Community Notes Moderate Engagement With and Diffusion of False Information Online

Isaac Slaughter[a], Axel Peytavin[b], Johan Ugander[b], and Martin Saveski[a,1]

[a]Information School, University of Washington
[b]Department of Management Science and Engineering, Stanford University
[1]Corresponding author: msaveski@uw.edu

Abstract

Social networks scaffold the diffusion of information on social media. Much attention has been given to the spread of true vs. false content on online social platforms, including the structural differences between their diffusion patterns. However, much less is known about how platform interventions on false content alter the engagement with and diffusion of such content. In this work, we estimate the causal effects of Community Notes, a novel fact-checking feature adopted by X (formerly Twitter) to solicit and vet crowd-sourced fact-checking notes for false content. We gather detailed time series data for 40,074 posts for which notes have been proposed and use synthetic control methods to estimate a range of counterfactual outcomes. We find that attaching fact-checking notes significantly reduces the engagement with and diffusion of false content. We estimate that, on average, the notes resulted in reductions of 45.7% in reposts, 43.5% in likes, 22.9% in replies, and 14.0% in views after being attached. Over the posts' entire lifespans, these reductions amount to 11.4% fewer reposts, 13.0% fewer likes, 7.3% fewer replies, and 5.7% fewer views on average. In reducing reposts, we observe that diffusion cascades for fact-checked content are less deep, but not less broad, than synthetic control estimates for non-fact-checked content with similar reach. This structural difference contrasts notably with differences between false vs. true content diffusion itself, where false information diffuses farther, but with structural patterns that are otherwise indistinguishable from those of true information, conditional on reach.

1 Introduction

The spread of false information on social media poses risks to public health [1], democratic processes [2], and social cohesion [3]. Social media has been broadly observed to preferentially support the spread of false news over true news [4, 5, 6, 7]. Scholars as well as social media platforms are actively working to design and test strategies to limit its transmission [8, 9, 10], including fact-check warning labels placed on individual sources or pieces of information [11, 12], educational interventions to boost users' competencies at identifying false information [13, 14, 15, 16], and a shift to design objectives other than user engagement [17, 18, 19].

Professional fact-checking is the most widely used intervention against misinformation, often implemented by attaching warning labels to fact-checked posts [20, 21, 22]. Studies investigating the effectiveness of these labels find that they decrease self-reported belief in and willingness to share misinformation [23, 24, 25]. However, even if effective, professional fact-checking is costly and difficult to scale both in speed and coverage [26], and increasingly viewed with skepticism by segments of the public [27]. Crowd-sourced fact-checking has emerged as a promising alternative, leveraging the "wisdom of the crowd," i.e., that aggregating judgments of groups of non-experts leads to accurate assessments even if the individual assessments are inaccurate [28]. Lab experiments investigating the feasibility of crowd-sourced fact-checking find that groups as small as 15 people can identify misinformation as accurately as professional fact-checkers [29, 30, 31, 32].

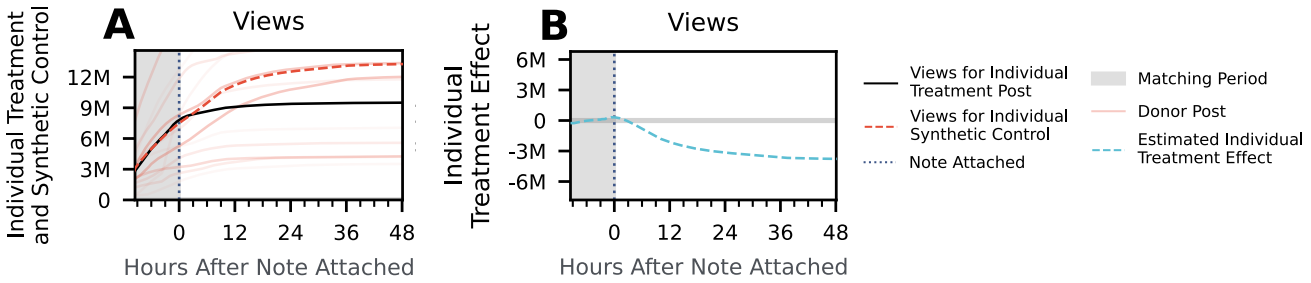

Figure 1: Illustration of synthetic control methodology for views. *(A)* Total number of views over time before and after receiving a community note for a sample post. The black line represents the post's observed view count and the dashed red line depicts the synthetic control's view count, an estimate of what would have occurred had no note been attached. Individual donor posts that contribute to the synthetic control are shown in solid red lines with intensity of color proportional to their weight. Note that in addition to view counts the synthetic control has a similar trajectory on all other engagement and diffusion metrics. *(B)* Estimated difference between views of the synthetic control without the note and observed views with the note, quantifying the decline in views attributable to note attachment at a given time.

Building on these findings, X (formerly Twitter) introduced a crowd-sourced fact-checking system called Community Notes [33]. The system enables ordinary users to propose fact-checking notes to be attached to potentially misleading posts and rate the helpfulness of proposed notes. The system uses a "bridging-based" matrix factorization algorithm to score the overall helpfulness of notes based on the individual ratings [34]. Notes rated helpful by many users with diverse views, as measured by estimated latent positions, are scored higher. Only notes that cross a certain helpfulness threshold are deemed helpful and displayed with the post.

Upon introducing the Community Notes program, X reported results from an A/B test that notes selected by the bridging-based algorithm reduced individual-level decisions to like and repost misinformation by 25–34% relative to a control group [34]. Through the lens of widely employed epidemiological models of information diffusion, changes in the probability that individual units will share content typically have a highly non-linear relationship with the overall number of people exposed to the content [35], a quantity that is not easily assessed through an A/B test due to ubiquitous network effects, i.e., interference between treatment and control units [36].

In this work, we investigate the causal effects of attaching community notes to posts on the engagement with and diffusion of the posts. We collect time-series data for 40,074 posts created between March and June 2023 for which community notes were proposed. We track key engagement metrics of the posts, including the number of reposts, likes, replies, and views over time. We also collect all reposts since the post was created, both before and after a note was proposed, and the follow graphs of the users that reposted them, which we use to reconstruct the diffusion cascades of the posts. These granular records concerning a post's engagement both before and after a community note appears allow us to provide precise estimates of the notes' effects, and investigate the conditions under which community notes are more or less successful at reducing the impact of misinformation.

We use synthetic control methods to estimate these causal effects [37, 38]. For each post with a note attached, we construct a synthetic control by averaging the engagement histories of multiple donor posts—those for whom a note was proposed, but not attached—such that the synthetic control closely matches the history of all metrics of the noted post during the period before note attachment. Then, we estimate the effect of attaching a note by comparing engagement metrics during the period after the note was attached between the post and its synthetic control. This analysis produces an individual treatment effect for each post where a community note was attached. Figure 1 illustrates the procedure for a sample post: Figure 1*A* shows the number of views over time for the community noted post along with its estimated views had the note not been attached, while Figure 1*B* shows the estimated treatment effect of the community note over time. To validate our estimation approach, we conduct an in-time placebo test [39] by artificially shifting the note attachment time one hour earlier and, as expected,

observe null effects in the period between the artificial and actual attachment time (further details can be found in Appendix D.)

We find that notes significantly reduce the number of views, reposts, likes, and replies. We interpret the impact on these metrics through two perspectives: a *growth perspective*, which quantifies the reduction in the additional growth of the metric after a note was attached, and an *overall perspective*, which quantifies the total reduction in the metric since the post's creation. The growth perspective measures effectiveness conditional on when the note was attached, while the overall perspective also accounts for the engagement that occurred before the note was attached. These varied perspectives are related to measures of the "prevented fraction" and "prevented fraction among the unexposed" in epidemiology [40].

Beyond engagement, we also consider impact of note attachment on the structure of information diffusion. Previous studies have found that fact-checked false news have larger, deeper, wider, and more viral diffusion cascades than fact-checked true news [5]. Subsequent analyses of the same data have shown that while there are significant differences in cascade size, the structural differences disappear after controlling for size, suggesting that mechanisms through which true and false fact-checked news diffuse are relatively similar [6]. In our setting, we find that note attachment qualitatively changes the structure of the post's diffusion cascade, relative to the same post without a note. Most significantly, it reduces the depth and structural virality more than would be expected given the overall reduction in size.

Our rich data on diverse engagement metrics (reposts, replies, likes, and views) as well as our reconstruction of diffusion cascades (enabling us to study how note attachment influences cascade structure) go far beyond earlier work studying the effects of Community Notes [41], which considered only the effects on reposts and deletions. That prior work also used difference-in-differences methods which, unlike our synthetic control methods, rely on strong "parallel trends" assumptions [42]. When comparable, our independent estimates also provide important corroboration of those prior estimated effects.

Since our synthetic control methods approach provides causal effect estimates at the individual post level, we can examine how average effects vary across different post subpopulations. Overall, we find that notes have the greatest absolute impact on reducing engagement when they are attached shortly after a post is created or attached to highly engaging posts. We find that notes on posts with embedded media, as opposed to text-only posts, are associated with larger reductions. We also see larger reductions on posts where concerns about altered media are presented as a reason for the note. In terms of differences across how notes are composed, we find that moderately long notes and notes written using simpler language are associated with larger reductions in engagement.

2 Data Collection

We collected data from March 16 to June 23, 2023, tracking $40,074$ posts for which a note was proposed. We continuously monitored the "New" tab of the Community Notes website, which provides the identifiers of the posts for which notes were recently proposed. When a community note was created, we immediately retrieved the associated post's engagement metrics using the X API. We then made API calls every five minutes to record the post's total number of reposts, likes, replies, and views for three weeks following note creation. To ensure that all post engagement histories are comparable when constructing synthetic control weights, we shifted the engagement metrics to a timeline aligned with post creation rather than note creation time by linearly interpolating to fifteen-minute intervals. All exposure and engagement measurements are all-cause measures, regardless of whether the content was delivered through an algorithmic or reverse-chronological feed.

We use public data available from X to determine when notes were attached to posts and which notes were never attached to any posts. Among the $40,074$ posts included for analysis, $6,753$ (16.9%) received helpful notes and constituted the treatment group. (Further details on the construction of the treatment group are provided in Appendix E.2.) The remaining $33,321$ posts, for which a note was proposed but no note reached a helpful status,

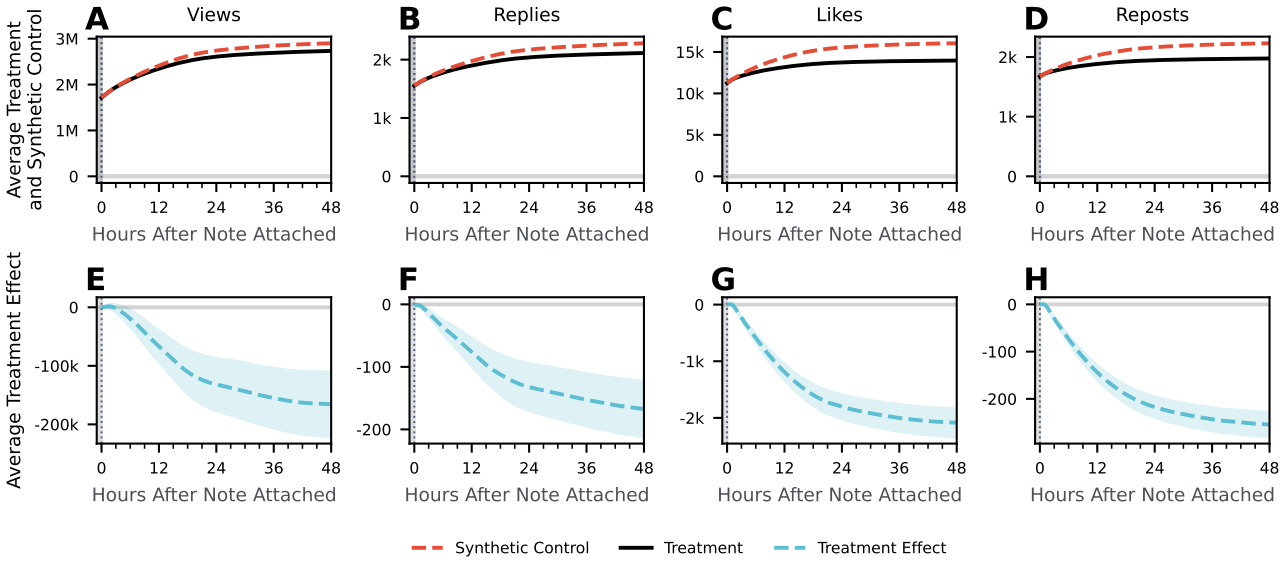

Figure 2: Effects of note attachment on views, replies, likes, and reposts. *(A-D)* Average treatment and synthetic control for all engagement metrics. Average treatment consists of the average value for a metric at a given time point among posts that received notes. Average synthetic control consists of the average value of the synthetic controls, estimating what the average would have been had notes not been attached. *(E-H)* Average treatment effect on the treated for all engagement metrics. Estimates the average difference between the treatment and control: the average decline in a metric due to note attachment.

and thus no intervention took place, constitute the donor pool for constructing our synthetic controls. We focus our analysis on the effects of community notes within a 48-hour window after a note is attached. Given the rapid decay of engagement on X [43], our 48-hour estimates closely approximate the lifetime effects for posts that remain on the platform.

In addition to collecting engagement metrics starting when a first community note was written about a post, we also collected each post's full public repost and reply history, extending back to its creation. However, we were unable to collect this data for deleted and private posts. In such cases, we relied on the repost and reply counts returned by the X API. We collected complete repost histories for 36,580 posts (90.8%) and complete reply histories for 30,879 posts (76.6%). We provide further details on the data cleaning and cascade data collection in Appendix E.

3 Results

3.1 Decline in Average Engagement

We estimate a synthetic control and individual treatment effect on the number of views, replies, likes, and reposts for each post that received a helpful note. We then aggregate these effects into average treatment effects on the treated population. Details on the construction of synthetic controls and the uncertainty quantification underlying the confidence intervals can be found in Section 5. Figure 2*A* shows the average number of views over the 48 hours after a note was attached for posts that received community notes, along with the average number of views for the same posts' synthetic controls. Figure 2*E* shows the average treatment effect across all noted posts. The equivalent figures for the number of replies, likes, and reposts are shown in Figure 2 *B-D* and Figure 2 *F-H*, respectively. These measures quantify the aggregate impact of the Community Notes program. We discuss the heterogeneities among posts in Section 3.3 and the distribution of individual treatment effects in Appendix C.

We first discuss the aggregate effect of a note on posts' number of views. We estimate that the average number of views for posts that received helpful notes decreased from 2.90 million (95% CI: [2.84M, 2.96M]) to 2.73 million views due to note attachment. These estimates represent the total number of views that noted posts received, or would have received without notes, 48 hours after they had community notes attached. The treatment effect at this time amounts to $-165,670$ views (95% CI: [$-221,446$, $-109,895$]). This corresponds to a -14.0% reduction in additional growth of number of views after note attachment or a -5.7% reduction in the total number of views, including those that occurred before the note was attached. Note that the first metric captures how the presence of a note affects a post's ability to gain new views, while the second measures the overall impact of community notes at the platform level.

While the decrease in views reflects how community notes limit the posts' reach, change in engagement metrics that require active participation from the users—specifically replies, likes, and reposts—capture how the notes affect the way that users interact with the posts. Replies on social media may signal some combination of agreement, disagreement, or confusion from a replier. We estimate that the average number of replies to posts receiving helpful notes decreased by a similar percentage as the number of views: from $2,280$ (95% CI: [$2,236$, $2,325$]) to $2,113$, a change of -167 replies (95% CI: [-212, -123]), which amounts to a -22.9% change in reply growth after attachment or -7.3% change in total number of replies.

Likes and reposts, on the other hand, are more frequently used as signals of positive engagement: likes on social media can indicate that the user finds the post enjoyable, useful, or interesting [44, 45], while reposts can signal agreement or serve to amplify a message [46]. We estimate that the average number of likes given to posts that received helpful notes during this period fell from $16,045$ (95% CI: [$15,794$, $16,297$]) to $13,961$ due to note attachment, an absolute change of $-2,084$ likes (95% CI: [$-2,336$, $-1,833$]). This amounts to a change of -43.5% in likes after note attachment and a change of -13.0% in total likes. Similarly, relative to the average synthetic control of $2,230$ (95% CI: [$2,204$, $2,256$]), we estimate that note attachment led to a change of -254 (95% CI: [-281, -228]) reposts, bringing the observed average down to $1,975$. This amounts to a percentage change in reposts after attachment of -45.7% and a percentage change in total reposts of -11.4%.

3.2 Altered Dynamics of Information Diffusion

Having found that community notes lead to sizable reductions in average engagement, we next examine their impact on how information spreads on the platform, specifically their effect on the structure of repost cascades. A repost cascade records the tree of reposts stemming from a post. The max depth of a cascade refers to the length of the longest chain of reposts it contains, max breadth refers to the maximum number of reposts at any level of depth, and structural virality refers to the average distance between any two nodes in the cascade, standardized by its size [47, 48]. A cascade with high breadth or low structural virality suggests that the post spread primarily through direct reposts of the original post. In contrast, high depth or high structural virality indicates the post spread more through multi-step, person-to-person reposting chains, a pattern often seen with rumors and viral content [48]. Additional details on the calculation of these metrics can be found in Appendix E.3.

As before, we construct a synthetic control for each post that received a helpful note, but now estimate treatment effects for structural metrics that characterize the post's repost cascade: its max breadth, max depth, and structural virality. We again take the mean of these synthetic controls, as shown in Figure 3 *B-D*, along with the average observed values under treatment, and calculate average treatment effects, shown in Figure 3 *F-H*. The effect on the total number of reposts, also referred to as cascade size, is repeated in Figure 3*A* and Figure 3*E* for reference.

As with engagement metrics, we estimate that note attachment leads to a decline in all structural metrics of the repost cascade after 48 hours, relative to what would have been expected had the note not been attached. The average max breadth of synthetic controls is $1,570$ reposts (95% CI: [$1,551$, $1,589$]), while the observed value under treatment was $1,388$ reposts, a change of -182 reposts (95% CI: [-200, -163]) or -45.7% in

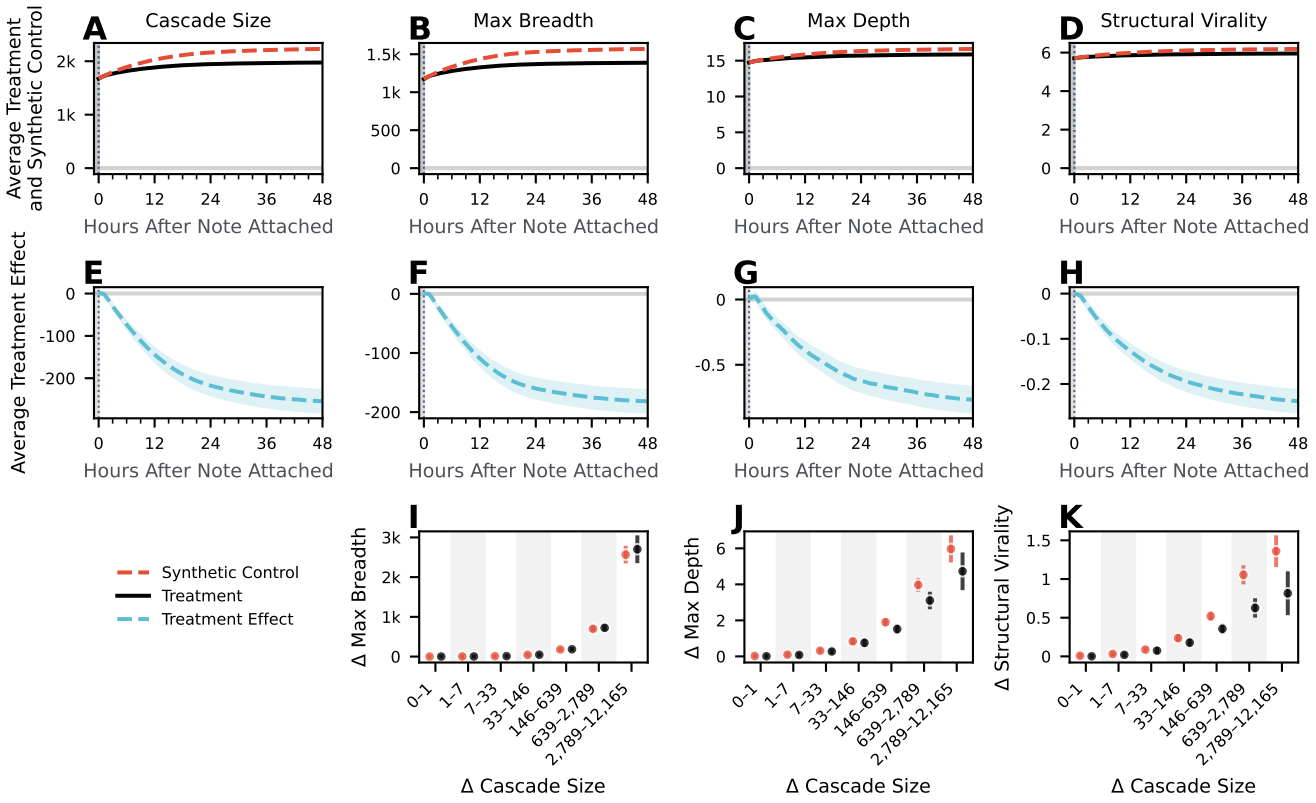

Figure 3: Effects of note attachment on the structure of repost cascades. *(A-D)* Average treatment effect for *(A)* the size of the repost cascade (equivalent to the total number of reposts), *(B)* the maximum breadth of the repost cascade (which frequently occurs at the first level of the cascade, and is hence a proxy for the number of direct reposts), *(C)* the maximum depth of the repost cascade (the longest chain of person-to-person reposting), and *(D)* the structural virality of the repost cascade (intended to measure the extent to which content was spread "virally," where larger numbers indicate more viral diffusion). *(E-H)* Average treatment effect on the treated for structural metrics. *(I-K)* Growth after note attachment in cascade size relative to growth in cascade structure. For individual treatment and control posts whose repost cascade grew by a given amount in the 48 hours between $t = 0$ and $t = 48$, the plot shows the average growth in a structural metric that occurred in that same time period. Bin edges evenly divide the positive range of Δ Cascade Size in logarithmic scale.

growth after note attachment, or -11.6% in total. The estimated average max depth under control is 16.64 reposts (95% CI: [16.54, 16.73]), compared to the observed average of 15.87 reposts, a treatment effect of -0.77 reposts (95% CI: [-0.86, -0.67]), or -40.5% growth after note attachment, or -4.6% change in total. Finally, the average structural virality under control is estimated to be 6.19 (95% CI: [6.17, 6.21]) while the observed value is 5.95, a change of -0.24 absolute units (95% CI: [-0.26, -0.21]), or -48.8% growth after note attachment, or -3.8% in total.

Smaller repost cascades tend to be both less broad and less deep than larger repost cascades [6]. To disentangle changes in size from changes in structural metrics, we perform the following matching procedure. For each post that received a note, we first calculate the growth in observed cascade size as well as in the observed structural metrics in the 48 hours after note attachment. We do the same for the estimated synthetic control values. Finally, we compare growth in structural metrics for treated and control posts whose repost cascades grew by the same amount in total size after note attachment. As shown in Figure 3 *J-L*, we find that while max breadth does not differ in the distribution between treatment and synthetic controls that grow by similar amounts, max depth and structural virality do differ. Relative to synthetic controls that grew by a similar amount, posts that receive notes do so with a smaller max depth and smaller structural virality, indicating less viral diffusion in the presence of an

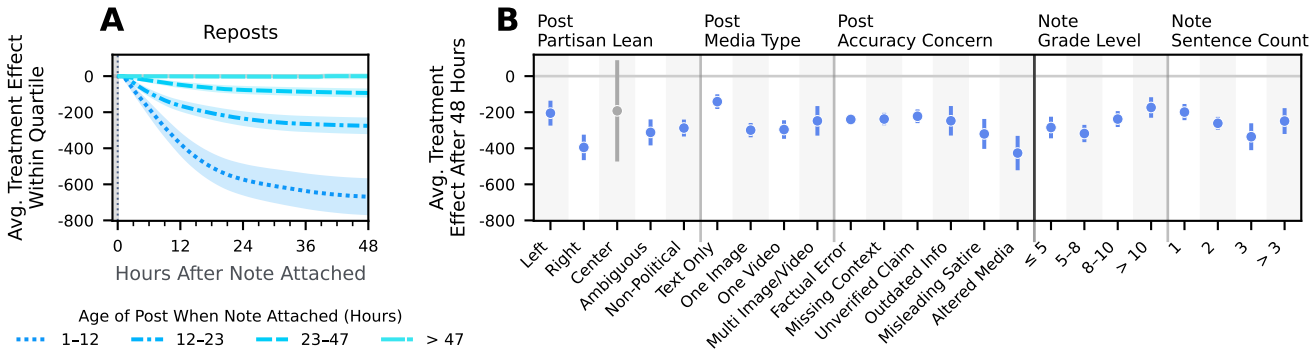

Figure 4: Factors associated with effects on reposts. *(A)* Average treatment effects for posts, stratified by quartiles of note attachment speed, i.e., the amount of time that elapsed between the post's creation and when it first had a note attached. *(B)* Average treatment effects after 48 hours, based on (i) the partisanship of the post, (ii) the number of images and videos included, (iii) the accuracy concern raised by the community note writer, (iv) readability of the community note, as measured by Flesch–Kincaid grade level, and (v) length of the note.

attached note. These results indicate that effects of attaching a community note on depth and structural virality cannot be simply explained by a change in cascade size and suggest that the attachment of a note significantly affects the mechanism through which the posts spread over the network. These structural results are consistent with community notes having a larger moderating effect on users when the post reaches them through a repost cascade, and less of an effect on users who receive the post directly from the original poster.

3.3 Factors Associated with Large Effects

Our previous analyses show that community notes significantly reduce the average engagement with misleading posts and change the diffusion patterns of such posts. However, their effects are not uniform across all types of posts or notes. Next, we perform exploratory analyses to identify the contexts in which attaching community notes has the largest impact.

In Figure 4*A*, we show the average treatment effect of community notes on reposts, stratified by speed of note attachment. We find that notes attached soon after a post is created are more effective at reducing reposts. Specifically, the absolute treatment effects for the first quartile (attached within 1–12 hours), second quartile (12–23 hours), third quartile (23–47 hours), and fourth quartile (47+ hours) are -669 reposts (95% CI: [-764, -574]), -276 reposts (95% CI: [-314, -237]), -93.5 reposts (95% CI: [-111, -75.8]), and -0.47 reposts (95% CI: [-10.6, 9.68]), respectively. The corresponding percentage reductions in repost growth following note attachment are -49.4%, -43.3%, -38.4%, and -1.53%, while the total repost reductions amount to -24.7%, -11.9%, -4.22%, and -0.03%, respectively.

As shown in Appendix B, other engagement metrics exhibit similar monotonically decreasing effects when stratified by speed of note attachment. One nuance is that for posts in the fourth quartile (47+ hours from post creation to note attachment), the percentage changes in average growth of views and replies are *positive*: an increase of 11.5% for views and an increase of 20.7% for replies 48 hours after note attachment. This increase suggests that community notes may draw attention to stale posts. However, it does not imply that the notes draw more endorsement or agreement, as the average growth for likes and reposts (as opposed to views) within this quartile are -9.89% and -1.53%, respectively.

Another major factor associated with the magnitude of the treatment effect is the volume of engagement a post received before a note was attached: posts that received more reposts prior to note attachment have larger treatment effects on average. In Appendix A, we show that while the most popular posts exhibit larger absolute drops in reposts after note attachment, their percentage changes are relatively similar, suggesting these declines

stem mostly from their larger baseline audience. In contrast, posts with low initial repost counts often experience small positive changes, indicating that note attaching can sometimes boost the visibility of less prominent content.

While the majority of variation in treatment effects is attributable to the speed of note attachment and the posts' popularity before treatment, we also find that both post content and note quality are associated with variability in treatment effects. Prior work has found that right-leaning media consumers tend to be more prone to motivated reasoning than left-leaning consumers [49] and that Democrats are more likely than Republicans to rate community notes as helpful [34]. These findings suggest that posts on right-leaning content, likely seen more often by right-leaning users [50], would have smaller treatment effects than those on left-leaning content. However, we do not find evidence that this is the case among the English-language posts that we label for partisanship. In fact, we estimate that notes on left-leaning content tend to be less effective at reducing reposts, both in absolute and relative terms. We describe the methodology for labeling post partisanship in Appendix E.4. The average treatment effect on right-leaning posts is -395 reposts (95% CI: [-466, -325]), a percentage change in average growth due to note attachment of -55.0%, relative to an average treatment effect on left-leaning posts of -205 reposts (95% CI: [-274, -136]), a percentage change of -42.6%. As shown in Appendix B, we also find similar effects for views, replies, and likes.

Beyond partisanship, we also find that the type of media included in the post is associated with note effectiveness. The estimated treatment effect for media posts—those containing a single image, single video, or multiple images/videos—are -299 (95% CI: [-339, -260]), -296 (95% CI: [-348, -245]), and -248 reposts (95% CI: [-330, -166]), respectively. In contrast, text-only posts have an estimated treatment effect of -141 reposts (95% CI: [-182, -101]). While media posts generally receive more engagement than text-only posts before receiving a note, we find that the difference in average effect persists even after normalizing by size. The estimated percent changes in average growth due to note attachment are -50.4%, -44.3%, and -49.5% for a single image, single video, and multi-image/video posts, respectively, compared to -38.4% for text-only posts.

The sizes of the treatment effects also vary depending on what type of concern the community note addresses. When proposing a community note, writers are asked to indicate their accuracy concern, i.e., what aspect of the post they consider misleading. They can select multiple concerns, e.g., that the post both contains outdated information and makes a factual error. As shown in Figure 4*B*, we find that concerns related to altered media and misleading satire are most strongly associated with large effects.

Considering the community notes themselves, we find that both readability and length are associated with effectiveness. We measure readability using the Flesch-Kincaid grade level score, an estimate of the minimum U.S. grade level required to comprehend a text [51]. We find that simpler notes tend to be more effective. Notes with grade levels less than or equal to five have an average treatment effect of -285 reposts (95% CI: [-345, -225]), corresponding to a -46.5% change in average growth after note attachment. In contrast, notes with a grade level above ten have an average treatment effect of -174 reposts (95% CI: [-231, -117]), which amounts to a change of -38.3%. Finally, we find that moderately long notes (two or three sentences) are more effective than both short (one sentence) and long (more than three sentences) notes at reducing reposts.

4 Discussion

As the problem of misinformation persists on social media, scalable interventions are necessary to prevent its spread, impact, and harm. Crowd-sourced fact-checking presents one such approach, which has demonstrated encouraging results in early tests and is now deployed as a core content management component on one of the world's largest social media platforms. The public nature of Community Notes' deployment on X, with freely available source code and rating data, as well as detailed information on the content and engagement of posts over time, has allowed us to estimate the impact of community-driven fact-checking through a completely independent audit.

Our results indicate that once community notes are attached, on average, they reduce the engagement with and diffusion of false information on X. Consistent with related work by Chuai *et al.* [41], which studied reposts and deletions, we find that community notes lead to a decline in the number of reposts that a post receives after attachment. Despite differences in causal identification strategies, our estimate of percentage decline in average reposts after note attachment, -45.7% during the period March 16 to June 23, 2023, is roughly comparable with estimates from Chuai *et al.* of -55.2%, -49.6%, -45.6%, and -47.5% in the months of March, April, May, and June of 2023. Our analysis of other outcomes, including structural metrics, finds that notes lead to a similar percentage reduction in likes (-43.5%) but smaller reductions in views (-14.0%) and replies (-22.9%).

These findings suggest that the impact of note attachment is strongest on public expressions of support for content (reposts and likes), while its effect is smaller on whether content reaches people in the first place (views) or whether they choose to engage in an online conversation about it (replies). While these differences might be taken to imply that the reduced support nullifies the effects of misinformation exposure after attachment, we caution against this interpretation. Viewing false information, even if the viewer initially doubts its validity, can increase their likelihood of agreeing with it later [52]. Thus, each view prevented by a community note is meaningful. We also note that the decline in reposts and likes may not necessarily reflect a decline in actual support of the content but rather a reduced willingness among users to signal their support publicly. An internal analysis by X does report that users are less likely to agree with the substance of potentially misleading posts when presented with a community note [34].

The decline in engagement is paired with notable structural changes in how posts diffuse across the platform. The changes are consistent with larger behavioral changes by users who do not follow the original post's author. Such larger changes can be attributed to homophily [53], where users closer to the root author may share similar beliefs or information evaluation approaches with the root author. As an alternative mechanism, these results are also consistent with dyadic social pressures [46] whereby users may feel greater loyalty obligations to close connections than those encountered through deeper network paths.

While we find that community notes effectively reduce engagement once attached (-14.0% views, -22.9% replies, -43.5% likes, and -45.7% reposts), we also find evidence that the system would be much more effective if notes were attached faster. Moreover, the reductions in views, replies, likes, and reposts are much more modest when measured as a percentage of overall engagement with the post, compared to only considering changes in engagement after note attachment. When measured this way, the overall percentage changes in views, replies, likes, and reposts due to note attachment are -5.7%, -7.3%, -13.0%, and -11.4%, respectively.

Our study has several limitations worth considering. First, our analysis can only estimate the effects of community notes on posts that had notes attached. Based on the data available to us, we cannot estimate the coverage of the Community Notes program, i.e., how many misleading posts on X received a note, or what the effectiveness of notes might be on that broader population of posts without notes. Such analysis is inherently challenging as it requires both access to all posts on the platform and a scalable method for identifying misleading posts. Second, we cannot test for implied truth effects [54], i.e., the potentially increased tendency of posts without community notes to be perceive as accurate or non-misleading, even when they are not. Finally, Community Notes is an evolving system and our analysis reflects the effects of the system during the study period, March–June 2023. Like all social media research, the rapidly changing environment makes temporal validity challenging [55]. Since we concluded our data collection, more volunteers have joined the program, extensions of the system have been proposed [56], and substantial updates to the system's implementation have been introduced. These updates include improvements in the time required to run the algorithm and display the notes [57], as well as the automatic attachment of a community note to posts with images and links that were previously included in other noted posts. Our findings suggest that these changes are likely to lead to significant additional reductions in engagement with misleading content on X. Nevertheless, community-based fact-checking is best viewed as one of several interventions [9] worth considering when aiming to reduce the spread of misinformation on social media.

5 Methods

5.1 Estimating Individual Treatment Effects

We use synthetic control methods (SCM) to estimate the effect of receiving a community note on a post's engagement. SCM are commonly employed to estimate the causal effects of interventions when detailed time series data is available for both units that received the intervention and those that did not [37, 38]. Under the SCM framework, each unit that receives the intervention (in our case note attachment, which we also refer to as "treatment") receives an individual synthetic control estimate, which is interpreted as what would have happened to the treated unit had the intervention not occurred (had it never received a community note). Observations of outcome for treatment unit i are denoted as $Y_{imt}(Z_i = 1)$, where m refers to the metric in question (e.g., reposts), t refers to the amount of time elapsed since note attachment (e.g., 48 hours after attachment), and the value of $Z_i = 1$ indicates that we are referring to unit i when it received treatment. We refer to estimates from the synthetic control as $\hat{Y}_{imt}(Z_i = 0)$.

Synthetic controls are weighted averages of donor posts—those where at least one community note was proposed, but which never had a note found to be helpful, and therefore never experienced note attachment. Weights are selected to minimize the total Euclidean distance between the metrics of the treated post and its synthetic control, up until the time at which the treatment unit had a note attached: $\frac{1}{|(m,t)\in M_i \times T_i : t<0|} \sum_{(m,t)\in M_i \times T_i : t<0} [Y_{imt}(1) - \hat{Y}_{imt}(0)]^2$, where M_i refers to the complete set of metrics for unit i, and T_i refers to the complete set of time points. We construct weights by minimizing this distance across the following metrics, when they are available for a post (see Appendix E.5): likes, views, reposts, replies, author follower count, repost cascade maximum depth, repost cascade maximum breadth, and repost cascade structural virality. In sum, fitting weights required solving $6,753$ linearly constrained least squares problems (one for each treated post), each such problem being a quadratic program. Due to the computational burden of solving each of these programs, we restrict the donor pool for each treated post to the $1,000$ control posts closest to that treated unit in Euclidean distance. This restriction greatly reduces the computational cost of solving each quadratic program, and can be viewed as a hard thresholding analog of the "penalized" synthetic control method [38]. To prevent variables with larger scales (e.g., views) from dominating the matching process, we standardize all variables by their sample standard deviation within the treated posts, following Abadie & L'Hour [38].

We refer to the true absolute individual treatment effect for unit i, metric m, at time point t as $\tau_{imt} = Y_{imt}(1) - Y_{imt}(0)$, and $\hat{\tau}_{imt}$ as our estimate of that quantity using synthetic controls. One approach to estimating this effect would be to consider the simple difference between the treated post's engagement and the corresponding engagement for its synthetic control, $Y_{imt}(1) - \hat{Y}_{imt}(0)$. This estimation method can induce bias in $\hat{\tau}_{imt}$ if the treatment unit and synthetic control do not closely match prior to treatment, which is not always possible in high-dimensional datasets [58]. We therefore employ the bias correction procedure recommended by Abadie & L'Hour [38] to address imperfect matches between treatment posts and their synthetic controls when estimating individual treatment effects. The bias correction procedure involves training a regression model (we use an ordinary least squares model) to predict donor posts' post-treatment outcomes based on their pre-treatment history. The value $\hat{\tau}_{imt}$ is then calculated as the difference between the treated post and its synthetic control in residuals from this model.

5.2 Aggregating Individual Treatment Effects

We calculate several statistics to summarize the estimated individual effects. First, we calculate the average estimated treatment effect as $\hat{\tau}_{\cdot tm} = \frac{1}{|N_{\cdot tm}|} \sum_{i \in N_{\cdot tm}} \hat{\tau}_{imt}$ (shown in Figure 2 *E-H*), where $N_{\cdot tm}$ refers to the complete set of treated units for a metric and time point. Referring to the average observed value under treatment as $Y_{\cdot mt}(1)$, and the average estimated synthetic control value as $\hat{Y}_{\cdot mt}(0)$, then $\hat{\tau}_{\cdot tm}$ is equivalent to $Y_{\cdot mt}(1) - \hat{Y}_{\cdot mt}(0)$. This means

that the average estimated treatment effect can be interpreted as the absolute change in average outcome due to note attachment.

To quantify the uncertainty of the synthetic control estimation procedure, we use standard Gaussian confidence intervals, given by $CI_{95\%} = \hat{\tau}_{\cdot mt} \pm z_{\alpha/2} \frac{\hat{\sigma}_{\cdot mt}}{|N_{\cdot tm}|}$, where $\hat{\sigma}_{\cdot mt}$ refers to the estimated standard deviation of the treatment effect for a given metric and time point. For analyses of heterogeneities, we use the equivalent confidence intervals, limited to the relevant subpopulations.

In addition to calculating this absolute change, we also calculate the percentage change in average outcome due to note attachment: $\frac{Y_{\cdot mt}(1) - \hat{Y}_{\cdot mt}(0)}{\hat{Y}_{\cdot mt}(0)}$. This metric normalizes by the average estimated outcome had notes never been attached to the posts considered, which speaks to the overall impact of the Community Notes program but also, in essence, includes the time that notes took to be written and rated as a part of treatment. To also summarize the effect of notes once they have been attached, we also consider the growth after treatment: $\Delta Y_{\cdot mt}(1) = Y_{\cdot mt}(1) - Y_{\cdot m0}(1)$ in the treatment unit, along with the comparable growth in the control unit, $\Delta \hat{Y}_{\cdot mt}(0) = \hat{Y}_{imt}(0) - \hat{Y}_{im0}(0)$. The percentage change in growth is then $\frac{\Delta Y_{\cdot mt}(1) - \Delta \hat{Y}_{\cdot mt}(0)}{\Delta \hat{Y}_{\cdot mt}(0)}$.

Acknowledgments

We thank Jennifer Allen, Michael Bernstein, Kayla Duskin, Dean Eckles, and Amy X. Zhang for helpful feedback and discussions. This work was supported in part by a University of Washington Information School Strategic Research Fund and Army Research Office Multidisciplinary University Research Initiative Award W911NF-20-1-0252.

References

[1] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780, 2020.

[2] Ullrich Ecker, Jon Roozenbeek, Sander van der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky. Misinformation poses a bigger threat to democracy than you might think. *Nature*, 630(8015):29–32, 2024.

[3] R Kelly Garrett, Jacob A Long, and Min Seon Jeong. From partisan media to misperception: Affective polarization as mediator. *Journal of Communication*, 69(5):490–512, 2019.

[4] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 101–110, 2014.

[5] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[6] Jonas L Juul and Johan Ugander. Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences*, 118(46):e2100786118, 2021.

[7] Sandra González-Bailón, David Lazer, Pablo Barberá, William Godel, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M Guess, et al. The diffusion and reach of (mis) information on facebook during the us 2020 election. *Sociological Science*, 11:1124–1146, 2024.

[8] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[9] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 6(10):1372–1380, 2022.

[10] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M Herzog, Ullrich KH Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, et al. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, 8:1044–1052, 2024.

[11] Hanaa' Tameez. Twitter wants to know if you read that article before you retweet it, 2020. `https://www.niemanlab.org/2020/06/twitter-wants-to-know-if-you-read-that-article-before-you-retweet-it/`. Accessed: 14 February 2025.

[12] Kevin Aslett, Andrew M Guess, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, 8(18):eabl3844, 2022.

[13] Jon Roozenbeek and Sander Van der Linden. The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5):570–580, 2019.

[14] Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545, 2020.

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

[16] Ryan C Moore and Jeffrey T Hancock. A digital media literacy intervention for older adults improves resilience to fake news. *Scientific Reports*, 12(1):6008, 2022.

[17] Smitha Milli, Luca Belli, and Moritz Hardt. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 714–722, 2021.

[18] Michael Bernstein, Angèle Christin, Jeffrey Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nicole Meister, Nathaniel Persily, Tiziano Piccardi, Martin Saveski, et al. Embedding societal values into social media algorithms. *Journal of Online Trust and Safety*, 2(1), 2023.

[19] Tom Cunningham, Sana Pandey, Leif Sigerson, Jonathan Stray, Jeff Allen, Bonnie Barrilleaux, Ravi Iyer, Smitha Milli, Mohit Kothari, and Behnam Rezaei. What we know about using non-engagement signals in content ranking. *arXiv preprint arXiv:2402.06831*, 2024.

[20] Adam Mosseri. Facebook: Addressing hoaxes and fake news, 2016. `https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/`. Accessed: 14 February 2025.

[21] Instagram. Combatting misinformation on instagram, 2019. `https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram`. Accessed: 14 February 2025.

[22] Yoel Roth and Nick Pickles. Updating our approach to misleading information, 2020. `https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information`. Accessed: 14 February 2025.

[23] Cameron Martel and David G. Rand. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, 8:1957–1967, 2024.

[24] Ethan Porter and Thomas J Wood. Political misinformation and factual corrections on the facebook news feed: Experimental evidence. *The Journal of Politics*, 84(3):1812–1817, 2022.

[25] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42:1073–1095, 2020.

[26] Morgan Wack, Kayla Duskin, and Damian Hodel. Political fact-checking efforts are constrained by deficiencies in coverage, speed, and reach. *arXiv preprint arXiv:2412.13280*, 2024.

[27] Mason Walker and Jeffrey Gottfried. Republicans far more likely than democrats to say fact-checkers tend to favor one side, 2019. `https://pewrsr.ch/2Fz9e22`. Accessed: 14 February 2025.

[28] Francis Galton. Vox populi. *Nature*, 75:450–451, 1949.

[29] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eabf4393, 2021.

[30] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[31] Paul Resnick, Aljohara Alfayez, Jane Im, and Eric Gilbert. Searching for or reviewing evidence improves crowdworkers' misinformation judgments and reduces partisan bias. *Collective Intelligence*, 2(2):26339137231173407, 2023.

[32] Antonio A Arechar, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro, Yunhao Zhang, Gordon Pennycook, and David G Rand. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, 7(9):1502–1513, 2023.

[33] Keith Coleman. Introducing birdwatch, a community-based approach to misinformation, 2021. `https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation`. Accessed: 14 February 2025.

[34] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723*, 2022.

[35] Daniel B Larremore, Bryan Wilder, Evan Lester, Soraya Shehata, James M Burke, James A Hay, Milind Tambe, Michael J Mina, and Roy Parker. Test sensitivity is secondary to frequency and turnaround time for covid-19 screening. *Science Advances*, 7(1):eabd5393, 2021.

[36] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1):20150021, 2017.

[37] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

[38] Alberto Abadie and Jérémy L'Hour. A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834, 2021.

[39] Alberto Abadie. Using synthetic controls: Feasability, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

[40] KJ Rothman. *Modern Epidemiology*. Lippincott Williams & Wilkins, 2008.

[41] Yuwei Chuai, Moritz Pilarski, Thomas Renault, David Restrepo-Amariles, Aurore Troussel-Clément, Gabriele Lenzini, and Nicolas Pröllochs. Community-based fact-checking reduces the spread of misleading posts on social media. *arXiv preprint arXiv:2409.08781*, 2024.

[42] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2009.

[43] Jürgen Pfeffer, Daniel Matter, and Anahit Sargsyan. The half-life of a tweet. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1163–1167, 2023.

[44] Florian Meier, David Elsweiler, and Max Wilson. More than liking and bookmarking? towards understanding twitter favouriting behaviour. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):346–355, 2014.

[45] Chih-Yu Chin, Hsi-Peng Lu, and Chao-Ming Wu. Facebook users' motivation for clicking the "like" button. *Social Behavior and Personality: an international journal*, 43(4):579–592, 2015.

[46] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10, 2010.

[47] Sharad Goel, Duncan J. Watts, and Daniel G. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 623–638, 2012.

[48] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J. Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2016.

[49] Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3):350–375, 2020.

[50] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

[51] JP Kincaid. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Chief of Naval Technical Training*, 1975.

[52] Gordon Pennycook, Tyrone D. Cannon, and David G. Rand. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12):1865–1880, 2018.

[53] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(Volume 27, 2001):415–444, 2001.

[54] Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11):4944–4957, 2020.

[55] Kevin Munger. The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media+ Society*, 5(3):2056305119859294, 2019.

[56] Soham De, Jay Baxter, Michiel Bakker, and Martin Saveski. Supernotes: Driving consensus in crowd-sourced fact-checking. In *Proceedings of the ACM Web Conference 2025*, pages 1086–1097, 2025.

[57] Community Notes. Ranking notes: Under the hood, 2024. Accessed: 14 February 2025.

[58] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021.

[59] Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, 1820.

[60] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

[61] James J. Heckman and V. Joseph Hotz. Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, 84(408):862–874, 1989.

[62] X. Note ranking algorithm, 2023. `https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes`. Accessed: 14 February 2025.

[63] Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American chemical society*, 69(1):17–20, 1947.

Appendix

A Heterogeneity in Effects on Reposts Based on Popularity Before Note Attachment

As noted in Section 3.3, we find that a post's popularity prior to receiving a community note is associated with the absolute treatment effect it receives. As in the main text, we center our analysis on reposts. We find that for posts with the fewest reposts prior to having a note attached (i.e., those in the first quartile, having 134 or fewer reposts prior to attachment), the mean absolute treatment effect after 48 hours was a small positive change of 2.66 reposts (95% CI: [-9.15, 14.5]) or 3.07%. In comparison, the absolute changes for the second (134–537 reposts), third (537–1,821 reposts), and fourth (1,821+ reposts) quartiles were -56.5 (95% CI: [-67.7, -45.2]), -228 (95% CI: [-260., -195]), and -743 reposts (95% CI: [-840, -647]), respectively, corresponding to percentage changes of -13.0%, -14.8%, and -10.8%.

The similarity in percentage changes between posts in the second, third, and fourth quartiles suggests that much of the difference in absolute changes (-56.5, -228, and -743) may be due simply to the number of users who see a post and can therefore be affected by note attachment. In other other words, community notes on posts of a large enough size have relatively similar effects on individual viewers' reposting behavior. However, community notes attached to posts in the smallest popularity quartile do appear to have a different aggregate effect on reposting behavior. Beyond the fact that this quartile has a positive average treatment effect, we also show in Figure S5 that posts in this quartile are simply more likely to have notes lead to increases in reposts: 44.2% of posts in the bottom quartile have positive treatment effects, relative to 35.2%, 29.4%, and 21.8% of posts in larger quartiles. These findings may indicate that attaching notes at times draws additional attention to low-visibility posts, resulting in a higher likelihood of positive changes. With that said, we also note that the coefficient of variation is much larger in magnitude for small posts: 89.6, compared to -4.02, -2.90, and -2.61 for posts in the next quartiles. The large amount of relative variation in this quartile suggests that notes attached to low-visibility posts tend to have a less consistent effect than notes on more visible posts.

B Factors Associated with Note Effectiveness on the Number of Views, Replies, and Likes

In addition to investigating factors associated with large effects on reposts, we also investigate factors associated with large effects on views, replies, and likes. We plot conditional treatment effects on views, replies, and likes in Figure 5, Figure 6, and Figure 7, respectively. Most of the factors associated with large effects on reposts tend to have similar associations with these additional metrics, however we note some nuances below. First, considering the partisanship of the post, we find that the treatment effects on views, replies, and likes are all larger in magnitude on right-leaning posts than left-leaning posts. This gap appears largest for replies and views, metrics that are not as clear of a positive signal of agreement from a user as a reply or a like. In fact, we estimate positive (although not statistically significant) average treatment effects on views and replies for left-leaning posts. Another trend we note is that while the treatment effect on reposts declines as notes become more readable, the association with views is in the opposite direction: the more readable a note is, the larger the effect it tends to have on a post's repost, but the smaller the effect it tends to have on views. This observation supports the hypothesis that unclear notes may act as a warning signal, causing users to quickly move past a post. In contrast, more comprehensible notes may engage users, making them less likely to later express agreement with the post.

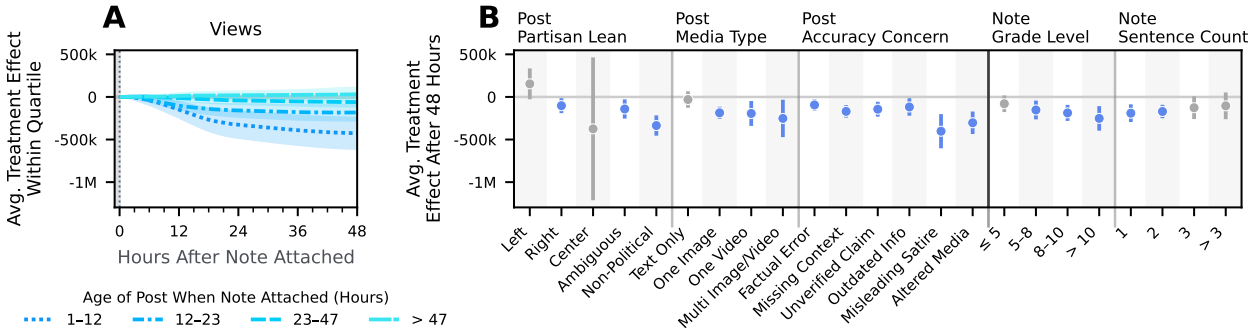

Figure 5: Factors associated with effect on views. *(A)* Average treatment effects for posts, stratified by the amount of time that elapsed between each post's creation and when it first had a note attached. Bins show note speed quartiles. *(B)* Average treatment effects after 48 hours, based on (i) the partisanship of the post, (ii) the number of images and videos included, (iii) the accuracy concern raised by the community note writer, (iv) readability of the community note, as measured by Flesch–Kincaid grade level, and (v) length of the note.

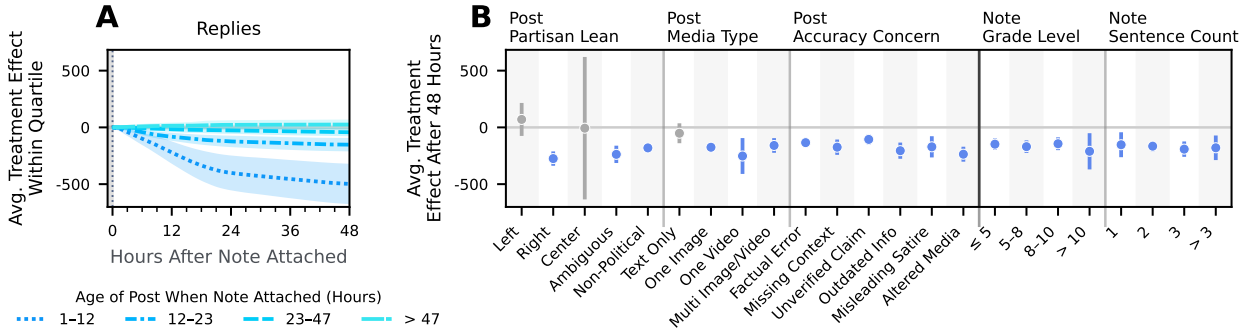

Figure 6: Factors associated with effect on replies. *(A)* Average treatment effects for posts, stratified by the amount of time that elapsed between each post's creation and when it first had a note attached. Bins show note speed quartiles. *(B)* Average treatment effects after 48 hours, based on (i) the partisanship of the post, (ii) the number of images and videos included, (iii) the accuracy concern raised by the community note writer, (iv) readability of the community note, as measured by Flesch–Kincaid grade level, and (v) length of the note.

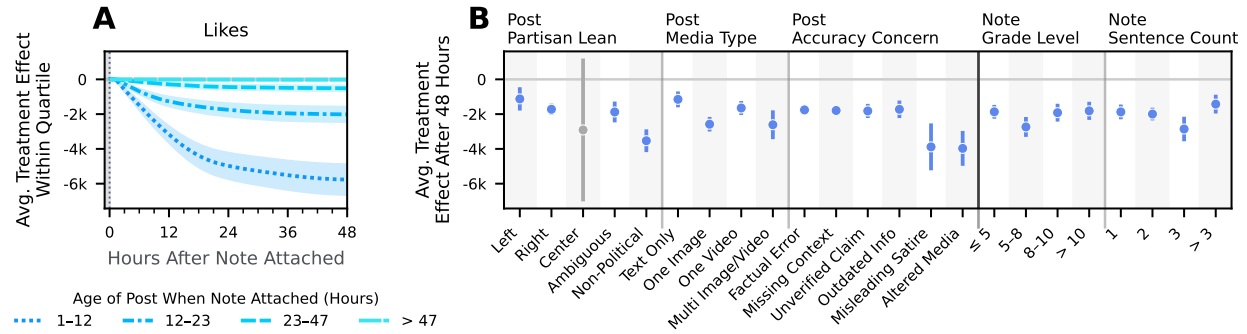

Figure 7: Factors associated with effect on likes. *(A)* Average treatment effects for posts, stratified by the amount of time that elapsed between each post's creation and when it first had a note attached. Bins show note speed quartiles. *(B)* Average treatment effects after 48 hours, based on (i) the partisanship of the post, (ii) the number of images and videos included, (iii) the accuracy concern raised by the community note writer, (iv) readability of the community note, as measured by Flesch–Kincaid grade level, and (v) length of the note.

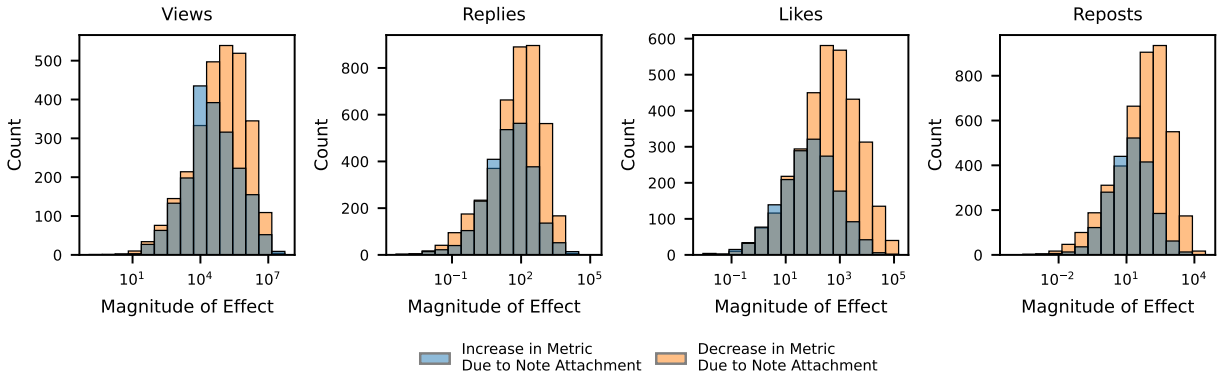

Figure 8: Distribution of individual treatment effects on each reach and engagement metric. Color indicates whether the treatment effect was positive (meaning the note resulted in an increase in the metric) or negative, while location on x-axis indicates the magnitude of the increase or decrease. X-axis is in log scale. Plots show treatment effects after 48 hours.

C Distribution of Treatment Effects

While we estimate that community notes result in declines in engagement on average, for many posts we find that note attachment leads to *increased* engagement relative to their synthetic controls. We estimate that 41.6% of posts see increases in views after 48 hours due to note attachment, while the comparable number for replies is 37.8%, for likes is 33.9%, and for reposts is 32.8%. However, the estimated increases are typically smaller in magnitude than the decreases, resulting in the sizable negative average effects we observe. We plot the distributions of individual treatment effects for these four metrics in Figure 8, showing increase and decreases separately. The plots show histograms of treatment effect magnitudes on a common log-scale axis, where the color indicates whether the treatment effect was positive or negative.

Considering positive and negative treatment effects separately, the average effects are 529,230 and -659,571 for views, respectively; 316 and -461 for replies; 774 and -3,550 for likes; and 117 and -436 for reposts. The medians show similar differences: the median positive treatment effect for views is 27,841, for example, about three times smaller in magnitude than the median negative treatment effect of -94,895. For replies, the medians are 35.6 and -93.6; for likes, the medians are 81.0 and -439; and for reposts, the medians are 15.2 and -73.9. (Medians are non-integer as synthetic control estimates need not be integer.) This analysis suggests that while community notes do at times result in increases in engagement, the increases tend to be both less frequent and smaller in magnitude than the decreases.

We state in the Section 3.1 the average declines in views and replies after note attachment are -14.0% and -22.9%, respectively, compared to -43.5% and -45.7% for likes and reposts. The difference between these sets of metrics suggests that community notes may have a stronger impact on engagement that clearly signals support for a post than on the number of people who simply view it or who engage with it in a more ambiguous way. The percentage of treatment effects that are positive (41.6% and 37.8% for impressions and replies, compared to 33.9% and 32.8% for likes and reposts) suggests that part of this effect is due to community notes more frequently leading to increases in views and replies than in likes and reposts. The distributions of magnitudes for positive and negative effects suggest another dimension to the differences in averages. When comparing positive and negative treatment effects that come from the same percentile of the magnitude distribution, we find that, for views and replies, the positive and negative treatment effects tend to be more similar to each other than when comparing percentiles for likes and reposts. Considering the median, for example, the median negative views and replies treatment effects are 3.41 and 2.64 times as large as their positive counterparts, respectively. In comparison, the median negative likes and reposts treatment effects are 5.42 and 4.87 times as large as their positive counterparts. When considering

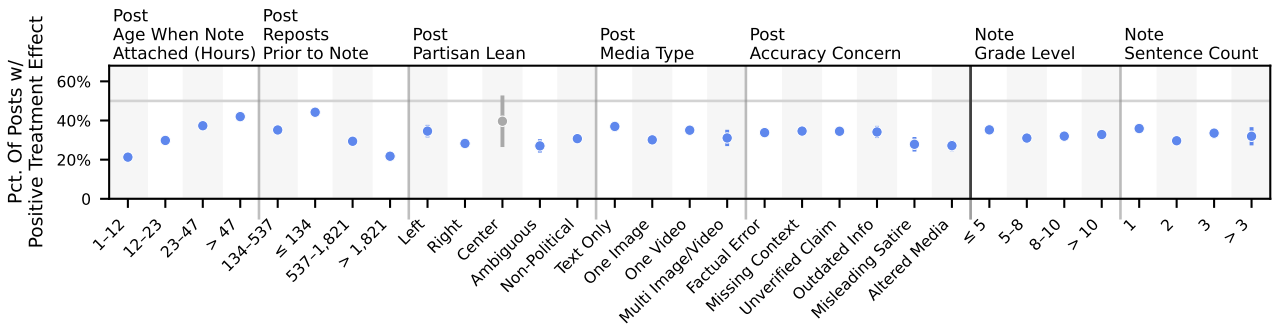

Figure 9: Factors associated with positive treatment effects on reposts. Considering all treated posts, the percentage that have positive individual treatment effects after 48 hours. Disaggregated by (i) the number of hours elapsed between post creation to note attachment (ii) the number of reposts the post had received at the time of note attachment (iii) the partisanship of the post, (iv) the number of images and videos included, (v) the accuracy concern raised by the community note writer, (vi) readability, as measured by Flesch–Kincaid grade level, and (vii) length of the community note. Plot uses binomial proportion interval [59], commonly referred to as Wald interval, rather than Gaussian.

the complete set of percentiles between the median and 99th, we in fact find that the maximum ratios of negative to positive treatment effect (again, how much larger the negative effect is than the equivalent positive effect) are 3.41 and 2.74 for impressions and replies, compared to maximum ratios of 6.53 and 6.09 for likes and reposts. The *minimum* ratios among these percentiles is in fact 4.58 and 3.05 for likes and reposts, compared to 0.877 and 0.817 for views and replies. These findings suggests that beyond positive signals of engagement being more likely than non-positive signals to decrease due to note attachment, when they do decrease, they also tend to do so by a relatively larger amount.

Considering reposts, as we do in Figure 4, we find that many factors associated with differences in average treatment effects, e.g., the amount of time between post creation and note attachment, are also associated with posts' probabilities of having a positive treatment effect. In Figure 9, we plot the percentage of treatment effects that are positive, disaggregated by variables previously found to be associated with differences in treatment effects. Based on the plot, the age of a post at note attachment appears to have one of the stronger pairwise associations with whether attachment will lead to an uptick in reposts: 21.8% of posts in the most rapid quartile (those noted within 12 hours of creation) have positive treatment effects, compared to 44.2% of posts in the final quartile (those noted after 47 hours). (In Figure 9, we also plot the percent positivity based on the number of reposts that the post had received prior to having a note attached, as previously discussed above in Appendix A.) Partisan lean, media type, accuracy concern, grade level, and sentence count show similar, if less strong, relationships with the positivity of the reposts treatment effect as they do with its value.

D In-Time Placebo Test

Traditional methods for performing statistical inference are difficult to adapt to Synthetic Controls, and as such, researchers typically validate SCM-based estimates using falsification experiments [60]. Similar to A/A tests, falsification experiments involve altering the data such that no treatment effect should be observed, before rerunning the synthetic control process to ensure that no effect is in fact estimated. One such exercise is the in-time placebo test [39], first introduced by Heckman and Hotz [61]. This test involves "backdating" the treatment time—artificially considering it to have occurred earlier than it did in reality—then matching only on observations up until this backdate time. In this test, any non-null treatment effects observed during the backdate period (between the backdate and the treatment time) would suggest that treatment effects are not due to treatment itself, and would therefore undermine the credibility of the SCM estimates.

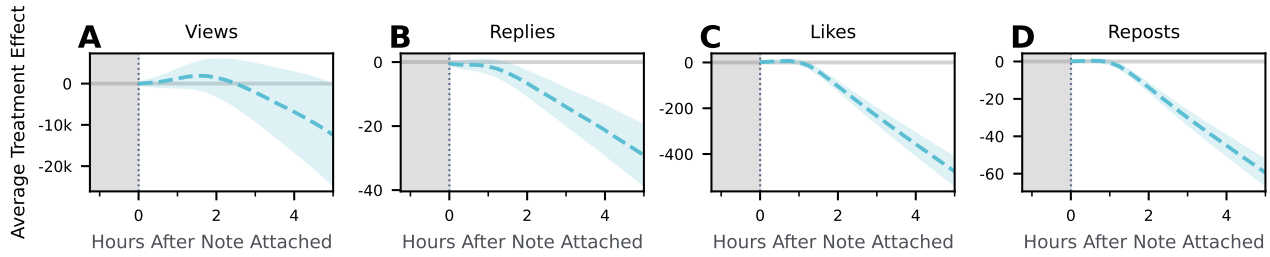

Figure 10: Effects of note attachment on views, replies, likes, and reposts during first five hours after note receiving a helpful rating.

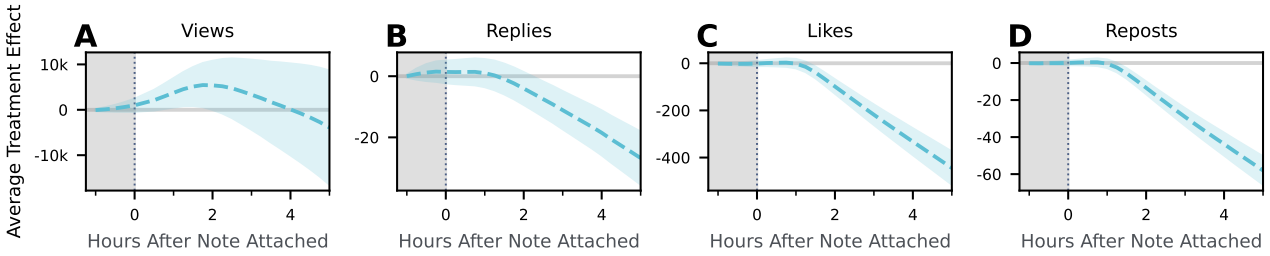

Figure 11: One hour placebo test: Effects of note attachment on views, replies, likes, and reposts when using a one hour backdate. The estimated effects at the end of the backdate period, when we would expect to find a null effect, are 990 views (95% CI: [-394, 2,374]), 1.40 replies (95% CI: [-2.21, 5.01]), -1.22 likes (95% CI: [-10.7, 8.26]), and 0.05 reposts (95% CI: [-1.03, 1.14]). The slight initial increase and later decrease in views is likely due to posts appearing in the "Rated Helpful" tab on the Community Notes webpage, which displays the most recently noted posts and is frequently visited by Community Notes volunteers.

Prior to October 23rd of 2023 there was a delay between a note being rated as helpful and the note being displayed to users [62]. There are two reasons for this delay: (1) the timestamps indicating when the note had a "helpful status" in the publicly available data correspond to the start of the Community Notes algorithm's computation, as documented in the open-source code, and (2) the time required for the algorithm's output to propagate through X's system and for the notes to start appearing on the platform. This delay typically lasts around one hour, however there is unknown variation in the length of the delay, and we therefore chose to consider the time the post first received a helpful note as its treatment time. (We use the term "note attachment" to refer to the treatment time throughout our work for ease of exposition, and differentiate between note attachment and treatment time in this section only.) Because treatment typically occurred around an hour after the actual treatment time used, our SCM estimation therefore includes a natural in-time placebo test.

While not easily visible in Figure 2 due to the axes' scales, we do find a null effect extending until around one hour after the first note was rated helpful. Figure 10 shows a subset of Figure 2, focusing on only the five hours after the first helpful rating, rather than the full 48 hours. As seen in the figure, we find that replies, likes, and reposts do not begin to decline meaningfully until around one hour after treatment, which is in fact when notes typically began to be displayed to users. The slight initial increase and later decrease in views is likely due to posts appearing in the "Rated Helpful" tab on the Community Notes webpage, which displays the most recently noted posts and is frequently visited by Community Notes volunteers.

In addition to the natural in-time placebo test, we also perform an in-time placebo test with a one hour backdate: matching on observations only until one hour prior to a post received its first helpful note. As shown in Figure 11 we again find that the treatment effect does not begin to decline until around one hour after treatment occurs. We also find that the confidence intervals we estimate after backdating include 0 for all metrics. One hour after the matching period ends (i.e., at the time of note attachment), the estimated effects were 990 views

(95% CI: [-394, 2,374]), 1.40 replies (95% CI: [-2.21, 5.01]), -1.22 likes (95% CI: [-10.7, 8.26]), and 0.05 reposts (95% CI: [-1.03, 1.14]).

E Additional Details Concerning Data Processing

E.1 Data Collection Pipeline

We collected data from March 16 to June 23, 2023. On June 23, 2023, the Academic Research API was discontinued, and a new pricing structure was implemented, making further data collection prohibitively expensive. We initially collected data for 41,310 posts that were the subject of community notes, using the `/2/tweets` X API end-point to collect counts of engagement views, likes, replies, and reposts over time. During data processing, we identified anomalies in engagement metrics returned by the API for 874 of these posts (2.1%). These anomalies consisted of sharp rises and declines in engagment metrics (usally only a single engagement metric) which we hypothesize were due to data processing issues on the X backend. We removed posts exhibiting such anomalies from our dataset, resulting in the inclusion of 40,436 posts after filtering. Full details concerning this process are available in Appendix E.6.

We applied two additional filters: posts needed at least one hour of data prior to note attachment to ensure reliable synthetic control construction, and at least 48 hours of data afterward to maintain consistency in the set of posts used for averages at different time points. These criteria excluded 355 posts (0.9%), reducing the dataset size to 40,081. Seven treated posts were later excluded due to unsolvable convex programs, leaving a final count of 40,074.

In addition to collecting engagement data from the `/2/tweets` end-point, we also used the full-archive search (`/2/tweets/search/all`), recent search (`/2/tweets/search/recent`), as well as the follow graph (`/2/users/:id/followers` and `/2/users/:id/following`) end-points. These endpoints provided information concerning which accounts had publicly reposted and replied to posts that received community notes, which allowed us both to calculate exact values of reposts and replies at any given time as well as to construct repost cascades. Because these counts were exact, we used them as our default repost and reply count metrics, falling back to the repost and reply counts returned by the `/2/tweets` endpoint when necessary.

E.2 Defining Treatment Status

The "bridging-based" matrix factorization algorithm that classifies community notes as "helpful" is run every hour. While, the majority of notes classified as helpful remain helpful for the 48 hour duration that we analyze note effects, some notes lose their helpful status. In our analysis, if a post had a helpful note at any point, even if it was later reclassified as unhelpful, we consider it to be a part of the treatment group for the entirety of its life. Of 6,753 treated posts in our analysis, 5,362 had a helpful note for the entire 48 hour period after they first received one (79.4%), meaning 20.6% of posts at some point experienced at least one additional change in status.

Helpful to unhelpful transitions tend to occur after many viewers have already seen a post. The median transition occurred 15.5 hours after posts were originally rated helpful and 65.9% of views for posts that made a helpful-to-unhelpful transition occurred while the post had a helpful note. Together, these facts result in most views for treated posts overall occurring when the posts have notes attached: among all treated posts, 93.4% of views in the 48 hours after note attachment occurred while posts were rated helpful.

Only one note can be attached to a post at a time, and in cases where multiple notes are found to be helpful for a given post, one is selected randomly by X to be shown. Of the 6,753 treated posts, 854 had more than one note found helpful in the 48 hours after initial note attachment (12.6%). For posts with multiple notes rated as helpful, we use the note that was first to receive a helpful rating to determine the treatment time. Considering all treated

posts, there were a total of 7,759 notes potentially shown with the 6,753 treated posts in the 48 hours following first note attachment.

Data was not available concerning which note was actually shown alongside a post at a given time. When calculating post-level statistics related to the note that a post received, e.g., post's reading grade level in Figure 4, we take a weighted average across the different notes that appeared with a post, where weights are proportional to the amount of time in the 48 hours after note attachment that each individual note spent rated as helpful.

E.3 Structural Characteristics of Repost Cascades

For a given post, we construct its repost cascade as a directed tree using time-inferred diffusion [47], a standard procedure for attributing diffusion pathways [5]. The post itself is the root node, and each repost constitutes another node in the cascade. Each repost has a directed edge to the likely source from which the reposter saw the post, according to the following procedure: For each repost, we scan the set of users whom they follow, to find the user that most recently shared the content. A directed edge is then added pointing follower's repost node to the followee's repost node. If no user the reposter follows has reposted the content previously, a directed edge is added to the root node.

The maximum depth of a repost cascade refers to the longest path between the original post and any individual repost in the cascade graph. The maximum breadth is defined as the maximum number of reposts occurring at any single level of depth. In addition to breadth and depth, we also calculate the structural virality of a post's repost cascade, a metric closely related to the Wiener index [63] that captures the extent to which a post was shared by people who follow one another. Structural virality is calculated as the average distance between all pairs of nodes in the repost cascade, taken to be undirected, and normalized by the size of the cascade [47]. We calculate the exact values of maximum depth, maximum breadth, and structural virality every 15 minutes after a post was created, based on the timestamps of reposts as given by the API. Since repost data was generally collected after the fact, deleted reposts are not included in the cascades. Reconstructing repost cascades requires data we were not able to collect for all posts, and as such we only analyzed cascades for the posts which did have this data available, (see Appendix E.5 for further details).

E.4 Post Partisanship Classification

We used the Claude 3.5-Sonnet V2 model ("claude-3-5-sonnet-v2@20241022") to label the partisanship of English-language posts, as well as to provide a number of other nominal and ordinal classifications of the post. The prompt used was "Please annotate the tweet below using the following schema. It is okay if you cannot view photos, videos, or links from the tweet, but please do your best to interpret the text and any available context. Ensure that your response ends with a correctly formatted JSON containing all requested columns. Make sure that you do not include any other text after the JSON." A full JSON schema was provided, in which the description of the partisanship class was given as "The political leaning of the tweet content. The category 'unknown' indicates the political leaning is unclear, and 'none' indicates the content is not political." (In plots, "unknown" was changed to "ambiguous" and "none" was changed to "non-political" for clarity.) There were 904 posts labeled as "left," 1,361 posts labeled as "right," 53 posts labeled as "center," 692 posts labeled as "unknown," 1,910 labeled as "none," 1,580 non-English posts not labeled, and 253 posts not labeled due to irretrievable post text. Full details for reproduction, including descriptions of the other classifications made by the model, are available in the code repository for this project.

E.5 Data Missingness

Along with data released by X as part of the Community Notes project, we also requested data from multiple X API end-points for analysis. While the API returned most requests, there were cases where one API end-point returned data concerning a post while another did not, or where data was returned from a single end-point with observations of one metric but not another. As we requested data each 5 minutes and later linearly interpolated observations to be at 15 minute intervals from their post's time of creation, this mostly did not pose an issue to analysis. However, for some posts we were never able to retrieve responses from a given end-point or found that responses from an endpoint were consistently missing observations of a metric. We also found cases where a single end-point would stop returning requests after a certain time while other end-points would continue returning requests, and similarly where some metrics would stop being returned at a given time, while others would not. Rather than remove these posts from analysis, we instead elected to simply use the metrics that we did have available. For posts missing views observations, for example, we did not include views in the synthetic control matching process, and did not estimate a views treatment effect. Below, we present the number of posts that had each metric available.

	Control		**Treatment**	
	Unavailable	Available	Unavailable	Available
Reposts	110	33,211	16	6,737
Replies	122	33,199	10	6,743
Likes	740	32,581	208	6,545
Impressions	1,664	31,657	368	6,385
Repost Cascade Depth	10,088	23,233	1,172	5,581
Repost Cascade Width	10,088	23,233	1,172	5,581
Repost Cascade Structural Virality	10,365	22,956	1,177	5,576

Table 1: Missingness by metric and treatment status. Number of posts that had at least one observation available for a metric, compared to number of posts for which observations were never returned.

	Treatment				
	Fully Available	Dropped Pre-treatment	Dropped Post-treatment	Only Post-treatment	Unavailable
Reposts	6,412	49	261	15	16
Replies	6,627	1	6	109	10
Likes	4,955	70	375	1,145	208
Impressions	4,843	69	366	1,107	368
Repost Cascade Depth	4,263	404	914	0	1,172
Repost Cascade Width	4,263	404	914	0	1,172
Repost Cascade Structural Virality	4,258	402	897	19	1,177

Table 2: Amount of data available for treatment posts. For each metric, the number of posts that had observations (i) available for at least 48 hours after treatment, (ii) available at some point before treatment, but dropped before treatment occurred, (iii) available at points both before and after treatment, but not for a full 48 hours, (iv) only after treatment had already occurred, and (v) never available

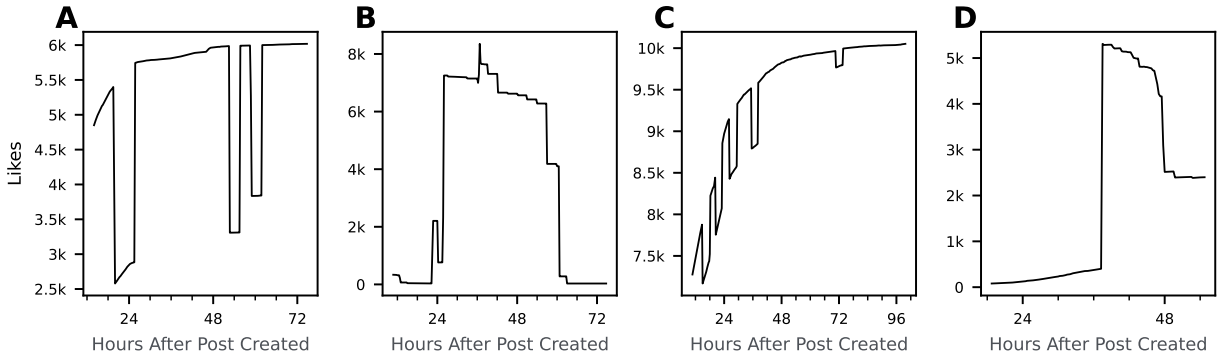

Figure 12: Likes from sample posts considered anomalous. These posts, along with 870 additional posts, were removed from analysis after investigation.

E.6 Anomalous Post Removal

When exploring our dataset, we identified several posts that we considered to be anomalous. These posts showed large rises and drops in a subset of the post's engagement metrics, while other metrics showed smooth increases for the same time period. In most cases only a single engagement metric was affected. Examples of such posts are shown in Figure 12. We did not believe the metrics returned by the X API for these posts accurately reflected the true number of people who had engaged with them, and as such elected to remove them from analysis. To identify posts for removal, we began by plotting a random sample of 300 posts that showed a more than 1% rise or drop in at least one metric between API calls, which must have also amounted to more than 10 absolute units, a permissive heuristic that captured all posts identified as anomalous during initial checks. Two authors (M.S. and I.S.) each labeled the reach and engagement metrics from 150 posts as either (1) accurate or (2) likely erroneous and worthy of removal. After discussing labeling criteria, they then conducted a second round of labeling on an additional random sample of 300 posts that exhibited between a 0.5% and 3% drop in a metric (amounting to between 20 and 100 absolute units), combined with a rise of the same magnitude. The authors each labeled 200 of the 300 posts, and gave the same label to 49 of the 50 overlapping records, (Krippendorff's $\alpha = 0.96$, 95% CI: $0.85, 1$).

After labeling a total of 600 posts, we tested various criteria for identifying erroneous posts. We performed a grid search to identify combinations of percentage rise and drop, and absolute rise and drop that minimized mislabeling. From the grid search we selected the thresholds that maximized the number of posts that were correctly removed, among solutions that did not incorrectly remove *any* posts from the labeled training data. The optimized thresholds were a rise of 25 units amounting to at least 3%, and at another point in time, a drop of 25 units amounting to at least 3%. We finally applied these thresholds to the complete dataset, removing a total of 874 posts.