-treatment.

I use MatchIt, a software package implemented by Ho et al. (2011), to match units 1-to-1 via Mahalanobis distance. Mahalanobis distance is a measure of unit similarity that uses covariates to calculate a measure resembling Euclidean distance[10] (Stuart 2010). In particular, Mahalanobis distance is known to have strong performance in covariate balance

[10]More specifically, for two units X_i and X_j and sample covariance Σ, Mahalanobis distance $D_{ij} = (X_i - X_j)'\Sigma^{-1}(X_i - X_j)$ (Stuart 2010).

when matching with relatively few covariates (Stuart 2010), as in this study. Tables 2 and 3 indicates the covariate balance achieved through Mahalanobis distance matching (for additional covariate balance measures, see Appendix, Tables 10-11).

	Treatment Means	Control Means	Std. Mean Diff.
Number of Ratings	33.75	32.46	0.04
Helpful Ratings	25.07	24.13	0.03
Not Helpful Ratings	4.78	4.67	0.02
Somewhat Helpful Ratings	3.33	3.10	0.05
Agree with Note	0.50	0.49	0.00
Disagree with Note	0.03	0.03	0.00

Table 2: Covariate Balance of Matches, All Tweets

	Treatment Means	Control Means	Std. Mean Diff.
Number of Ratings	38.47	38.84	-0.01
Helpful Ratings	25.84	25.53	0.02
Not Helpful Ratings	7.47	8.21	-0.10
Somewhat Helpful Ratings	4.53	4.47	0.01
Agree with Note	0.58	0.58	0.00
Disagree with Note	0.05	0.05	0.00

Table 3: Covariate Balance of Matches, Tweets by Politicians

Summary Statistics

	n	Mean Likes	SD Likes	Mean Retweets	SD Retweets
Control, pre-Treat	422	19274.33	50284.32	3531.45	8407.87
Control, post-Treat	633	19256.19	50226.95	3527.66	8390.25
Treatment, pre-Treat	422	33494.17	88502.76	4869.35	8670.11
Treatment, post-Treat	633	33451.43	88399.06	4866.03	8662.39

Table 4: Summary Statistics, All Tweets

For the full dataset of study, in order to balance the use of a difference-in-difference model to draw inferences against the overly constraining the study sample, I study only Tweets with at least 2 pre-treatment observations and 3 post-treatment observations. This leaves 422

candidate Tweets for study (211 treated, 211 control). Summary statistics for this dataset are seen in Table 4.

	n	Mean Likes	SD Likes	Mean Retweets	SD Retweets
Control, pre-Treat	44	31234.63	34282.78	5129.24	4693.24
Control, post-Treat	66	31185.60	34083.65	5119.21	4663.17
Treatment, pre-Treat	44	37736.04	50054.86	7424.74	9085.64
Treatment, post-Treat	66	37657.11	49726.09	7417.19	9043.34

Table 5: Summary Statistics, Tweets by Politicians

Next, for the subset of political Tweets, the dataset is filtered further for only Tweets by politicians (as defined in the Politwoops dataset). After this filter, we are left with 44 observations (22 treated, 22 control). Summary statistics for this subset of the datasets are available in Table 5. Notably, constraints on both datasets severely limit the number of analyzable Tweets; this is discussed further in the Limitations section.

The basic summary statistics presented for both datasets reveal several characteristics. First, the data observed here experience extreme variability with respect to the outcome variables (likes, retweets). This is both expected yet problematic: since Community Notes are crowd-sourced and submitted at the discretion of users, note authors can ultimately choose to prioritize both high-profile Twitter accounts as well as Twitter accounts with weak follow-ings.

Second, in both treatment and control groups, the mean number of likes and retweets are trending downward. Though the exact cause of this is harder to ascertain, given the high variability of the data, this decrease nonetheless suggests that in both the treatment and control groups, Tweets have already reached peak engagement and are trending downwards.

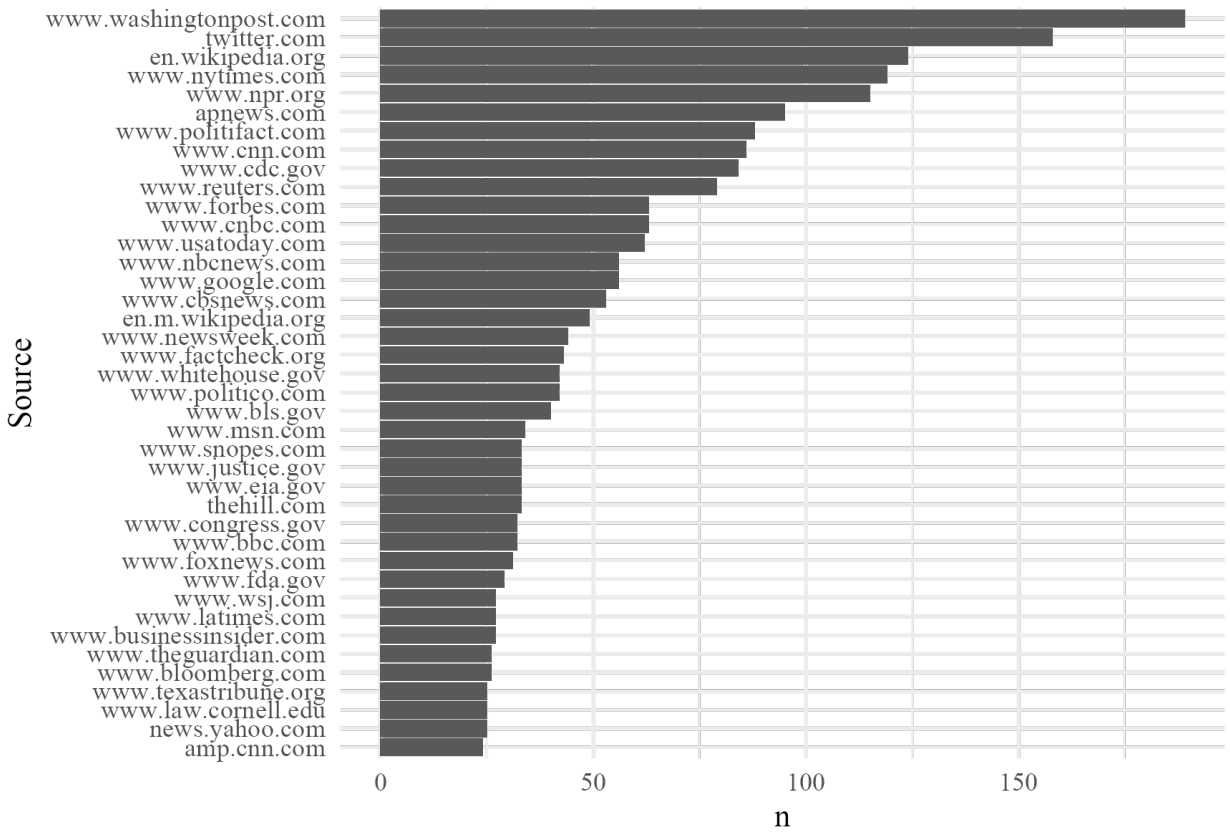

Figure 4: Most Commonly Cited Sources on Notes on Political Accounts

In addressing Hypothesis 2, I note several revealing characteristics of cited sources in the Twitter Community Notes program. As seen in Figure 4, the majority of cited sources are fairly centrist, though notably left-leaning, with the most commonly cited news sources being the Washington Post and the New York Times.

Still, the most commonly cited source that is considered substantially left-leaning is CNN; the most commonly cited source that is considered substantially right-leaning is Fox News (Pew Research Center 2014). Notably, citations of CNN nearly tripled those of Fox News; though this is consistent with survey studies that Twitter users are more likely to be left-leaning, the magnitude of this disparity is far greater than that observed in survey studies (where about 60% of Twitter users identify as left-leaning while 35% identify as right-leaning) (Wojcik and Hughes 2019). Still, to simplify my analysis of Hypothesis 2, I focus my efforts

on analyzing the interaction between citations of CNN and Fox News as partisan sources.

Analysis

To test Hypothesis 1, I analyze the data using a fixed-effects difference-in-difference regression model on both the full dataset and the subsetted dataset. Depending on the outcome variable, one of the following regressions was run (for a given matched treatment-control pair w):

$$Likes_w = \beta_0 + \beta_1(treatmentGroup_w) + \beta_2(treatedAtTime_w) + \beta_3(DiD_w) + \gamma_w + \epsilon_i \quad (3)$$

$$Retweets_w = \beta_0 + \beta_1(treatmentGroup_w) + \beta_2(treatedAtTime_w) + \beta_3(DiD_w) + \gamma_w + \epsilon_i \quad (4)$$

where $treatmentGroup$ is an indicator for whether the Tweet received a note, $treatedAtTime$ is whether the Tweet is in a period where a representative note would have been received, DiD is the interaction $treatmentGroup * treatedAtTime$, and γ_w is fixed effects for a treatment-control pair. In estimating these models, I use standard errors that are robust to heteroskedasticity.

A cursory look at the data reveal several patterns of note (Figures 5 and 6). First, the data (after de-meaning) in Figures 5b and 6b have different implications for the parallel trends assumption; the former appears to be consistent with it, while the latter appears to violate it. The ramifications of this are discussed in the Limitations section. Still, the control Tweets do behave as anticipated in comparison to treatment Tweets; that is, likes on control Tweets appear to decline steadily and consistently, whereas likes on treatment Tweets behave erratically, dropping substantially at certain periods and overall experiencing a higher magnitude of effect. Nevertheless, from data visualization alone it is unclear whether the effect is significant (a strong enough decline in slope) or substantive (a strong decline overall).

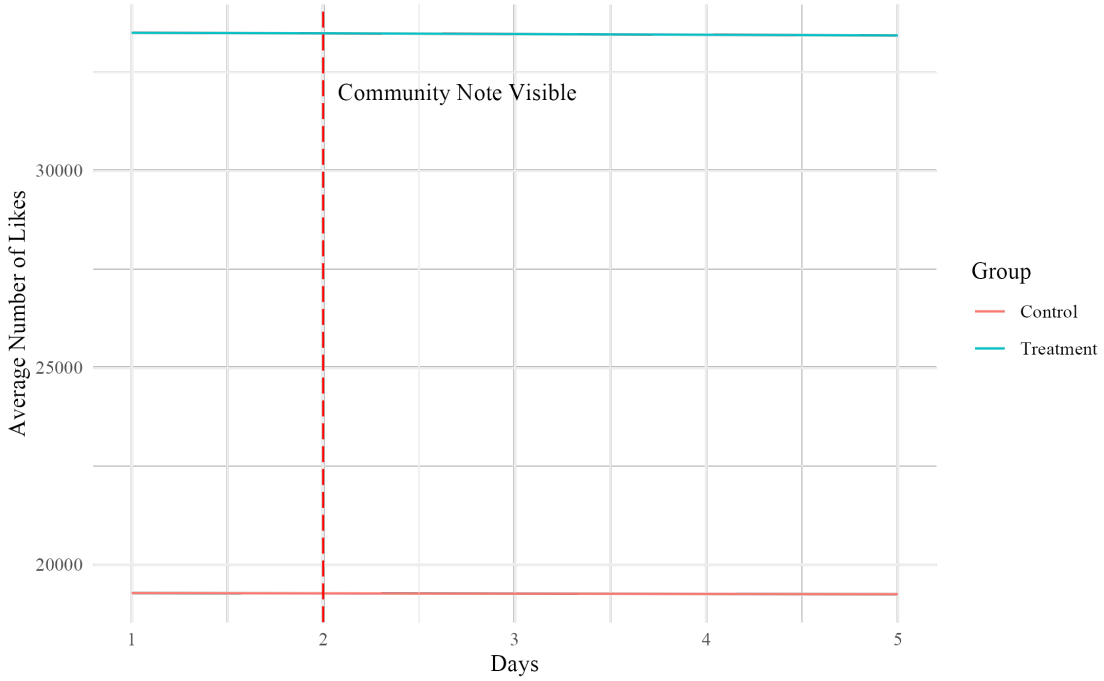

Figure 5a: Average Likes over Time, Treatment vs. Control, All Tweets

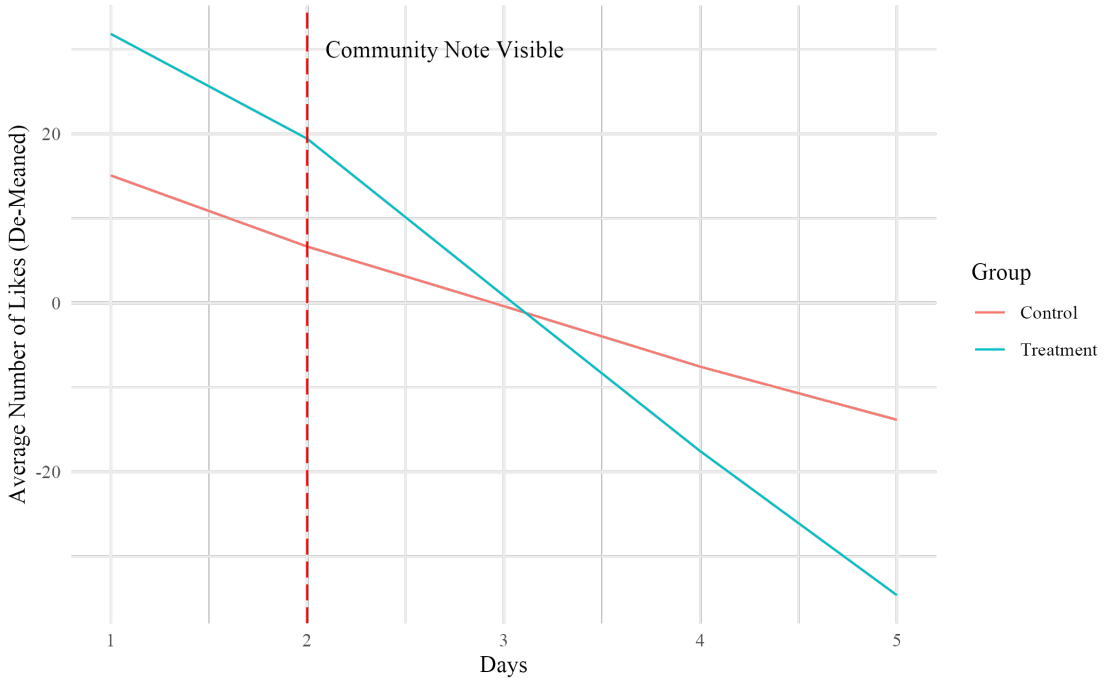

Figure 5b: De-Meaned Average Likes over Time, Treatment vs. Control, All Tweets

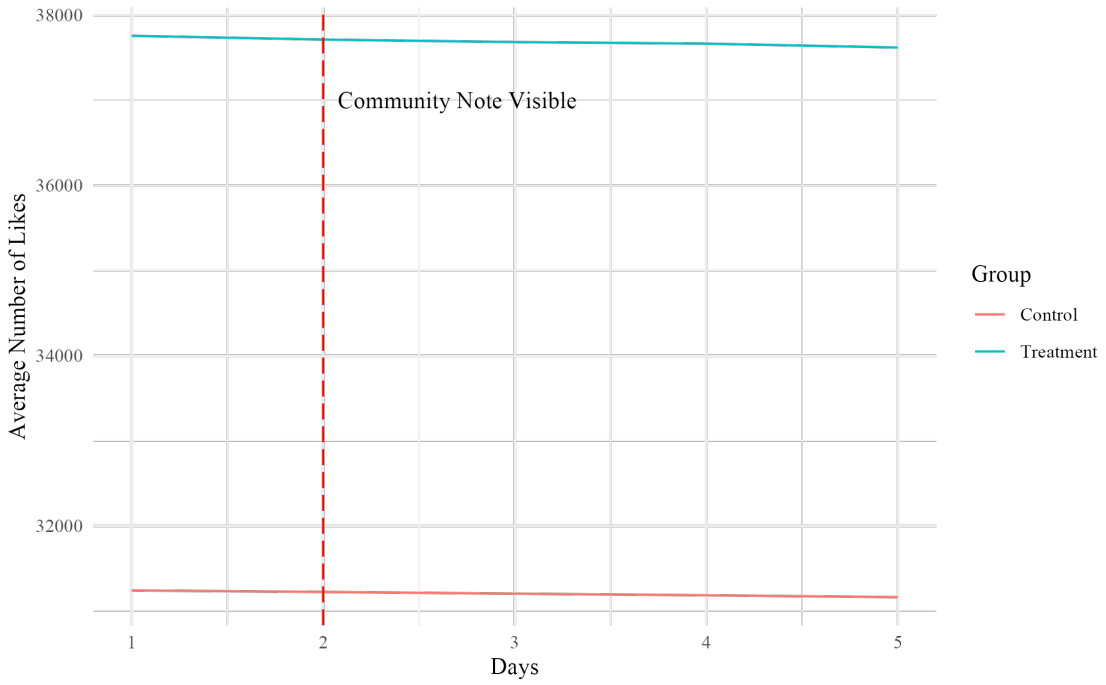

Figure 6a: Average Likes over Time, Treatment vs. Control, Tweets by Politicians

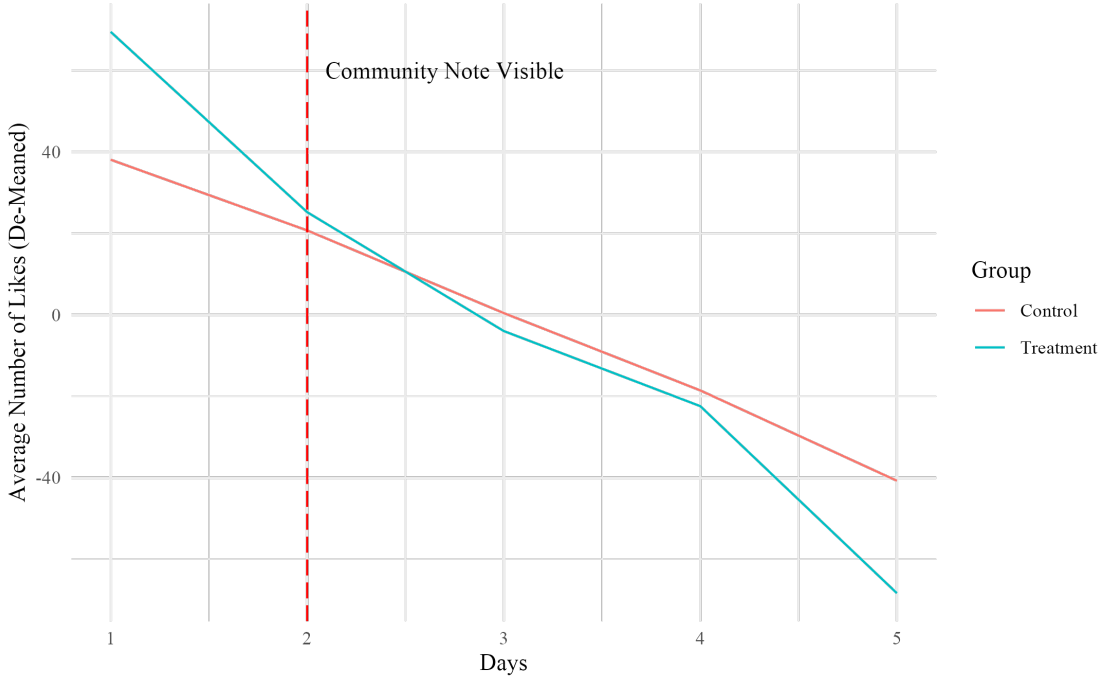

Figure 6b: De-Meaned Average Likes over Time, Treatment vs. Control, Tweets by Politicians

Due to the constraints placed on observable Tweets, a direct test of Hypothesis 2 using the fixed-effects difference-in-difference model described in Equations 3 and 4 was not possible. Instead, I indirectly test Hypothesis 2 by estimating the length of time a note was active, which I argue reflects greater potential for certain notes to have effects on individuals moreso than others. I estimate this by:

$$timeActive = \beta_0 + \beta_1(CNN) + \beta_2(Fox) + \beta_3(Dem) + \beta_4(Dem * CNN) + \beta_5(Dem * Fox) + \epsilon_i \tag{5}$$

where *timeActive* is the number of periods t (measured in days) the note was active, CNN is an indicator for whether a note mentioned cnn.com, Fox is an indicator for whether a note mentioned foxnews.com, and Dem is an indicator for whether the Tweet was posted by a Democratic politician. Because the data for *timeActive* were left-skewed, I applied a log transformation and estimated another model as follows:

$$log(timeActive{+}1) = \beta_0 + \beta_1(CNN) + \beta_2(Fox) + \beta_3(Dem) + \beta_4(Dem{*}CNN) + \beta_5(Dem{*}Fox) + \epsilon_i \tag{6}$$

In both Equations 5 and 6, I use heteroskedasticity-robust standard errors.

Though I do not provide a direct test of Hypothesis 3, I instead indirectly test this hypothesis by contrasting the results from estimating Equations 3 and 4 on the full dataset and the subset of data tweeted by politicians. The success of this approach in testing Hypothesis 3 depends on the increased likelihood that a politician tweets about issues that agents have stronger priors p about (i.e., given $p = P(\text{state is } H)$, $P(\text{state is } H \mid \text{politician Tweet}) > P(\text{state is } H \mid \text{non-politician Tweet})$). Though I do not test this empirically, I assume this to be true, given the likely correlation between politicians tweeting on a topic and agents having greater knowledge of that topic ($Cov(p, \text{follows Politician}) > 0$).

Results

Table 6: Equations 3 and 4, Estimated on All Tweets

	Dependent variable:			
	Likes	Retweets	Likes	Retweets
	(1)	(2)	(3)	(4)
treatmentGroup	14,219.840***	1,337.908**	14,219.840**	1,337.908
	(4,953.894)	(587.778)	(7,001.308)	(814.497)
treatedAtTime	−18.135	−3.786	−18.135***	−3.786***
	(3,158.407)	(527.905)	(3.478)	(1.154)
DiD	−24.600	0.465	−24.600***	0.465
	(6,393.484)	(758.491)	(8.202)	(1.389)
Constant	19,274.330***	3,531.446***		
	(2,447.221)	(409.192)		
Fixed Effects?	No	No	Yes	Yes
Observations	2,110	2,110	2,110	2,110
R^2	0.010	0.006	0.019	0.013
Adjusted R^2	0.008	0.005	−0.091	−0.098
Note:				*p<0.1; **p<0.05; ***p<0.01

In my test of Hypothesis 1, estimating the models in Equations 3 and 4, as seen in Table 6, I find that the intervention of a Community Note causes an average decrease of about 24.6 likes on a Tweet ($p < 0.01$), though there is no similar effect on retweets ($p > 0.1$). By design of the difference-in-difference estimation and the matching procedure, I interpret these results causally; issues with causal inference are discussed further in the Limitations section. Still, at face value, these results provide preliminary evidence that Hypothesis 1 holds true: the introduction of a Community Note indeed causes a significant decline in like counts on a Tweet.

Table 7: Estimating Equations 3 and 4 on Tweets by Politicians

	Dependent variable:			
	Likes	Retweets	Likes	Retweets
	(1)	(2)	(3)	(4)
treatmentGroup	6,501.405	2,295.503	6,501.405	2,295.503
	(9,125.088)	(1,538.086)	(11,561.490)	(2,067.443)
treatedAtTime	−49.030	−10.031	−49.030**	−10.031**
	(6,651.592)	(910.370)	(22.751)	(4.771)
DiD	−29.900	2.484	−29.900	2.484
	(11,768.750)	(1,984.727)	(35.002)	(2.647)
Constant	31,234.630***	5,129.237***		
	(5,156.347)	(705.893)		
Fixed Effects?	No	No	Yes	Yes
Observations	220	220	220	220
R^2	0.006	0.025	0.014	0.054
Adjusted R^2	−0.008	0.012	−0.107	−0.063
Note:				*p<0.1; **p<0.05; ***p<0.01

However, when the data is constrained to Tweets by politicians (Table 7), I find no significant causal effect of the Community Notes intervention on either the number of likes or number of retweets garnered by a Tweet ($p > 0.1$). This result speaks to some of the limitations of the Community Notes program, as well as lending support to Hypothesis 3: on issues that are more salient and/or on Tweets by political elites/politicians, the program is unlikely to have any significant causal effect.

Table 8: Estimating Equations 5 and 6

	Dependent variable:	
	Active Period	log(Active Period + 1)
	(1)	(2)
Cites CNN	−0.252	−0.019
	(0.490)	(0.056)
Cites Fox	1.183	0.130
	(1.149)	(0.116)
Dem	0.927***	0.097***
	(0.247)	(0.023)
Dem * Cites CNN	2.487	0.178
	(2.199)	(0.179)
Dem * Cites Fox	−3.351**	−0.347**
	(1.230)	(0.121)
Constant	0.897***	0.096***
	(0.120)	(0.011)
Observations	4,223	4,223
R^2	0.006	0.006
Adjusted R^2	0.004	0.005
Note:		*p<0.1; **p<0.05; ***p<0.01

Finally, in estimating Equations 5 and 6 in Table 8, I find several results of note. First, Democrats on average have notes that last about .927 periods longer (about 1 day), at a significant level ($p < 0.01$). Second, citing CNN is not associated with any significant change in the length of time a note is active for either Democrats or Republicans ($p > 0.1$). Finally, and in contrast to CNN, citing Fox News is associated with a significant decline in the length of time a note is active for Democrats only, with an estimated decline of about 3.35 periods (or about 3.35 days) ($p < 0.01$). Due to a lack of relevant controls available in the data and significant potential for confounding, I do not interpret these results causally. Still, the

results (particularly the coefficient on Dem * Cites Fox) lend some support to Hypothesis 2.

Discussion

Community Notes Effects At-Large

As seen in Table 6, the data reveal that the Community Notes program on average causes a decrease of about 24 likes. Notably, a difference of 24 likes does not appear substantive: Tweets regularly reach thousands of likes, and an arguably more direct measure of information spread, retweets, appears to have no significant effect. Still, this effect must be evaluated in context: first, even Twitter users and political "influencers" with thousands of followers struggle to reach 24 likes, and such a penalizing decrease on these users would effectively "zero-out" any influence they gain via Twitter (see Figure 7). Second, predicting the viral spread of Tweets without further knowledge of Twitter's internal recommendation models is difficult to do with precision, much less predicting the possible substantive effect of a 24 like reduction in any part of an informational cascade. Finally, in the dataset itself, most of the Tweets observed have substantially fewer likes than the mean, and are thus Tweets for which 24 likes are likely to have a much greater impact (see Figure 8). Altogether, I contend that this causal effect is somewhat substantive, and in any case, the comparative statics of this effect, rather than the point estimate, are more important in evaluating this effect.

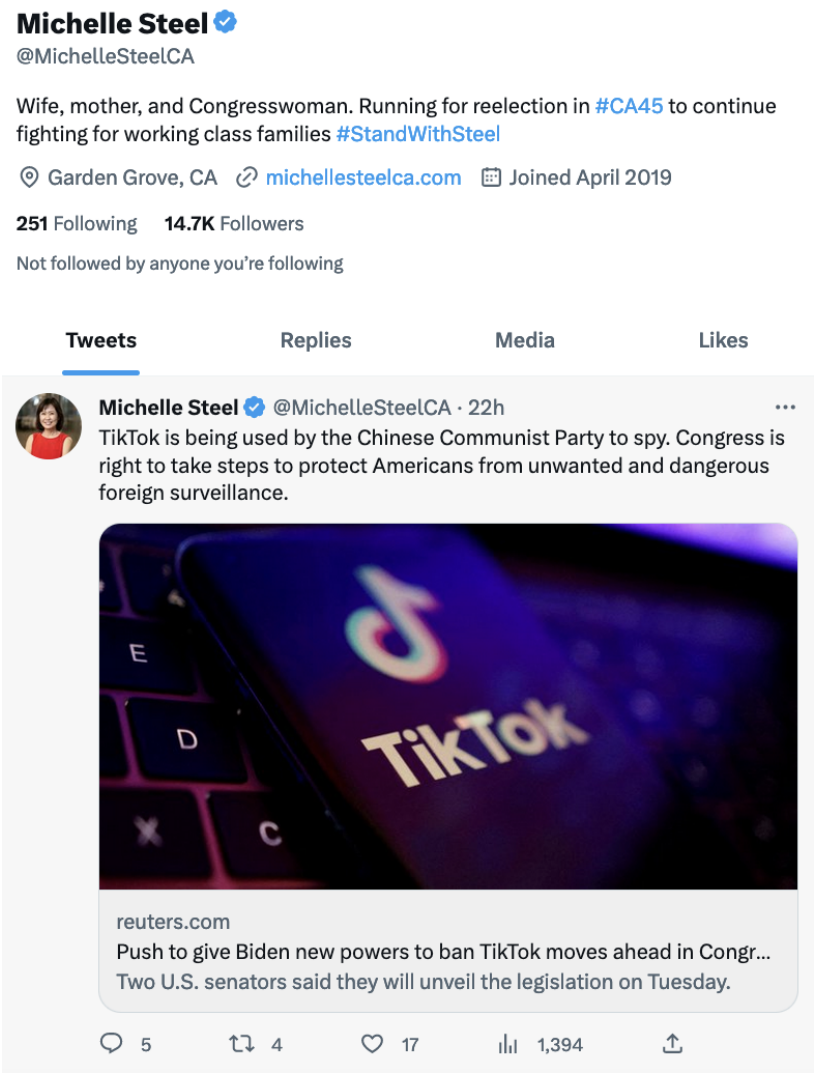

Figure 7: @MichelleSteelCA Twitter Engagement

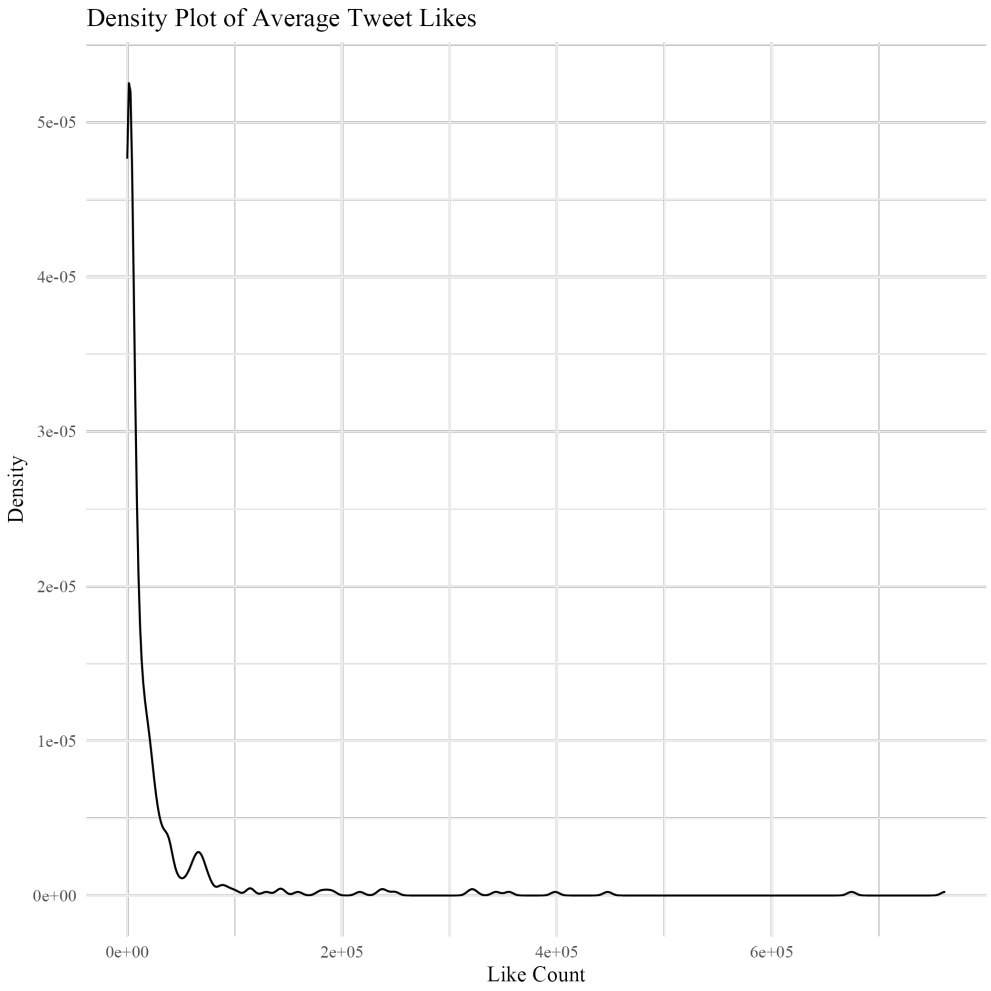

Figure 8: Density Plot of Likes in Dataset Tweets

More generally, the significant causal effects observed in Community Notes are consistent with previous evidence of similar interventions in the literature. Most preeminently, the results confirm that evidence of lower engagement observed by Wojcik et al. (2022), Twitter's own publication on the topic, holds true outside of an experimental setting. Similarly, and in concert with results found by Barrera et al. (2019) and Clayton et al. (2019), "soft" content moderation schemes likely have some substantive effect on social media engagement. Likewise, as previously evidenced from Table 1, Community Notes itself appears remarkably effective in keeping up with both current-day events and common sources of disinformation. All told, the significant effect here is both consistent with prior research on content modera-

tion, appears relevant to modern sources of political disinformation, and carries considerable ecological validity, an important trait given many previous studies that operated only in an experimental setting.

This effect is coupled with several other desirable attributes of the program: the implementation of the program is open-source, and the very design of the program allows users greater control over content moderation on the platforms they use. Thus, not only is the program low-cost, but it may be also effective in addressing consumer trust in social media platforms over viewpoint censorship (Vogels et al. 2020). Though these results are more nuanced (as will be discussed shortly), for the social planner, Community Notes approaches a "free lunch", providing effective content moderation that is unlikely to be controversial as well as providing open-source code for cheap, large-scale deployment.

Finally, it is noteworthy to point out that the causal effect of Community Notes is only limited to likes, not retweets. Though this is contrary to my expectations in Hypothesis 1, it is not altogether unexpected: the functions of retweets and likes are very different on the Twitter platform, and so it is difficult to express this conditional effect solely through a single scalar C. In any case, since both likes and retweets are measures of engagement that are both significant inputs to Twitter's recommendation systems, I do not scrutinize this difference in effect.

Politicians & Community Notes

In contrast to the previous result, Table 7 reveals that Community Notes has no significant causal effect on engagement for Tweets by politicians. Naturally, this lends itself to troubling conclusions about Community Notes; after all, politicians are undoubtedly some of the most influential social media users, and a key unstated benefit of Community Notes itself is to allow for hands-off content moderation of these high-profile but controversial individuals by social

media platforms[11]. Turning back to the strategic model, social media platforms undeniably face a careful balancing act in content moderation: platforms must maintain a reputation of unbiased promotion of freedom of speech (Kemp and Ekins 2021) but ultimately garner greater profits by implementing stronger content moderation policies (Liu et al. 2022). This tradeoff has only grown in significance: Twitter likely faces an extraordinary cost should its lowered reputation attract policymaker scrutiny, in the form of platform immunity reforms via the increasingly controversial Communications Decency Act Section 230 (Alizadeh et al. 2022). Community Notes is undoubtedly an end towards Twitter's strategy, providing content moderation in a way that is not directly managed by the platform itself. It then appears concerning that Community Notes is ineffective towards subjects whose indirect moderation poses the greatest benefit for the platform (i.e. politicians).

Still, the actual nature of politicians' interactions with Community Notes is likely more nuanced. First, a major benefit of crowdsourced content moderation is its ability to address a wide range of authors and content types, leaving resources for professional fact-checkers to address more high-profile cases. That is, instead of devoting resources to moderate both general user A and politician B, Twitter can rely on Community Notes to moderate user A and dedicate more resources to scrutinizing and evaluating content by politician B. Second, consistent with Hypothesis 3, it is likely that individuals following politicians' Twitter accounts were unlikely to be convinced by such a content moderation scheme in the first place (the implications of which are discussed shortly).

Partisan Priors & Political Effects

By contrasting the two effects above, we see the importance of priors in this model of Bayesian consumers. Thus, taking into account the previous results, as well as further research on Bayesian media consumption, I outline the ramifications of the results above and the role of

[11] see "Permanent suspension of @realDonaldTrump," Twitter (2021).

priors on real-world political effects.

First, the models above reveal the benefits of Community Notes for agents with weaker priors. In another model of Bayesian consumers, Gentzkow and Shapiro (2011) reveal that media sources face much stronger incentives to distort truths if Bayesian consumers have weaker priors or lack methods of verifying signals. Ultimately, then, if Community Notes is indeed effective only for agents for weaker priors, Community Notes effectively protects the population most vulnerable to misinformation in the first place. Indeed, Aslett et al. (2022) confirm that such soft content moderation schemes are more effective in heavy consumers of disinformation. In turn, the evidence here thus suggests that in the general space of political communications, Community Notes may discourage the spread of misinformation in the first place, since disseminators of misinformation targeting those with the weakest priors must now weigh being exposed by the signals posed by Community Notes.

Second, as pointed out previously, individuals following politicians' Twitter accounts were likeliest to have the strongest priors regarding political issues and were thus unlikely to have any substantial reaction to Community Notes in the first place. Indeed, Community Notes seems best poised to aid those with weaker priors (and therefore most vulnerable to misinformation), so the null effect observed here is unsurprising. Likewise, even if substantive Bayesian updating did occur, evidence from Barrera et al. (2022) indicates it may not have translated to a change in engagement or belief in a policy. All told, although we observe a null effect conditional on Tweets by politicians, under the assumption that agents following politicians are more likely to have stronger partisan priors, the agents who are unaffected are those for which an appreciable impact of media on electoral choice are least relevant.

Together, these observations leave wide open the possibility that Community Notes and similar crowdsourced content moderation schemes may influence real-world electoral outcomes. Though evidence on partisan priors, policy choice, and media consumption is far from conclusive[12], the results indicate that if agents' priors correlate with susceptibility to change

[12]see Fujiwara et al. (2021); Guess et al. (2021); Malzahn and Hall (2023).

in policy belief, the Community Notes program may have a substantial impact on electoral choice for agents with the weakest priors.

Partisan Trust in Notes

As seen in Table 8, and in line with aforementioned strategic models of strategic disinformation, the data suggest effects of Community Notes are also conditional on both the partisan affiliation of Tweet authors and the perception of a source by note viewers. In particular, the presence of Fox News as a cited source on a Tweet by a Democratic author appears to decrease trust in the Community Note drastically, with such notes lasting more than 3 days less on average.

Though these results cannot be interpreted causally, they nonetheless shed further light on the strategic behaviors of agents in Community Notes, as well as the effectiveness of content moderation generally. As conjectured previously, the partisan behavior and usage of Community Notes differs across parties: notes appear to last longer on Tweets by Democratic candidates, and in general only notes on Democratic sources appear to be affected by the citation of certain media sources. This provides some evidence that Democrats may place higher trust in the Community Notes program, allowing them to linger for greater periods on Democratic politicians, but nonetheless "police" the program for sources they deem untrustworthy, such as Fox News.

Though I do not describe a definitive causal effect, the data here show that Democratic users of Community Notes display patterns of media consumption consistent with prior studies of homophily in media consumption (Halberstam and Knight 2016) and selective exposure to partisan communications (Dejean et al. 2022), more commonly known the theory of "echo chambers". Indeed, exposure to partisan news sources that challenges an agent's partisan priors is not a stated goal of Community Notes, but it remains a potential outcome (albeit, as evidenced in the data, an unlikely one). Nevertheless, given the extremely contentious

literature on echo chambers and their effects on electoral outcomes and affective polarization as a whole (Haidt and Bail Ongoing), I do not provide a definitive conclusion on whether Community Notes contributes to or helps prevent partisan echo chambers.

Interestingly, the same effect is not seen for Republicans. While this may be indicative of an overall Republican apathy towards Community Notes (in terms of our model, $q_{Rep} < q_{Dem}$), we might nonetheless conjecture why Republican voters do not display strategic incentives to reject certain sources over others. For example, Republicans may indiscriminately treat all Community Notes with suspicion, resulting in both lowered note durations relative to Democrats and lack of a conditional effect relative to partisan sources.

In fact, the results corroborate prior findings of Republican rejection of fact checking schemes (Shin and Thorson 2017) and asymmetric polarization of Republican agents relative to Democratic agents more broadly (Russell 2018), offering additional evidence that the model tested here, though based on observational evidence, is indeed the correct one. Likewise, Freelon et al. (2020) readily find evidence of strategic methods behind the spread of right-wing political disinformation, and this is consistent with my observation that notes on Republican politicians tend to persist for shorter periods of time.

Finally, there are several noteworthy nuances behind the examination of partisan trust in notes. First, by Twitter's own documentation, the initial display of a Community Note is contingent upon its approval by users of opposing partisan views ("Community Notes Guide"). Thus, partisan strategies need not occur among all Twitter users; rather, the select users involved in the initial display of notes have both considerable influence and partisan incentives that may result in outcomes consistent with those described above. Likewise, aside from restrictions on account age and engagement, the authorship of notes is not tightly controlled by Twitter. Thus, even with substantial controls on note authorship implemented by Twitter ("Community Notes Guide"), the results consistent with, for example, citation of Fox News on Democrats account may simply be due to lower-quality or "troll" notes being

authored, rather than a rejection of intended "fact-checking" that happens to cite Fox News. In any case, I find substantial evidence for Hypothesis 2: particularly among Democrats, user engagement with Community Notes is conditional on the partisan background of the source.

Limitations

Biased Estimation & Causal Inference

From an econometric perspective, several issues in model estimation emerge by design of the sampling procedure and method of analysis.

First, classic assumptions behind causal inference using a difference-in-difference design may be violated. As outlined in Lechner (2011), several classic assumptions like stable unit treatment value (SUTVA) and exogeneity apply to difference-in-difference models; by design of the experiment, these assumptions are likely to hold, and in any case are unlikely to be the primary cause of bias in the estimators.

More importantly, difference-in-difference relies on the "common trend" assumption (Lechner 2011), also called the "parallel trends assumption". Violating this assumption results in a biased causal effect estimator, and as seen in Figure 6b, this assumption indeed may not hold. However, it is difficult to evaluate the parallel trends assumption in light of the two pre-treatment periods; for example, in Figure 5b, the same plot for all Tweets in the dataset appears in line with the parallel trends assumption.

I also attempt to mitigate other sources of bias somewhat, primarily through the use of Mahalanobis distance matching. As specified previously, treatment is purely a function of ratings data and note authorship (i.e., whether a note was composed and therefore appeared in the dataset in the first place), such that we may match units with similar propensity scores with a heightened degree of accuracy. However, focusing only on these data can produce

inherent differences on other measures; for example, Tweets may have very disparate numbers of likes/retweets even if the relative number of ratings are the same (hence why de-meaned plots are necessary to represent the data). Nevertheless, I contend matching on ratings is most appropriate, since most other covariates of interest must otherwise be subjectively constructed by the researcher (for example, Tweet text, location, authorship, etc.).

All told, though I interpret these results causally, there are significant econometric challenges to evaluating these data. In particular, the study's strength in fairly exacting propensity score estimates are undermined by a weak evaluation of the parallel trends assumption and other possible covariates, problems unique to such a novel dataset. Nevertheless, I argue my approach is appropriate given constraints on available data and methods, and my results are interpreted in the context of supporting research in the field regardless.

Administration of Treatment

As mentioned in the Data & Analysis section, the treatment period is defined as the period after the first "representative note" is shown. Notably, this does not correspond with the actual treatment of interest; that is, for any time period $t > t_{treat}$ after the treatment period, there is no guarantee the representative note is actually shown at t.

This is a challenge of the design of Community Notes itself. That is, since notes are subject to ongoing review, a representative note could (in the worst case) appear for one period (day) t_{treat}, and never again $\forall\, t > t_{treat}$. In practice, the data are more nuanced: though many notes do indeed waver between being shown and not being shown, for any treated Tweet, there are usually multiple notes submitted on the same Tweet, such that at least one note is generally active on a treatment group Tweet at any given time (though not necessarily the representative note).

Thus, in context, the treatment cannot be interpreted exactingly as the "causal effect of the representative note being shown". Rather, the exact interpretation is most accurately the

"causal effect of entering the period after the representative note being shown", and this period is characterized by a possible combination of the representative note being shown, a different note being shown, or no note being shown at all. With the last possibility being less likely, I leave the interpretation as the "causal effect of the Community Notes intervention" for brevity's sake.

Constraint of the Study Sample

As mentioned previously, the analyzed sample in Tables 4 and 5 consisted of only 422 and 44 Tweets, respectively, down from 26,904 Tweets observed during the sample period. This is primarily due to the requirements of a difference-in-difference study design: that is, the analysis performed here requires data on Twitter engagement for a certain number of pre-treatment and post-treatment periods. The Community Notes dataset is helpful, here, in that data on notes are provided as soon as the note is authored, rather than only after the note itself is observable.

Still, as mentioned previously, the time-lag through which data is provided by Twitter on Community Notes places considerable constraints on the analyzable data. Notwithstanding the many notes that had already been active once the study period began, as seen in Figure 9, in the best case, notes must persist for at least 9 periods (days) before they can be analyzable as data. This includes at least 6 pre-treatment periods (during which the note cannot become visible) and 3 post-treatment periods, severely constraining the amount of data that can be analyzed using a difference-in-difference design. Adding on the constraint that Tweets must be authored by politicians, the data collected over a 2-month period shrink down to the dataset analyzed in Tables 4 and 5.

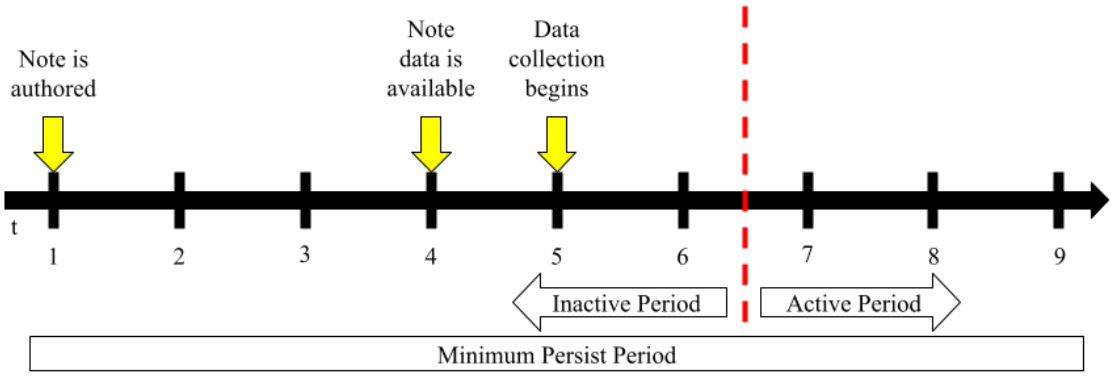

Figure 9: Observed Note Data Collection Timeline

Thus, constraints by both computational resources and Twitter data release policies contribute to the small sample size, and concerns over expanded standard errors and unrepresentative data abound. Even still, these data may represent the best possible route of analysis: sidestepping Twitter's time-lagged data shares would require active collection and logging of political Tweets on a massive scale, and barriers to usage of the Twitter API only appear to increase over time[13].

The results here therefore rely on the assumption that those Tweets exhibiting this "ideal" behavior are representative of treated Tweets in general. Though this is difficult to test empirically, in Table 9, I present a corresponding table of unique terms in Community Notes in the study sample. In particular, several topics match well to their corresponding month-year pairs (for example, "election" in Nov 2022, "rittenhouse" in Nov 2021), while others do not (for example, "ukraine" in Feb 2022 is not mentioned in the study sample).

More generally, treated Tweets in this sample are by construction more likely to have a Note that takes longer for approval. This is concerning in light of the fact that this trait may be correlated with other influential properties: for example, Tweets with especially egregious disinformation or Tweets that are quickly growing in engagement may be more likely to fall

[13]see post by @TwitterDev, February 2, 2023, 1:05AM, https://twitter.com/TwitterDev/status/1621026986784337922.

Month-Year	Rank 1	Rank 2	Rank 3
Jun 2021	president	data	trump
Jul 2021	ban	communist	democrats
Aug 2021	israeli	soldiers	alleged
Sept 2021	covid	tax	pistol
Oct 2021	administration	kids	families
Nov 2021	rittenhouse	price	prices
Dec 2021	minneapolis	1995	attack
Jan 2022	trump	president	properties
Feb 2022	president	checker	companies
Mar 2022	court	supreme	critical
Apr 2022	barack	checking	fake
May 2022	trump	rallies	resolution
Jun 2022	national	guard	button
Jul 2022	christian	court	nationalism
Aug 2022	committee	default	documents
Sept 2022	ivermectin	pubmed.ncbi.nlm.nih.gov	covid
Oct 2022	musk	vote	2020
Nov 2022	election	employees	notes
Dec 2022	whelan	paul	inflation

Table 9: Most Commonly Mentioned Unique Terms in Study Sample, June 2021 - Dec 2022

into the category of Tweets with notes that get approved quickly, in which case the effects estimated here are likely underestimated. Thus, the estimates presented in both studies (that is, Tables 6 and 7) may represent a lower bound on the true coefficient.

Determination of Variables

Important variables of interest, particularly party identification of Tweet authors and Tweet topics, are vital to the study of political communication yet imperfectly determined. In particular, the former is determined solely by whether a Tweet is posted by a politician with a given party ID; more accurately, thus, the results of this study are constrained only to current politicians, rather than other partisan influencers.

This is especially concerning given that many substantial political communications are not fostered by these current politicians, but rather former politicians or other media influencers. For example, popular Twitter users that are undoubtedly partisan and produce substantial

political content, but are not included in the dataset, include former presidents Barack Obama (@BarackObama, 133M followers) and Donald Trump (@realDonaldTrump, 87M followers), television hosts Rachel Maddow (@maddow, 10.4M followers) and Tucker Carlson (@TuckerCarlson, 5.8M followers), commentators Kara Swisher (@karaswisher, 1.4M followers) and Ben Shapiro (@benshapiro, 5.5M followers), and many more.

In turn, constraining the dataset to only include politicians in the Politwoops dataset provides a convenient, wider-use subset for analysis, but undoubtedly restricts analysis and indirectly inflates standard errors (by way of a smaller dataset, as discussed above). Nevertheless, I maintain that the focus of the Politwoops dataset provides compelling real-world analyses, particularly for tracking the behaviors of current politicians.

Developments in Community Notes

As a novel program deployed in an ever-changing social media environment, many changes have taken place in Community Notes since its release, such that it is important to note distinctions between Community Notes during the study period (October 2022 to December 2022) and Community Notes as of early 2023.

For example, in December 2022 (post-study period), Twitter implemented two major changes to the program. First, Twitter further restricted the authorship of notes, requiring that Community Note authors first engage with notes as a reviewer only, restricting the authorship of actual notes further in a move that appears to improve note quality[14]. Notably, I do not expect this to substantially affect any model estimates, though it may alter the results in Table 8 should the citation of CNN or Fox News be associated with a lower-quality note. Second, Twitter began notifying users of specific trending Tweets that may require review[15]; thus, given our concerns that Community Notes may be implemented after Tweets have

[14]see post by @CommunityNotes, December 20, 2022, 8:10PM, https://twitter.com/CommunityNotes/status/1605369962750582784.

[15]see post by @CommunityNotes, December 12, 2022, 8:48PM, https://twitter.com/CommunityNotes/status/1602480513125560320.

already peaked in engagement (Tables 4 and 5), the results from the study period may under-estimate the true effects of the current implementation of Community Notes.

Likewise, in early 2023, several Community Notes changes were implemented that may also affect the strategic interactions agents have with Community Notes. In January 2023, Twitter announced that the status of a note will be locked after two weeks[16]. Thus, Community Notes raters now face a "deadline" to either persist or reject a note, and the behavior of partisan raters may alter in turn. Additionally, in February 2023, Twitter implemented a program to notify users who have already engaged with a Tweet that the Tweet has received a note[17]. This change directly affects our cascade model; in particular, Twitter now actively encourages users to revise their actions, ultimately reducing the cost of a wrong belief C.

All told, this study of Community Notes is by no means conclusive, and no study of Community Notes can be fully accurate as the program receives continued internal and open-source development. Rather, the results are wholly accurate only as an evaluation of a certain "snapshot" of Community Notes during a particular period. Instead, I encourage the interpretation of these results as a contribution to the broader literature of "soft" content moderation schemes and crowdsourced content moderation (*a la* Aslett et al. (2022); Allen et al. (2021)), as well as providing background knowledge for further development and policymaking ventures around similar programs.

Further Study

Unique to such a novel dataset, and to the the study of political communication and disinformation on social media overall, my analysis provides several key avenues for further research.

[16]see post by @CommunityNotes, January 20, 2023, 10:40PM, https://twitter.com/CommunityNotes/status/1616641919173685254.

[17]see post by @CommunityNotes, February 21, 2023, 5:22PM, https://twitter.com/CommunityNotes/status/1628158167006994436.

First, as discussed previously, the imperfect criteria for which variables of interest are determined (in particular, the party identification of Tweet authors) prevents the determination of more generalizable results to political communication and disinformation. However, as the use of of large language models (LLMs) grows increasingly popular in the social sciences, the potential of LLMs to aid in the study of political communication and disinformation is undeniable. For example, in this study alone, further development in LLMs like PoliBERT (Gupta et al. 2019) could allow for classification of Tweets to better identify all Tweets that contain political content, that are spread by users of a certain party identification, or otherwise maliciously attempt to spread disinformation.

More generally, econometricians and other social scientists involved in text analysis and other "big data" approaches have much to learn from these data. Crowdsourced content moderation is not solely involved with the production of content moderation schemes; rather, they assist in the actual identification of content, the production of opinions regarding such content, and the responses to certain trends in political communication. Researchers and industry developers alike could identify, for example, which content a majority of US citizens finds most problematic, which pieces of disinformation are spreading most dramatically during certain periods, or even perform partisan sentiment analysis towards certain topics, disinformation, or news sources, as my cursory analysis in Table 1 attempts.

Another possible avenue of research is the effects of such content moderation methods on communications that do not in fact require moderation. For example, n their investigation of soft content moderation schemes, Zannettou (2021) observes, but does not comment further on, inconsistencies within soft content moderation schemes, including the misapplication of source credibility labels. Community Notes takes these concerns to a new level: strategic actors could author, propagate, and vote on Community Notes that are otherwise inappropriate or unnecessary, with the possibility of damaging political engagement and the reputation of the Community Notes program as a whole. Indeed, the study of such adversaries is essential to the success of crowdsourced communications as a whole, and while this analysis readily

recognizes this possibility, I leave rigorous analysis of this phenomenon up to question.

On a similar note, other issues of reputation and perception remain up to question. As conjectured previously, from Twitter's perspective, it may well be of no relevance whether Community Notes is effective; rather, Twitter need only gain a reputation of active development of content moderation, and Twitter's reputation, and the reputation of social media platforms generally, will be evaluated in turn. Likewise, there is lacking research in the evaluation of Community Notes from the consumers' perspective: is citizens' active participation in content moderation, as construed here or in similar programs, sufficient to assuage free speech and other concerns? In choosing a policy of social media regulation, the social planner must be able to evaluate these as well; spread of disinformation and other political effects are only one of several tradeoffs for political economists to consider in this ever-growing field.

Finally, though I emphasize possible effects on real-world political outcomes in terms of electoral choice, I leave up to question another real-world political effect that has recently garnered much scholarly and media attention: political violence. Indeed, I propose that Community Notes and similar content moderation schemes are likely to affect real-world outcomes because they are particularly effective for agents with weaker priors; however, this interpretation also implies that such content moderation schemes are ultimately ineffective against individuals with the strongest priors, who are most likely to incite political violence on social media (Wahlstrom and Tornberg 2019). Ultimately, this may well be a normative problem: should the priority of content moderation be to ensure voters (particularly the most influential ones) are well-informed in their electoral choices, or should it instead prioritize the mediation of disinformation and non-extremism among the most partisan individuals? In any case, as the link between political misinformation and political violence grows (Piazza 2020), it is clear electoral choice is far from the only real-world political effect of social media and content moderation, and the ability of Community Notes to affect change in this context is far from optimistic.

Conclusion

Altogether, Community Notes is a unique program relevant to policymakers and researchers alike. From the social planner's perspective, the benefits of Community Notes are apparent but deeply nuanced: though it appears effective in reducing engagement with Tweets in general, it appears to have no significant effect on moderating Tweets by politicians, and may well even be altogether less effective for Republicans. Even so, Community Notes may serve further policymaking goals as a preventative measure: as conjectured previously, the program may have some effect in dissuading the spread of disinformation particularly among uninformed agents, and this effect will likely only increase as the Community Notes program continues to expand. Likewise, Community Notes as a whole would be a relatively costless program to implement across platforms, given that Twitter publicly posts its open-source implementation of crowdsourced content moderation. Nevertheless, other tradeoffs are duly mentioned but not yet fully quantified: the potential for mislabeling of content, the exacerbation of selective media exposure, and the actual perception of Community Notes as an intervention that preserves free speech rights remain unexplored.

From a researchers' perspective, Community Notes provides additional evidence in the contentious study of the political economy of social media consumption, yet poses many more questions. The analysis here confirms significant yet nuanced effects of soft content moderation schemes in reducing engagement with disinformation, but a rigorous analysis of real-world political effects, particularly for political violence, is left up to question. Likewise, I confirm the existence of evidence of certain strategic interactions with Community Notes and content moderation interventions more broadly, but to evaluate direct effects, I provide only limited observational evidence. Even further questions on echo chambers and platform/user reputation are also left unanswered.

All told, there is much evidence to be gained from computational research methods to study the political economy of social media consumption. Indeed, the influence of social media on

politics is becoming increasingly apparent, with increasing evidence of social media effects on electoral competition, political polarization, and even dysfunction of political systems as a whole (Haidt and Bail Ongoing). In turn, as policymakers struggle to examine consumer surplus in social media consumption, regulation of social media platforms, and interpretation of free speech rights as a whole, social scientists will undoubtedly continue research on Community Notes and similar social media interventions. As of current, the effects of social media on political discourse are deeply contentious, though continued advancements in both computational social science and microeconomic theory will continue to paint a clearer picture.

Appendix

Appendix 1 (Deriving μ (due Iaryczower (2021)))

WLOG, let there be two agents, 1 and 2, who are not in a cascade, and there are equal numbers of agents prior to agent 1 picking H and L. Further, let $V = 1$. *If* $p > C$, *it is easy to see that agent 1 will always believe H or L according to their private signal* x_1. *In turn, agent 2 will always know agent 1's signal, and acts according to strategy* $\sigma_i = P(i \text{ picks } H)$:

$$\sigma_2^*(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 = x_2 = H \\ \frac{1}{2} & \text{if } x_1 = H, x_2 = L \end{cases}$$

Note that an H cascade begins only if two individuals consecutively believe H. So, $P(H \text{ cascade} \,|\, V = 1) = p^2 + \frac{1}{2}p(1-p) = \mu$.

Appendix 2 (Deriving Equation 1)

We already know that for agent j to pick reject during an H cascade, it must be:

$$P(V = 1 \,|\, x_j = L, y = L, H \text{ cascade}) < C$$

By Bayes' Rule:

$$P(V = 1 \,|\, x_j = L, y = L, H \text{ cascade}) = \frac{P(x_j = L, y = L, H \text{ cascade} \,|\, V = 1) * P(V = 1)}{P(x_j = L, y = L, H \text{ cascade})}$$

Note the events are independent, so:

$$P(x_j = L, y = L, H \text{ cascade} \,|\, V = 1) = P(x_j = L \,|\, V = 1)P(y = L \,|\, V = 1)P(H \text{ cascade} \,|\, V = 1)$$

Using the above, we have:

$$P(V = 1 \,|\, x_j = L, y = L, H \ \textit{cascade}) = \frac{(1-p)(1-q)\mu}{(1-p)(1-q)\mu + pq(1-\mu)}$$

Appendix 3 ($P(V = 1 \,|\, x_j = L, y = L, H$ cascade) is decreasing in q)

Given:

$$P(V = 1 \,|\, x_j = L, y = L, H \ \textit{cascade}) = \frac{(1-p)(1-q)\mu}{(1-p)(1-q)\mu + pq(1-\mu)}$$

Taking derivatives:

$$\frac{\partial}{\partial q} P(V = 1 \,|\, x_j = L, y = L, H \ \textit{cascade}) = -\frac{(p-1)p(\mu-1)\mu}{(p(q-\mu) - q\mu + \mu)^2}$$

So, for $p \in (0,1)$, $q \in (0,1)$, *and* $\mu \in (0,1)$, $\frac{\partial}{\partial q} P(V = 1 \,|\, x_j = L, y = L, H \ \textit{cascade}) < 0$.

	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
Number of Ratings	1.33	0.01	0.02	0.07
Helpful Ratings	1.28	0.01	0.02	0.05
Not Helpful Ratings	1.06	0.00	0.02	0.15
Somewhat Helpful Ratings	1.35	0.01	0.02	0.06
Agree with Note	1.03	0.00	0.00	0.00
Disagree with Note	1.00	0.00	0.00	0.00

Table 10: Additional Measures for Covariate Balance of Matches, All Tweets

	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
Number of Ratings	1.00	0.01	0.11	0.08
Helpful Ratings	1.01	0.01	0.05	0.06
Not Helpful Ratings	0.86	0.01	0.11	0.14
Somewhat Helpful Ratings	1.04	0.01	0.05	0.07
Agree with Note	1.00	0.00	0.00	0.00
Disagree with Note	1.00	0.00	0.00	0.00

Table 11: Additional Measures for Covariate Balance of Matches, Tweets by Politicians

Bibliography

Alizadeh, Meysam, Fabrizio Gilardi, Emma Hoes, K. Jonathan Kluser, Mael Kubli, and Nahema Marchal. 2022. "Content Moderation As a Political Issue: The Twitter Discourse Around Trump's Ban." *Journal of Quantitative Description: Digital Media* 2. https://doi.org/10.51685/jqd.2022.023.

Allcott, Hunt and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31, no. 2 (2017): 211-236.

Allen, Jennifer, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. "Scaling up fact-checking using the wisdom of crowds." *Science Advances* 7, no. 36. https://www.science.org/doi/full/10.1126/sciadv.abf4393.

Allen, Jennifer, Cameron Martel, and David G. Rand. 2022. "Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowd-sourced fact-checking program." *Association for Computing Machinery: CHI Conference on Human Factors in Computing Systems*: 1-19. https://doi.org/10.1145/3491102.3502040.

Arceneaux, Kevin and Martin Johnson. 2013. *Changing Minds or Changing Channels? Partisan News in an Age of Choice*. Chicago and London: The University of Chicago Press.

Aslett, Kevin, Andrew M. Guess, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2022. "News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions." *Science Advances* 8, no. 18. https://doi.org/10.1126/sciadv.abl3844.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya. "Facts, alternative facts, and fact checking in times of post-truth politics." *Journal of Public Economics*

182 (2019). https://doi.org/10.1016/j.jpubeco.2019.104123.

Bikhchandani, Sushil, David Hirschleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Information Cascades." *Journal of Political Economy* 100, no. 5: 992-1026. http://www.jstor.org/stable/2138632.

Clayton, Katherine et al. 2020. "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reudcing Belief in False Storeis on Social Media." *Political Behavior* 42: 1073-1095. https://doi.org/10.1007/s11109-019-09533-0.

Coleman, Keith. 2021. "Introducing Birdwatch, a community-based approach to misinformation." *Twitter Blog*, January 25. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch- a-community-based-approach-to-misinformation.

"Community Notes: Download data." *Twitter*. Accessed December 29, 2022. https://twitter.com/i/birdwatch/download-data.

"Community Notes Guide." *Twitter*. Accessed December 29, 2022. https://twitter.github.io/communitynotes/.

Dejean et al. 2022. "Partisan selective exposure in news consumption," *Information Economics and Policy* 60.

Easley, David and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge: Cambridge University Press.

Freelon, Deen, Alice Marwick, and Daniel Kreiss. 2020. "False equivalencies: Online activism from left to right." *Science* 369, no. 6508: 1197-1201. https://doi.org/10.1126/science.abb2428.

Friggeri, Adrien, Lada A. Adamic, Dean Eckles, and Justin Cheng. 2014. "Rumor Cascades." *Proceedings of the International AAAI Conference on Web and Social Media* 8, no. 1: 101-110.

Fujiwara, Thomas, Karsten Muller, and Carlo Schwarz. 2021. "The Effect of Social Media on Elections: Evidence from the United States." *NBER Working Paper* no. 28849. Accessed January 25, 2023. https://www.nber.org/system/files/working_papers/w28849/w28849.pdf.

Gentzkow, Matthew and Jesse M. Shapiro. 2006. "Media Bias and Reputation." *Journal of Political Economy* 114, no. 2: 280-316. https://www.jstor.org/stable/10.1086/499414.

Greene, Kenneth F. 2011. "Campaign Persuasion and Nascent Partisanship in Mexico's New Democracy." *American Journal of Political Science* 55, no. 2. https://doi.org/10.1111/j.1540-5907.2010.00497.x.

Guess, Andrew, Pablo Barbera, Simon Munzert, and JungHwan Yang. 2021. "The consequences of online partisan media." *Proceedings of the National Academy of Sciences* 14, no. 118. https://www.pnas.org/doi/full/10.1073/pnas.2013464118.

Guess, Andrew. 2022. "Lecture Notes from POL327: Mass Media, Social Media and American Politics." *Princeton University.*

Gupta, Schloak, Sara E. Bolden, Jay Kachhadia, Ania Korsunska, and Jennifer Stromer-Galley. 2019. "PoliBERT: Classifying political social media messages with BERT." *SBP-BRIMS 2020.* https://news.illuminating.ischool.syr.edu/2020/11/24/polibert-classifying-political-social-media-messages-with-bert/.

Haidt, Jonathan and Chris Bail. Ongoing. "Social Media and Political Dysfunction: A Collaborative Review." *New York University.* Accessed January 23, 2023. https://docs.google.com/document/d/1vVAtMCQnz8WVxtSNQev_e1cGmY9rnY96ecYuAj6C548/edit.

Halberstam, Yosh and Brian Knight. 2016. "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter." *Journal of Public Economics* 143 (2016): 73-88.

"Helpful Birdwatch notes are now visible to everyone on Twitter in the US." 2022. *Twitter Blog*, October 6. https://blog.twitter.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us.

Iaryczower, Matias. 2022. "Lecture Notes from POL347: Game Theory in Politics." *Princeton University*.

Kemp, David and Emily Ekins. 2021. "Poll: 75% Don't Trust Social Media to Make Fair Content Moderation Decisions, 60% Want More Control over Posts They See." *Cato Institute*. https://www.cato.org/survey-reports/poll-75-dont-trust-social-media-make-fair-content-moderation-decisions-60-want-more.

Kim, John. 2023. "Birdwatch Data Collection." *GitHub*. https://github.com/benidjones/birdwatch-data-collect.

Lasswell, Harold D. 1971. *Propaganda Technique in World War I*. Cambridge: MIT Press.

Lazarsfeld, Paul F., Bernard Berelson, and Hazel Gaudet. 1988. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Columbia University Press.

Lechner, Michael. 2011. "The Estimation of Causal Effects by Difference-in-Difference Methods." *Foundations and Trends in Econometrics* 4, no. 3: 165-224. http://dx.doi.org/10.1561/0800000014.

Liu, Yi, Pindar Yuldirim, and Z. John Zhang. 2022. "Implications of Revenue Models and Technology for Content Moderation Strategies." *Marketing Science* 41, no. 4: 831-847. https://doi.org/10.1287/mksc.2022.1361.

Malzahn, Janet and Andrew B. Hall. 2023. "Election-Denying Republican Candidates Underperformed in the 2022 Midterms." *Stanford Graduate School of Business* Working Paper no. 4076. Accessed March 11, 2023. https://www.gsb.stanford.edu/faculty-research/

working-papers/election-denying-republican-candidates-underperformed-2022-midterms.

Mitchell, Amy, Jeffrey Gottfried, Galen Stocking, Mason Walker, and Sophia Fedeli. 2019. "Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed." *Pew Research Center*. https://www.pewresearch.org/journalism/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.

Pew Research Center. 2014. "Political Polarization & Media Habits: From Fox News to Facebook, How Liberals and Conservatives Keep Up with Politics." *Pew Research Center*. https://www.pewresearch.org/journalism/2014/10/21/section-1-media-sources-distinct-favorites-emerge-on-the-left-and-right/.

Piazza, James A. 2022. "Fake news: the effects of social media disinformation on domestic terrorism." *Dynamics of Asymmetric Conflict* 15, no. 1: 55-77. https://doi.org/10.1080/17467586.2021.1895263.

"Politicians tracked by Politwoops." 2019. *ProPublica Data Store*. https://www.propublica.org/datastore/dataset/politicians-tracked-by-politwoops.

Russell, Annelise. 2018. "US Senators on Twitter: Asymmetric Party Rhetoric in 140 Characters." *American Politics Research* 46, no. 4: 695-723.

Shin, Jieun and Kjerstin Thorson. 2017. "Partisan Selective Sharing: The Biased Diffiusion of Fact-Checking Messages on Social Media." *Journal of Communication* 67, no. 2: 233-255. https://doi.org/10.1111/jcom.12284.

Tucker, Joshua A. et al. 2018. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." *Hewlett Foundation*. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3144139.

Twitter. "Twitter API Documentation." *Development Platform*. Accessed December 29, 2022. https://developer.twitter.com/en/docs/twitter-api.

Twitter. 2021. "Permanent suspension of @realDonaldTrump," *Twitter Blog*, January 8. https://blog.twitter.com/en_us/topics/company/2020/suspension.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science* 9, no. 6380: 1146-1151. https://doi.org/10.1126/science.aap9559.

Wahlstrom, Mattias and Anton Tornberg. 2019. "Social Media Mechanisms for Right-Wing Political Violence in the 21st Century: Discursive Opportunitites, Group Dynamics, and Co-Ordination." *Terrorism and Political Violence* 33, no. 4: 766-787. https://doi.org/10.1080/09546553.2019.1586676.

Wojcik, Stefan and Adam Hughes. 2019. "Sizing Up Twitter Users." *Pew Research Center*. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

Wojcik, Stefan et al. 2022. "Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation." *Twitter*. https://github.com/twitter/communitynotes/blob/main/birdwatch_paper_2022_10_27.pdf.

Zannettou, Savvas. 2021. " 'I Won the Election!': An Empirical Analysis of Soft Moderation Interventions on Twitter." *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media* 15, no. 1: 865-876. https://doi.org/10.1609/icwsm.v15i1.18110.

This paper represents my own work in accordance with University regulations.