Can Crowdchecking Curb Misinformation? Evidence from Community Notes

Yang Gao
Gies College of Business, University of Illinois Urbana-Champaign, ygao1@illinois.edu,

Maggie Mengqing Zhang
McIntire School of Commerce, University of Virginia, Charlottesville, Virginia 22903 ffx3ab@virginia.edu,

Huaxia Rui
Simon Business School, University of Rochester, huaxia.rui@simon.rochester.edu,

To battle against rampant misinformation on social media, many platforms are experimenting with crowdsourced fact-checking—systems that rely on social media users' annotations of potentially misleading content. This paper investigates the efficacy of such systems in curbing misinformation in the context of Community Notes, a pioneering crowdsourced fact-checking system from Twitter/X. Utilizing a regression discontinuity design, we empirically identified the positive effect of publicly displaying community notes on an author's voluntary retraction of the noted tweet, demonstrating the viability of crowdsourced fact-checking as an alternative to professional fact-checking and forcible content removal. Our findings reveal that the effect is primarily driven by the author's reputational concern and perceived social pressure, and there is considerable heterogeneity of such effect, depending on specific tweet- and user-level characteristics. Platforms, therefore, can exploit the underlying mechanism and explore the use of contextual factors to harness the full potential of crowdsourced fact-checking. Furthermore, results from discrete-time survival analyses show that publicly displaying community notes not only increases the probability of tweet retractions but also accelerates the retraction process among retracted tweets, thereby improving platforms' responsiveness to curb misinformation. This study offers important insights to both social media platforms and policymakers on the promise of crowdsourced fact-checking and calls for the broad participation of social media users to collectively tackle the problem of misinformation.

Key words: Misinformation, Fact-checking, Content Moderation, Crowdsourcing, Community Notes

---

"*If it is valid, it cannot be retracted, any more than the dead can be brought to life.*"

— Abraham Lincoln (1863)

1. Introduction

The rampant spread of misinformation on social media poses critical challenges to democratic processes, public health, and social stability (Lazer et al. 2018), highlighting a pressing need for effective countermeasures. For instance, during the 2016 U.S. presidential election, approximately 6% of people who shared political URLs included fake news sources on Twitter (Grinberg et al. 2019), and misinformation significantly undermined compliance with health guidelines and vaccine acceptance during the COVID-19 pandemic (Roozenbeek et al. 2020). Traditionally, social media platforms have relied on professional fact-checking organizations like FactCheck.org to counteract misinformation. However, this approach is limited by its scalability and is sometimes criticized for its potential bias in evaluating content.

To overcome these limitations, social media platforms have increasingly turned to crowdsourced fact-checking (henceforth *crowdchecking*) to leverage the "wisdom of crowds." This approach not only democratizes the fact-checking process by engaging a broader community with more diverse perspectives (Borwankar et al. 2022), but also enhances scalability and responsiveness. Twitter/X, the pioneer of crowdchecking, introduced its Community Notes (formerly known as Birdwatch) in 2021, a system that enables users to collaboratively add annotations to potentially misleading tweets, which are then publicly displayed beneath the tweets when deemed helpful. Inspired by this model, other major platforms have introduced similar initiatives. YouTube launched a pilot program that allows viewers to add contextual notes to videos, helping clarify potentially misleading content.[1] Meta began replacing professional fact-checking with a crowdchecking system in early 2025, a move that has sparked heated public debate.[2] TikTok introduced a new feature called Footnote, which integrates community-contributed annotations to contextualize viral posts.[3] Similarly, Weibo, a leading microblogging platform in China, has implemented a crowdchecking function to empower users to identify misinformation.[4]

[1] For details, see https://blog.youtube/news-and-events/new-ways-to-offer-viewers-more-context/

[2] For details, see https://transparency.meta.com/features/community-notes

[3] For details, see https://newsroom.tiktok.com/en-us/footnotes

[4] For details, see https://www.sixthtone.com/news/1013647

Advocates argue that crowdchecking is a better alternative to traditional content moderation strategies, such as content removal, which often faces criticism for infringing on freedom of speech (Hwang and Lee 2025). Crowdchecking relies on a broad base of users to participate in the process, thereby circumventing issues related to censorship and enhancing transparency[5] while safeguarding information accuracy at the same time. For example, Allen et al. (2024) found that community notes on Twitter were generally accurate and cited high-quality sources, particularly on contentious issues like COVID-19 vaccines. More importantly, voluntary retraction by authors of misinformation, whether under pressure or out of reputation concern, is a more civilized approach to resolving disagreement than forcible removal of content which often sparks concerns of censorship and may result in further polarization.

On the other hand, opponents of crowdchecking raise concerns about its actual effectiveness in curbing the spread of misinformation. Since crowdsourced fact-checking is performed by peers rather than professionals, many doubt its reliability and credibility. Indeed, Draws et al. (2022) revealed cognitive bias in crowdchecking, with users more likely to overestimate the truthfulness of claims they ideologically agreed with. This behavior could skew the perceived neutrality of fact-checking efforts, making them less effective at fostering informed discussions and more likely to be viewed as vehicles for partisan conflict. Insiders of Twitter's Community Notes expressed[6] similar concerns about partisan bias and preferential flagging, which may worsen the problem of misinformation and further lead to polarization. Even if crowdchecking is objectively accurate, whether authors of misleading posts would perceive so and voluntarily retract their content is unclear. Humans are vulnerable to confirmation bias and social media often exacerbate the problem by creating echo chambers where similar beliefs are reinforced (Garrett 2009). Worse still, some authors may even consider crowdsourced corrections personal attacks (Tang et al. 2024), which can trigger emotional responses and result in further entrenchment of their positions.

[5] For example, see https://www.bloomberg.com/opinion/articles/2024-05-22/elon-musk-s-community-notes-feature-on-x-is-working

[6] For details, see https://www.wired.com/story/x-community-notes-disinformation/

Inspired by this important debate and motivated by the urgent need to curb the spread of misinformation, we empirically evaluate the effectiveness of crowdchecking by addressing the following research question: *Does crowdchecking facilitate the voluntary retraction of misleading social media posts?* Specifically, we leverage Twitter's Community Notes as the research context to empirically investigate the effect of publicly displayed community notes on tweet retractions. Our full sample consists of tweets that received community notes during two distinct periods: June 11 to August 2, 2024, and January 1 to February 28, 2025. We systematically monitor the status of these tweets on a daily basis, capturing both the presence of any publicly displayed notes and the retraction status of tweets. In addition, we leverage a comprehensive dataset of notes and ratings released by Community Notes and apply its open-source ranking algorithms on a high-performance computing cluster to calculate the helpfulness scores of community notes.

Given the threshold mechanism of Community Notes (i.e., a note must achieve a helpfulness score of at least 0.4 to be publicly displayed below a tweet), we employ a regression discontinuity (RD) design. Results indicate that tweets with publicly displayed community notes are more likely to be retracted by the author. We then conduct a series of subsample RD analyses to further understand the underlying mechanism. Results from mechanism tests reveal that the observed effect stems from users' reputational concerns and perceived social pressure. Our analysis also identifies heterogeneous effects of community notes across different tweet- and user-level characteristics, highlighting the importance of contextual factors in the effectiveness of crowdchecking. Importantly, discrete-time survival analyses suggest that displayed notes also accelerate the retraction process itself for retracted tweets. Finally, our analysis suggests that Community Notes not only addresses contemporary misinformation but also engages with misinformation of a more general or persistent nature.

To the best of our knowledge, this is the first study to examine the effectiveness of crowdchecking from the perspective of misinformation spreaders, thereby enriching the literature on countermeasures against misinformation. Practically, our study highlights the effectiveness of crowdchecking systems like Community Notes in mitigating misinformation on social media, presenting a cost-effective alternative to traditional fact-checking methods. The mechanism tests and the heterogeneity of treatment

effects over tweet- and user-level characteristics suggest that platforms need to tailor their approaches to harness the full potential of crowdchecking. For policymakers, our findings are particularly meaningful because crowdchecking strikes a balance between protecting First Amendment rights and the urgent need to curb misinformation. Given its effectiveness on voluntary retraction, policymakers should encourage its use by all large social media companies.

The rest of the paper is organized as follows. We provide a literature review in Section 2 and develop our hypothesis in Section 3. Section 4 describes our research context and data collection process. In Section 5, we present our identification strategy. In Sections 6 and 7, we report our empirical findings. Section 8 concludes the paper by discussing its implications and limitations.

2. Literature Review

This study relates to the broad literature on the dissemination, detection, and countermeasures of misinformation on social media. The pervasiveness of misinformation across digital platforms has heightened the need to understand its spread and curtail its impact. We first summarize existing research on misinformation dissemination mechanisms and detection strategies, then transition to a detailed review of countermeasures, directly aligning with our research question in this paper.

2.1. Dissemination and Detection of Misinformation on Social Media

The dissemination of misinformation on social media has been explored from various angles, including diffusion patterns, audience susceptibility, and spreader motivations. From the diffusion perspective, Vosoughi et al. (2018) found that false news spreads significantly farther, faster, deeper, and more broadly than the truth, especially for false political news. Other studies, such as the one by Mostagir and Siderius (2022), explore factors that influence users' susceptibility to misinformation, revealing that individuals with higher cognitive sophistication and communities with unequal access to learning resources are more prone to misinformation. Regarding spreader motivations, Allcott and Gentzkow (2017) identified two drivers for sharing fake news on social media—pecuniary motivation refers to the economic incentive derived from increased audience engagement with fake news and ideological motivation stems from the desire to advance the political agenda. Additionally, Osmundsen et al. (2021)

noted that partisanship and political polarization often drive individuals to share misinformation to derogate political opponents, while Talwar et al. (2019) found that social pressures, like the fear of missing out and the desire to maintain group identity, also play a role. The role of algorithm-driven bots in amplifying misinformation dissemination also draws increasing attention from researchers. For example, Salge et al. (2022) leveraged the conduit brokerage perspective to understand how bots strategically disseminate (mis)information on social media.

The prevalence of misinformation has made efficient detection a longstanding challenge for social media platforms. Given the limitations of traditional professional fact-checking, which struggles to scale up to the vast amounts of content generated daily, there has been a significant shift toward technological solutions. Machine learning models have been at the forefront of this shift, offering scalable and efficient tools for automating the detection of false information. Leveraging the potential synergies between machine intelligence and human judgment, Wei et al. (2022) developed a framework that integrates machine learning algorithms with crowdsourced assessments to detect false news more effectively. Lee and Ram (2024) proposed a computational model that bolsters the ability of machine learning systems to discern between genuine and misleading claims by understanding claim-evidence relationships. Besides content-based approaches, researchers have explored account-based methods for misinformation detection. For example, Schoenmueller et al. (2024) utilized users' historical posting behaviors on social media platforms to predict their propensity to share misinformation, highlighting the potential of integrating user-specific contextual data in misinformation detection.

2.2. Countermeasures Against Misinformation on Social Media

Misinformation poses a significant threat to public discourse, prompting both platforms and government agencies to devise strategies to mitigate its spread. These strategies can be broadly categorized into pre-exposure and post-exposure countermeasures, aligning with the framework proposed by Lazer et al. (2018), which distinguishes between interventions that prevent users' initial exposure to misinformation and those that empower individuals to evaluate content they have encountered.

Pre-exposure countermeasures aim to curb the diffusion of misinformation before it reaches users. These proactive strategies include policy changes to reduce the visibility of misleading content and

interface nudges to redirect users toward credible information. For instance, Chiou and Tucker (2018) investigated Facebook's ban on advertisements from fake news sites and found that it significantly reduced the spread of misleading articles compared to Twitter. Likewise, Hwang and Lee (2025) examined Twitter's interface-level nudges, demonstrating that promoting credible sources during searches on high-risk topics reduces misinformation diffusion.

Post-exposure countermeasures, by contrast, address misinformation after it has already circulated. These approaches aim to improve content evaluation and discourse quality through user engagement and community governance. Common mechanisms include flagging or reporting content (Jiménez Durán 2021), applying credibility labels to sources (Kim et al. 2019), and encouraging cognitive reflection through prompts or cues (Moravec et al. 2022). For example, Jiménez Durán (2021) conducted two field experiments leveraging Twitter's reporting tool, finding that flagged tweets were significantly more likely to be removed. Kim et al. (2019) analyzed three different rating mechanisms: expert ratings, user article ratings, and user source ratings, with the last approach proving to be particularly effective in mitigating the spread of fake news. Similarly, Moravec et al. (2022) showed that reflective prompts can reduce belief in misinformation, while Kim and Dennis (2019) found that encouraging users to consider content authorship increased skepticism toward misleading claims.

Crowdchecking, exemplified by Community Notes, represents a novel form of post-exposure intervention with several features that distinguish it from earlier approaches. First, unlike prior approaches that provide generalized *source*-level credibility signals (e.g., Kim et al. 2019), Community Notes operate at the *content* level. This distinction is particularly important in polarized environments, where perceptions of source credibility are highly subjective. Moreover, credible sources can still disseminate misleading content, and conversely, less credible sources may share factually accurate information. Therefore, by targeting the content itself rather than the source, Community Notes can provide more nuanced and contextually relevant corrections. Second, Community Notes are *community-driven* and *consensus-oriented*, relying on a scoring algorithm that surfaces only notes rated as helpful by ideologically diverse contributors, thereby prioritizing bipartisan agreement. Whereas top-down expert

interventions may carry perceptions of partisan bias, the crowdsourced vetting process enhances the system's perceived legitimacy and scalability. Third, Community Notes engage with misinformation in a *non-punitive* manner. Instead of completely removing posts or labeling them as harmful through platform-issued warnings, the system appends informative notes that provide corrective information while preserving the original content. This approach promotes accountability among content creators without relying on coercive moderation techniques.

Prior academic work on Community Notes has largely focused on note contributors. Borwankar et al. (2022) examined the impact of participating in the Community Notes program on contributors' posting patterns. Shan et al. (2022) found that while the identity anonymization policy did not increase the number of contributions, it did enhance the quality of fact-checking. Borwankar et al. (2024) reported that a privacy-related policy intervention improved both the frequency and quality of contributions. Shan and Qiu (2025) exploited a natural experiment on Community Notes and revealed that peer recognition reduces the likelihood of labeling misinformation as containing unverified claims, increases the use of politically charged language, and increases the number of trustworthy fact-checking contributions.

In contrast, limited attention has been paid to how Community Notes affect the authors of misinformation. Zhou et al. (2025) found that authors labeled by Community Notes may experience a temporary increase in audience engagement, but the effects dissipate over time. However, to our knowledge, no existing study has empirically examined authors' retraction behavior. This study addresses such a gap by investigating whether the display of a community note prompts tweet authors to voluntarily retract their posts. Our analysis shifts the focus from audience-level effects to author behavior, offering new insights into how community-driven, content-specific corrections may function as a scalable tool for mitigating misinformation at its source. In doing so, this work contributes to the literature by highlighting a novel mechanism of accountability and providing evidence for the effectiveness of crowdchecking as a behavioral nudge toward curbing misinformation.

3. Hypothesis Development

Crowdchecking provides a democratic approach to annotating misleading posts, which could influence the audience of these posts through group and individual factors, as outlined in the framework by Scheufele and Krause (2019). At the group level, homogeneous social networks significantly contribute to the spread of misinformation; repeated exposure to specific beliefs within these networks increases familiarity with false information, thereby boosting its perceived credibility (Lewandowsky et al. 2012). Introducing counter-narratives via crowdchecking can expose these homogeneous networks to corrective information, potentially disrupting echo chambers. On the individual level, Scheufele and Krause (2019) suggest that the "uninformed" can easily become "misinformed," as a lack of knowledge about information veracity increases susceptibility to misinformation. This aligns with the ignorance or inattention theory proposed by Pennycook and Rand (2019), which argues that misinformation spread often stems from an individual's failure in critical thinking rather than intentional deception. Therefore, by adding corrective content to misinformation, crowdchecking could prompt the audience to evaluate misleading posts critically and therefore reduce their influence.

Importantly, the consequences of crowdchecking are not limited to audience reactions—they may also influence the behavior of those who authored the flagged content (hereafter, "misinformation spreaders"). Publicly displayed notes highlight factual inaccuracies or misleading claims, thereby signaling to the broader audience that the content—and, by extension, its author—is untrustworthy. As a result, such public signals may significantly damage the perceived reputation of the author.

In the context of social media, reputation refers to the extent to which a user is recognized by others, reflecting their perceived social standing. Reputation theory (Eisenegger and Imhof 2008) posits that individuals evaluate others' reputations through various lenses, including competence-based functional reputation, which indicates perceived ability and expertise, and normative social reputation, which reflects perceived adherence to community norms and expectations (Kim et al. 2019). On social media, the assessment of both types of reputation is typically based on observable platform cues (Baccarella et al. 2018). For instance, metrics such as follower count, audience engagement,

and content reach reflect users' influence and content quality, thereby serving as proxies for their functional reputation. Platform-conferred status indicators (e.g., verification badges) signal formal recognition by the platform and are often interpreted as markers of normative social reputation.

A major reputational risk on social media arises from sharing problematic content (Baccarella et al. 2018). Indeed, experiments show that sharing misinformation hurts a spreader's reputation and reduces future audience attention, even when the misinformation aligns with the audience's pre-existing beliefs (Altay et al. 2022). While the problematic nature of content may not always be obvious, given that it is often subject to varying audience interpretations, the presence of a publicly displayed community note makes it explicit and highly visible. Such public notes could damage both dimensions of reputation for the misinformation spreader: they undermine functional reputation by implying the author's inability to distinguish misinformation, and they harm normative social reputation by suggesting that the author has violated social norms on the platform. When facing public scrutiny, failure to conform to social norms often leads to social sanctions, including embarrassment, public criticism, or reputational loss, thereby exerting significant social pressure on the misinformation spreader.

These reputational concerns and social pressures act as motivators of people's behavioral change (Panagopoulos 2010), prompting misinformation spreaders to take corrective actions. While several corrective options may be available, such as publishing follow-up debunking posts, content retraction stands out as the most definitive and effective approach. Retraction not only stops the spread of misinformation by removing the content but also limits exposure for all users, including those who might encounter it indirectly through quotes or replies (Arif et al. 2017). Moreover, retracting the flagged tweet removes the associated public note as well, thereby eliminating the signal of the author's connection to misinformation and mitigating the reputational threat it poses. As a result, misinformation spreaders could be motivated to retract the tweet as a reputation-preserving response.

Although we believe the proposed theoretical mechanisms are strong enough to trigger content retraction, whether they are effective in practice has to be tested empirically. Indeed, humans are

often resistant to correction, especially by peers. In our context, social media users might question the validity of crowdchecking, sometimes rightly so, and continue to hold their positions. Even if they realize the flaws in their content, they may still refuse to admit and retract, either because they interpret the correction as an attack on their values and belief system (Tang et al. 2024), or because they are blinded by their emotion and pride. Therefore, we propose the following hypothesis for empirical testing.

Hypothesis: *Displaying notes from crowdchecking increases the likelihood of voluntary content retraction.*

4. Research Context and Data

4.1. Community Notes

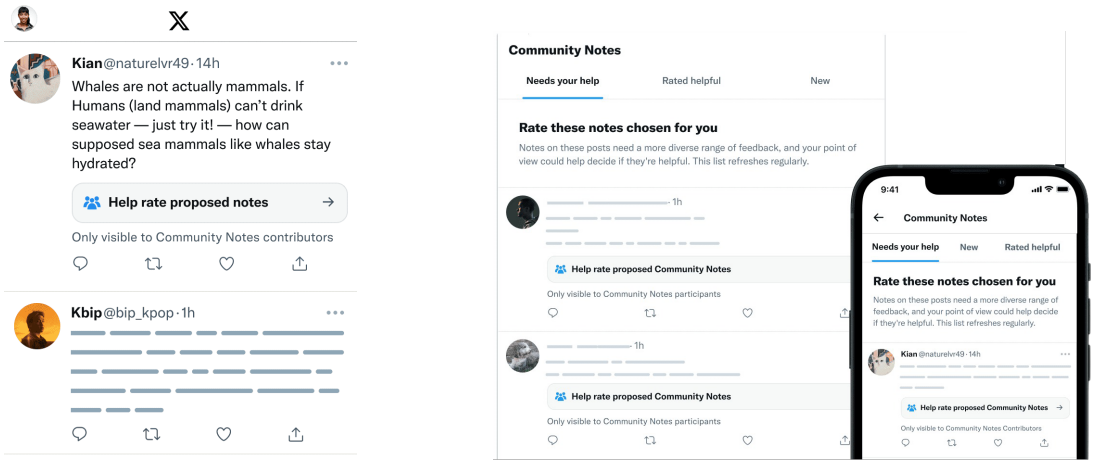
Our research context is Twitter's Community Notes, a crowdchecking system launched on January 25, 2021, formerly known as the Birdwatch program. Operating independently from Twitter, Community Notes employs a community-driven approach that enables users to collaboratively attach annotations or contextual information to tweets that are potentially misleading. To participate in adding and evaluating notes, users must enroll as contributors on the Community Notes platform. Initially, all notes are marked with a *Needs More Ratings* status on the platform and are hidden from general Twitter users. Only contributors can view these notes, equipped with rating prompts, directly on Twitter as depicted in Figure 1(a). They also have access to a dedicated rating system through the Community Notes interface, shown in Figure 1(b).

Twitter implements a matrix factorization-based ranking algorithm to process all notes and determine the status of a note (*Helpful*, *Not Helpful*, or *Needs More Ratings*) every day.[7] Specifically, for notes that receive at least five ratings, the algorithm generates an intercept term to represent the helpfulness score, which crucially influences the status of notes. According to the policy of Community Notes,[8] a note must achieve a helpfulness score of at least 0.4 to be deemed *Helpful*. However,

[7] The algorithm is publicly available at https://github.com/twitter/communitynotes

[8] For details, see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#note-status

Figure 1 Community Notes



(a) Note Visible Only to Contributors (b) Rating Interface of Community Notes

this criterion alone is not sufficient to acquire the *Helpful* status. To be classified as *Helpful*, a note must also receive broad support across diverse perspectives, assessed by another algorithm.[9] Meanwhile, an outlier filtering mechanism by Community Notes imposes additional requirements on note scores. In summary, having a score of at least 0.4 is necessary but not sufficient for a note to earn the *Helpful* status.

Once a note achieves the *Helpful* status, it will be displayed directly under the associated tweet and become visible to all Twitter users. Figure A.1 in Appendix A shows two tweets from Elon Musk with publicly displayed notes. After the public display of community notes, users who have engaged with the focal tweet—through replies, likes, or retweets—will receive a notification from Twitter.[10] However, passive viewers of the tweet, who only viewed the tweet without engaging, will not be notified. In addition, tweet authors are not notified when a community note is written for their post; they are only notified once the note becomes publicly visible.

The scoring algorithm updates the status of each note every day for two weeks following the note's creation. Once a note reaches two weeks of age, its status is locked and no longer changes, even though

[9] For details, see https://communitynotes.x.com/guide/under-the-hood/ranking-notes#matrix-factorization

[10] For details, see https://communitynotes.x.com/guide/contributing/notifications

new ratings may continue to be submitted. To ensure that changes in the *Helpful* status reflect a clear shift in consensus and avoid frequent status changes, Community Notes require that the note score drops below the applicable threshold by more than 0.01 before the note loses *Helpful* status. In other words, a note that has previously earned a *Helpful* status would have its score drop below 0.39 to lose its *Helpful* status.[11] A note rated as *Helpful* one day could potentially be reclassified as *Not Helpful* the following day if its score subsequently drops below 0.39 during the first two weeks, resulting in its removal from public display. Accordingly, a note's display status is time-varying during its initial two-week period, and remains fixed thereafter.

4.2. Data Collection

Twitter releases a public dataset[12] of Community Notes which is structured into four primary tables: notes, ratings, note status history, and user enrollment status. From this dataset, we compile a list of tweets, each associated with at least one community note, and monitor their status daily. To alleviate concerns about generalizability and enhance the external validity, we conduct data collection during two distinct periods: June 11 to August 2, 2024, and January 1 to February 28, 2025.[13] Specifically, we initiate our data collection by downloading the daily snapshots and conducting a day-to-day comparison to identify newly created notes. For each new note, we extract the associated tweet ID and monitor its status daily using an automated system we developed. The system captures comprehensive data for tweets, including textual content, engagement metrics, and author characteristics. It also records the presence and content of any notes displayed under the tweet. Table B.1 in Appendix B reports the summary statistics and definitions for the key variables used in the analysis.

Each tweet is tracked from the day its associated note first appears in the dataset until it is either retracted or reaches the end of our observation window (August 2, 2024, or February 28, 2025). We

[11] This rule also explains the three points to the immediate left of the threshold line at 0.4 in Figure 2.

[12] The dataset is released daily at https://x.com/i/communitynotes/download-data with tables provided as cumulative files in daily snapshots where each update incorporates all data from previous releases.

[13] The first period is before the 2024 US election while the second period is after the election. We thank the editors and reviewers for this suggestion.

combine the two datasets for our main analysis to increase the statistical power although we also conduct separate analyses using each dataset to increase external validity. Note that since Twitter releases notes and ratings with a 48-hour delay, it is possible for tweets to be retracted before our system starts collecting detailed information on these tweets and their authors. However, for these tweets, notes are not removed from the dataset, hence we can still compute their note scores.

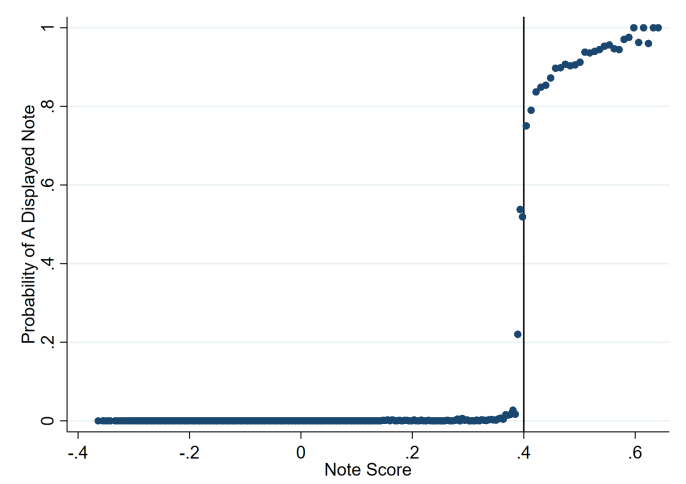
To compute the daily note scores, we implement Twitter's open-source ranking algorithms. Specifically, we process the daily snapshot data—including notes, ratings, note status history, and user enrollment status—using a Python script executed on a high-performance computing (HPC) cluster equipped with 8-way NVIDIA A100 GPUs and substantial memory capacity. This computationally intensive environment is necessary due to the large volume of daily data, which averages around 14 GB, and the demanding nature of the algorithm, which relies on matrix factorization and requires significant resources. The Python script runs across two compute nodes within the HPC cluster, each configured with 600 GB of RAM, one A100 GPU, and four CPU cores per task. Using this configuration, we can generate the daily note scores within around 10 hours, and this process is repeated for each day within our study window.

5. Identification

In this study, we employ a regression discontinuity design (RDD) to estimate the impact of publicly displaying community notes below a tweet on the likelihood of that tweet being retracted. This design helps identify the causal effect because it leverages the threshold mechanism of Community Notes, where a helpfulness score of at least 0.4 is required for a note to be categorized as *Helpful*. This scoring threshold introduces a locally exogenous variation in the probability of a note's public display, as slight variations in scores near the threshold can abruptly affect the display status. Figure 2 illustrates the relationship between the note scores and the likelihood of their public display. As indicated, there is a noticeable increase in the probability of a note being publicly displayed once its score surpasses the 0.4 threshold. As discussed, while achieving a score above the threshold is a necessary condition for a note's public display, it is not sufficient on its own. This subtlety is crucial as

it suggests the presence of other determinants influencing the final display status. Indeed, as Figure 2 shows, surpassing this threshold does not guarantee display (i.e., the probability equals 1). As such, instead of a sharp RDD, we adopt a fuzzy RDD, where the helpfulness score serves as the running variable. Since a fuzzy RDD does not rely on the characteristics of tweets and users, tweets that were

Figure 2 Discontinuity Plot on the Probability of a Displayed Note



Note. The bin size is 0.007 for note scores less than 0.4 and 0.014 for note scores larger than 0.4. Community Notes implements an inertia mechanism: a note's score needs to drop below the threshold by more than 0.01 before the note loses *Helpful* status which explains the three points to the immediate left of the threshold line at 0.4.

retracted within 48 hours of the first note creation are also included in the analysis.

Retract is the dependent variable indicating whether a tweet has been retracted by the end of the tracking process. *NoteScore*, as the running variable, is a note's helpfulness score calculated by Twitter's ranking algorithms.[14] It is a continuous variable ranging from -1 to 1, with 0.4 as the threshold. Notes with a score of at least 0.4 are much more likely to have a *Helpful* status (see Figure 2) than notes with scores less than 0.4. *NoteDisplayed* is a binary variable indicating whether a note has ever been rated as *Helpful* and publicly displayed under a tweet.

Although a note's status (*Helpful*, *Not Helpful*, or *Needs More Ratings*) can vary during the initial two-week period after its creation, and thus its display status could be time-varying, in our RD

[14] It is possible that a tweet receives multiple notes. For these tweets, we focus on the note with the highest score.

analysis, we take a cross-sectional approach and treat each note's status as static. Specifically, we capture whether a note was ever publicly displayed under the tweet during the observation window. If a note attained *Helpful* status at any point, we consider it as *Public*. The rationale is that even a brief period of public display of notes can exert meaningful influence on tweet authors. For instance, a note that is publicly displayed for even a single day can trigger notifications to all users who previously engaged with the tweet. Therefore, we code *NoteDisplayed* as 1 if the note was ever *Helpful*.[15]

5.1. RDD Specification

For a fuzzy RDD, a two-stage least squares (2SLS) estimation procedure is appropriate, as it operates within the standard instrumental variable framework (Hahn et al. 2001). In the first stage, we estimate the probability of a tweet having a publicly displayed note as a function of the note's helpfulness score and the threshold using the following model specification:

$$\begin{aligned} NoteDisplayed_i = &\alpha_0 + \pi AboveThreshold_i + \sum_{k=1}^{K} \rho_k (NoteScore_i - Threshold)^k \\ &+ AboveThreshold_i \sum_{k=1}^{K} \sigma_k (NoteScore_i - Threshold)^k + \eta_i \end{aligned} \tag{1}$$

where i denotes a tweet. $Threshold$ is a constant of 0.4. The dummy variable $AboveThreshold_i$ equals 1 if tweet i's highest-scored note has a helpfulness score above the threshold of 0.4 and 0 otherwise. We include polynomial functions of $(NoteScore_i - Threshold)$ up to an order of K. When $AboveThreshold_i$ is 0, the coefficient ρ_k represents the effect on the left side of the cutoff. When $AboveThreshold_i$ is 1, the coefficient $\rho_k + \sigma_k$ represents the effect on the right side of the cutoff.

[15] The majority (about 85%) of our sample consists of tweets with stable *Helpful* status. In addition, as we treat flipping-status notes as publicly displayed ($NoteDisplayed = 1$), and these cases ideally should exert a weaker influence than stably public notes, our estimation is conservative—i.e., the true effect may be stronger than our estimated effect. To further demonstrate that flipping-status notes do not compromise the validity of our results, we conducted a subsample analysis using tweets with stable note status only. The results remain consistent and are available upon request.

In the second stage, estimates of $NoteDisplayed_i$ from Equation (1), denoted as $\widehat{NoteDisplayed}_i$, are used to predict the probability of tweet retraction. We use the following model specification:

$$\begin{aligned} Retract_i = &\alpha_1 + \beta \widehat{NoteDisplayed}_i + \sum_{k=1}^{K} \gamma_k (NoteScore_i - Threshold)^k \\ &+ \widehat{NoteDisplayed}_i \sum_{k=1}^{K} \delta_k (NoteScore_i - Threshold)^k + \epsilon_i \end{aligned} \tag{2}$$

The coefficient of interest, β, captures the local average treatment effect of having a publicly displayed note on tweet retraction.

5.2. Identification Assumption

The main identification assumption for the validity of an RD design is that individual units cannot precisely manipulate their running variable to be above or below the threshold. The running variable (i.e., the helpfulness score in our context) is allowed to have arbitrary correlation with tweet or note characteristics, or the potential outcomes of retraction in treated or non-treated conditions. As Lee (2008) formally demonstrates, the key condition for valid identification is the existence of some randomness—factors outside the control of each individual unit—that determines the treatment assignment, which is often justified through the imprecise manipulation argument. In our context, we argue that individual units (i.e., misinformation spreaders or tweet authors) cannot precisely control whether the note score of their tweet falls above or below the threshold. Indeed, for Community Notes, precise self-selection or manipulation is nearly impossible in practice. According to Twitter,[16] "*Community Notes doesn't work like many engagement-based ranking systems, where popular content gains the most visibility and people can coordinate to mass upvote or downvote content they don't like or agree with. Instead, Community Notes uses a bridging algorithm—for a note to be shown on a tweet, it needs to be found helpful by people who have tended to disagree in their past ratings,*" thereby "*reducing the potential attempts at coordinated manipulation.*" Therefore, it is rather difficult

[16] For details, see https://communitynotes.x.com/guide/en/about/challenges#preventing-coordinated-manipulation-attempts

for tweet authors to even manipulate their note scores relative to the threshold, making RDD a valid identification strategy in our context.

With the technical assumption of continuous distribution of the running variable conditional on unobservable type, the above identification assumption implies the continuity of the distribution of the running variable around the threshold value. Therefore, we test whether the running variable (i.e., *NoteScore* in our context) is distributed without a discontinuity around the threshold. In particular, we use the McCrary density test (McCrary 2008), implemented by the Stata package *rddensity* (Cattaneo et al. 2018), to formally check for discontinuity. The t-statistic for discontinuity is -0.2520 with a p-value of 0.8010, indicating that the null hypothesis of no sorting or manipulation cannot be rejected at standard statistical confidence levels. Thus, the results of the McCrary density test suggest that there is no evidence of precise manipulation of note scores around the threshold, supporting the validity of the fuzzy RDD in this context.

Another consequence of the RD identification assumption (and the technical assumption of distribution continuity) is that pre-determined features should be continuous around the threshold as well. To empirically evaluate this, we conduct a series of covariate continuity checks, examining whether tweet-level engagement metrics (views, likes, replies, and shares) and user-level characteristics (number of followers, followings, and total tweets posted) change discontinuously at the 0.4 note score threshold. These covariates reflect baseline tweet and user attributes that are predetermined relative to treatment assignment and, as such, should not be influenced by it. Therefore, while they may influence retraction behavior, they should remain continuous at the cutoff if the RD design is valid. The results of these checks are presented in Figure C.1 of Appendix C. Across all examined covariates, we find no statistically significant evidence of discontinuities at the threshold. These findings support the plausibility of the continuity assumption and reinforce the credibility of our identification strategy based on the fuzzy RDD framework.

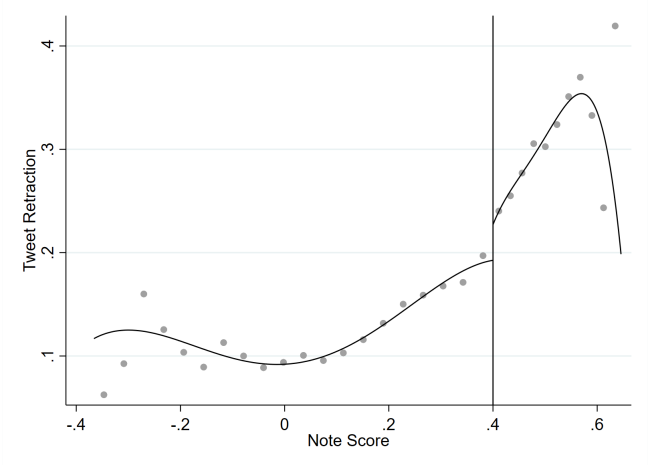
6. Empirical Results

6.1. Main Results

We begin our analysis by visually examining the relationship between note helpfulness scores (*NoteScore*) and tweet retraction rates (*Retract*) using the Stata package *rdplot*. Figure 3 presents

an RD plot of retraction rates against note scores, offering a visual demonstration of the effect. A fourth-order polynomial is fitted to the full dataset, with note scores divided into bins of width 0.038 on the left and 0.022 on the right of the cutoff. Tweets with note scores falling within the same bin are grouped, and the average retraction rate is calculated for each bin. The plot reveals a strong discontinuity in retraction rates at the note score threshold of 0.4. Specifically, in the local neighborhood of the threshold, the retraction rate increases significantly as the note score moves from below 0.4 to above, thereby supporting the hypothesis.

Figure 3 Graphical Discontinuity Analysis



We now present our fuzzy RDD findings using the 2SLS estimation approach. The Stata package *rdrobust* is used to estimate the effects. To demonstrate the robustness of the findings, we apply two different polynomial orders: linear ($K = 1$) and quadratic ($K = 2$). For each polynomial order, we use the optimal bandwidth separately on each side of the cutoff, due to the observed slope differences in Figure 3.[17] Table 1 reports the first-stage and second-stage results. As shown, the coefficient of interest, $NoteDisplayed$, is positive and statistically significant across all polynomial orders, indicating

[17] This approach follows practice in the regression discontinuity literature. Calonico et al. (2017) demonstrate that optimal bandwidths often differ substantially across the cutoff in empirical applications, while Arai and Ichimura (2016, 2018) show that using asymmetric bandwidths reduces asymptotic mean squared error when slopes differ across the threshold.

that having a displayed note significantly increases the likelihood of tweet retraction. Therefore, the hypothesis is supported. In terms of magnitude, the estimate in Column 1 suggests that tweets with note scores just above the 0.4 threshold are 32% more likely to be retracted compared to those with scores just below the threshold.

A potential concern of the observed effect is its dependence on contextual factors. In particular, the mid-2024 observation window coincided with the U.S. presidential election, a period during which the effectiveness of crowdchecking as a misinformation countermeasure may differ due to election-driven dynamics. To alleviate concerns related to temporal variation, including election-related factors, platform policy changes, or user composition shifts over time, we replicate the fuzzy RD analysis separately for the 2024 and 2025 datasets to assess whether the treatment effects persist across different time frames. Both the 2024 and 2025 analyses follow the same empirical procedures as the main analysis. As reported in Appendix D, the results are qualitatively consistent across years. We observe comparable discontinuities in note display probability and tweet retraction rates at the 0.4 threshold (see Figures D.1 and D.2). The 2SLS estimates further suggest that the effect of note display on tweet retraction remains positive and statistically significant in both subsamples (see Tables D.1 and D.2). These replications reinforce the robustness and generalizability of our findings.

6.2. Misinformation Egregiousness of Tweets Flagged by Community Notes

While we observe a positive effect of public note display on tweet retraction, an important remaining question is whether tweets flagged in Community Notes actually contain misinformation. This distinction matters because if a substantial portion of crowdchecked tweets do not involve misinformation, the observed retractions may simply reflect users responding to public pressure rather than taking corrective actions. In such cases, crowdchecking functions less as a misinformation countermeasure and more as a mechanism for exposing content to public targeting and scrutiny. To investigate this issue, we analyzed the distribution of misinformation in two datasets—one in retracted tweets, to assess whether retraction reliably signals misinformation and thus retraction reflects corrective behavior, and another in all tweets submitted to Community Notes, to evaluate the overall prevalence of misinformation and the extent to which crowdchecking targets it.

To evaluate the presence of misinformation in retracted tweets, we conduct a manual annotation of 500 randomly sampled retracted tweets from our dataset. Each tweet is independently reviewed by a trained research assistant, who examines the corresponding Community Note and consults external fact-checking sources (e.g., professional fact-checking websites) to determine whether the content was misleading or false. This evaluation reveals that approximately 80% (399 out of 500) of the retracted tweets are classified as misinformation, suggesting that tweet retraction is strongly associated with misinformation and thereby supporting its use as a proxy in our analysis. Representative examples of misleading tweets, along with the corresponding community notes explaining why they are deemed misleading, are presented in Table E.1 of Appendix E.[18]

In addition to the manual check of misinformation in retracted tweets, we construct a misinformation egregiousness classifier to automatically assess the factual accuracy of all English tweets in our sample. To ensure the classifier aligns with professional standards, we validate it against accuracy ratings from PolitiFact, a widely recognized professional fact-checking organization.[19] PolitiFact evaluates statements submitted by users or surfaced through media monitoring, rating them on a six-point Truth-O-Meter scale: True, Mostly True, Half True, Mostly False, False, and Pants on Fire.[20] We collect the latest 900 fact-checks published by PolitiFact in 2024, along with their corresponding ratings, to serve as our ground truth dataset. Statements rated as Half True, Mostly False, False, or Pants on Fire are coded as high misinformation egregiousness, while those rated as True or Mostly True are coded as low misinformation egregiousness. We then prompt the GPT-4.1 model to assess

[18] The remaining 20% (101 tweets), despite not being categorized as misinformation, are not benign or harmless in nature. Many of these tweets are associated with other forms of problematic content. Table E.2 of Appendix E lists several example tweets. For instance, some tweets appear to be designed primarily for engagement farming (see row 2), while others violate platform rules by promoting implicit advertisements (see rows 3 and 5) or targeting other users through personal attacks (see row 1).

[19] For details, see https://www.politifact.com/

[20] For details, see https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

the factual accuracy of each statement and categorize it as either high or low misinformation egregiousness. The classifier achieved an overall accuracy of 85%, indicating strong alignment with expert evaluations. We subsequently apply this classifier to all English tweets in our sample. Of the 114,666 English tweets analyzed, 95,406 (83.2%) are classified as high misinformation egregiousness, while 19,260 (16.8%) are classified as low misinformation egregiousness. This distribution closely mirrors the results of our manual labels in retracted tweets, suggesting that the majority of tweets contained misinformation in Community Notes.

6.3. Mechanism Tests

As discussed in Section 3, public notes can cause reputational concerns and exert social pressure, thereby triggering retraction. As a result, their effectiveness is expected to vary with the reputational status of the misinformation spreader. Users with high functional reputation, such as those whose tweets attract substantial visibility and engagement or who have large follower bases, face heightened reputational risks due to their social influence. Likewise, users with elevated normative social reputation, indicated by platform-conferred recognition (e.g., verification badges), tend to be more sensitive to norm violations. Accordingly, these users are more likely to retract flagged content as a strategy to mitigate reputational damage.

To test the mechanism, we conduct a series of subsample fuzzy RD analyses. We first split the sample at the median level of engagement, measured by the sum of likes, replies, and shares, and run the 2SLS estimation separately for each subgroup. As shown in Table 2, tweets with higher engagement show a significantly larger response to Community Notes $(0.1028, p < 0.01)$ compared to those with lower engagement $(0.0320, p < 0.1)$. A similar pattern emerges for view count (Table 3). Using the median number of views as the cutoff, we find that the treatment effect is substantially larger for tweets with high view counts $(0.1203, p < 0.01)$ compared to those with low views $(0.0211, p > 0.1)$. These findings suggest that when tweets attract greater public attention, the reputational stakes and perceived social pressure on misinformation spreaders intensify due to their amplified social influence. As a result, users may feel a heightened sense of accountability for misinformation dissemination, making them more likely to retract the flagged content. We also split the sample based on

the median number of followers. Table 4 shows that users with more followers are more responsive to Community Notes ($0.0773, p < 0.01$) than those with fewer followers ($0.0341, p > 0.1$). This result is consistent with the idea that audience size magnifies reputational stakes, increasing pressure to remove misinformation when publicly challenged.

In addition, we examine the role of user verification status in shaping the effect. Verified users on Twitter are marked with a blue checkmark, indicating that the account meets eligibility criteria, including the absence of misleading or deceptive behavior,[21] and thus serves as a platform-conferred signal of normative social reputation. We conduct subsample RD analyses for verified and unverified users. The results in Table 5 show that publicly displayed notes have a significant and positive effect on tweet retraction for verified users ($0.0733, p < 0.01$), while the effect is statistically insignificant for unverified users ($-0.0139, p > 0.1$). This finding suggests that, for verified users, flagged tweets present a more salient reputational risk due to their greater concern with maintaining credibility on the platform, thereby increasing the likelihood of tweet retraction as a harm-mitigation strategy. We further examine the interaction between user verification and tweet visibility to assess how concerns for normative social reputation and social pressure stemming from public attention jointly shape retraction behavior. Specifically, we conduct a subsample fuzzy RD analysis by verification status and engagement level as shown in Table F.1, and by verification status and views as shown in Table F.2 in Appendix F. The results consistently show that verified users are more likely to retract tweets when the tweet receives higher visibility—indicated by more views and engagements, suggesting that reputational concerns and social pressure jointly drive and amplify users' likelihood of removing misinformation.

6.4. Heterogeneous Effects of Community Notes

We examine factors that influence users' reputational concerns and exposure to social pressure, which may explain variation in their responses to community notes. First, we examine whether users have enabled direct messaging from the public. Table 6 reports the results of the subsample RD analysis.

[21] For details, see https://help.x.com/en/managing-your-account/about-x-verified-accounts

Among users who have enabled DMs, we find a significant effect on retraction following the display of a community note ($0.1060, p < 0.01$). By contrast, the effect is not significant for those who have disabled DMs ($-0.0028, p > 0.1$). Enabling DMs creates a channel for audiences to directly express disapproval and exert influence on the misinformation spreader, prompting them to retract tweets more readily to avoid further critiques or conflict.

Second, we explore the role of tweet characteristics, including misinformation egregiousness and tweet tenure. We split our data into tweets with high and low misinformation egregiousness and conducted separate fuzzy RD analyses. Table 7 reports the results. Among tweets containing highly egregious misinformation, *NoteDisplayed* has a positive and statistically significant effect on tweet retraction ($0.0580, p < 0.1$), suggesting that misinformative content is more responsive to public correction. In contrast, we find no statistically significant effect among tweets with low levels of egregious misinformation ($-0.0204, p > 0.1$), suggesting that authors of publicly noted tweets containing contentious but not necessarily misleading content are not susceptible to public note. This evidence indicates that more egregious content may pose a greater threat to the user's reputation, leading to the increased likelihood of tweet retraction. In addition, we assess whether the tweet is relatively new or old, using the median of tweet tenure as the cutoff. As shown in Table 8, the retraction effect is large and marginally significant for newer tweets ($0.1902, p < 0.1$), but effectively zero and not significant for older tweets ($-0.0059, p > 0.1$). New tweets are more likely to still be circulating on users' timelines when a community note is added, making their visibility and reputational damage more immediate. In contrast, older tweets may no longer be prominent, leading to less perceived need for the author to take corrective actions.

Finally, we examine user tenure as a proxy for the user's established status on the platform. Table 9 reports the results of the subsample RD analysis using the median of user tenure as the cutoff. We find that new users are significantly more likely to retract tweets after a community note is displayed ($0.1546, p < 0.01$), while long-tenured users show no significant response ($-0.0049, p > 0.1$). This pattern suggests that users interpret and act on reputational signals differently depending on

how established they are on the platform. For newer users, each public signal—such as being publicly noted—carries more weight in shaping how others perceive them because their online identity is still forming. Hence, they may be more sensitive to reputational cues and more motivated to correct actions that could undermine their credibility. In contrast, long-tenured users have already established a stable presence. A single note may not substantially shift how they believe others view them, much like the decreasing weight of each new signal in Bayesian updating. As a result, they are less responsive to such signals.

7. Additional Analysis

7.1. Do Community Notes Accelerate Tweet Retractions?

The empirical findings so far strongly support our hypothesis that publicly displayed community notes can trigger the retraction of misinformation on social media. A different but also important question is whether publicly displayed community notes can accelerate the retraction process, conditional on retraction. The promptness with which misinformation is curbed is crucial in diminishing its potential negative impact. If community notes can accelerate the retraction process, they could serve as a powerful tool not only in correcting individual instances of misinformation but also in enhancing the platform's overall responsiveness to emerging falsehoods. To shed light on this, we conduct a survival analysis to investigate whether publicly displayed community notes accelerate tweet retraction. To avoid data censoring issues, we exclude tweets posted well before the observation window. From our dataset, we identify 17,533 tweets, 2,431 of which have a community note displayed beneath them. The average lifecycle of these 17,533 tweets is approximately 32 days. We construct panel data at the tweet-day level to analyze the dynamics of tweet retractions with and without community notes.

Previous literature has suggested that discrete-time survival analysis is more appropriate when data is interval-censored (Singer and Willett 1993). This method has also been applied in information systems (IS) studies (see e.g., Gao et al. 2022). Two discrete-time survival analysis specifications have been widely used: logit hazard and complementary log-log hazard (*cloglog*). We employ both specifications to demonstrate the robustness of the findings. The following model is on the *cloglog* of the hazard or conditional probability of tweet retraction at time t given survival up to that time:

$$cloglog\lambda_{i,t} = \delta_t + \beta_0 + \beta_1 NoteDisplayed_{i,t} + \beta_2 X_{i,t} + \beta_3 Z_{i,t} \tag{3}$$

where $\delta_t = cloglog\lambda_{0,t}$ is the complementary log-log transformation of the baseline hazard. In the analysis, we specify the baseline hazard function as the logarithm of time and its square to check robustness. β_1 is the coefficient of interest. $X_{i,t}$ refers to the tweet-level characteristics, such as the number of views and received likes. $Z_{i,t}$ includes the user-level characteristics, such as their time-varying online profiles. Standard errors are clustered at the tweet level.

Table 10 reports the results of the discrete-time survival analysis. Columns 1 and 2 detail the results where the cloglog function is employed as the link function, while Columns 3 and 4 utilize the logit link function. For the baseline hazard function, Columns 1 and 3 apply the logarithm of time, aiming to capture a linear increase over time in the log-hazard rate. Conversely, Columns 2 and 4 utilize the square of time to assess the impact of a quadratic time trend on the hazard function, which helps explore nonlinear effects in the timing of tweet retractions. Across different empirical specifications, the coefficients of *NoteDisplayed* are consistently positive and statistically significant, suggesting that receiving a displayed community note indeed increases the hazard rate of being retracted and thus decreases the time to tweet retraction. These results confirm that publicly displayed community notes significantly accelerate the retraction process of misinformation, thereby enhancing the platform's ability to mitigate the spread of misinformation promptly.

7.2. Contemporariness of Tweets Flagged by Community Notes

To better understand the nature of tweets flagged by Community Notes, we examine whether these tweets are tied to contemporary events. Specifically, we develop a measure of *contemporariness* to assess whether each tweet addresses issues actively discussed by the public at the time it was flagged.

We construct this measure using large language models (LLMs) in two complementary ways. The first approach leverages the LLM's internal knowledge base. We prompt GPT-4.1 to assess whether a given tweet pertains to ongoing public debates, major news events, or widely discussed topics that attracted significant social media attention at the time the tweet was flagged. The second approach adopts a retrieval-augmented generation (RAG) framework to enrich the model's temporal context. We compile daily trending terms from Twitter's trending list[22] during our observation window and

[22] For details, see https://x.com/explore/tabs/trending

use them as proxies for active public discourse. For each tweet, we aggregate trending terms from the three days before, the day of, and the three days after its first appearance in the Community Notes dataset. These terms are then supplied as external context to GPT-4.1, which then determines whether the tweet topic aligns with the surrounding trending discussions. To compare the results, we apply both methods to a random sample of 400 English-language tweets. The classification outcomes match in 86% of cases, suggesting that LLMs, even without additional context, possess reasonably strong temporal awareness. Nonetheless, to ensure explicit incorporation of public discourse, we rely on the RAG-enhanced method for the analysis.

Applying the RAG-based classifier to all English tweets in the full dataset, we find that 55,979 tweets (48.8%) are related to contemporary topics, while 58,687 tweets (51.2%) are not. These findings suggest that although community notes frequently target content tied to ongoing events, they also address a wide spectrum of misinformation beyond immediate news cycles or trending debates. This broader coverage indicates that crowdchecking, as implemented through Community Notes, is not limited to the correction of misinformation on contemporary issues but also addresses general forms of misinformation. As such, our findings may have broader applicability beyond the context of trending or time-sensitive content.

8. Conclusion

This study leverages Twitter's Community Notes to empirically assess the effectiveness of crowdchecking in mitigating the spread of misinformation. Utilizing the fuzzy RDD, we find that receiving a public community note increases the likelihood of tweet retraction. Our mechanism tests reveal that this positive effect mainly stems from reputational concerns and perceived social pressure. Additionally, we identify heterogeneous effects of community notes across various tweet and user characteristics, emphasizing the need to consider these contextual factors to fully harness the power of crowdchecking. Results from discrete-time survival analyses indicate that displaying community notes not only increases the probability of misinformation retraction but also accelerates the retraction speed among retracted, thereby enhancing social media platforms' responsiveness to emerging misinformation. Lastly, our analysis suggests that Community Notes not only touches on contemporary topics but also engages in misinformation of a more general or persistent nature.

8.1. Contributions to Literature

This study contributes to the literature on misinformation, particularly concerning countermeasures against misinformation in the following ways. First, this study introduces a new angle to the literature by examining the effectiveness of crowdchecking, a post-exposure countermeasure focusing on *producers* rather than *consumers* of misinformation. Previous studies have predominantly focused on how audiences respond to misinformation corrections, often revealing limited effectiveness in changing beliefs or behaviors. For instance, Lewandowsky et al. (2012) found that misinformation can have continued and long-lasting effects on people's cognitive processes even after being debunked. The limitations of audience-based approaches highlight the grave challenge of countering misinformation and suggest the urgent need for more approaches. By studying the retraction of misleading content, our work provides valuable insight into an alternative approach that tackles the problem of misinformation at its source.

Second, our study demonstrates the potential of voluntary retraction from misinformation spreaders as an effective alternative to forcible removal of misinformation. Previous research has identified several challenges with platform-driven removal strategies. For example, Broniatowski et al. (2023) found that Facebook's removal of antivaccine misinformation led to remaining content becoming more misinformative and politically polarized. Furthermore, content removal faces the challenge of balancing misinformation regulation with free speech principles (Kozyreva et al. 2023). Given these challenges, our focus on voluntary retraction by misinformation spreaders offers a potentially more sustainable approach to tackling misinformation and therefore complements the existing literature on content moderation.

Third, our study uncovers the mechanisms through which crowdchecking influences misinformation spreaders to voluntarily retract their content. While the existing literature (see e.g., Arif et al. 2017) provides only limited and largely descriptive insights into people's retraction behavior, our study empirically demonstrates the role of reputational concerns and social pressures. Specifically, we find that higher reputational stakes prompt misinformation spreaders to take corrective actions

as their reputation repair strategy. This empirical validation deepens our understanding of how countermeasures like crowdchecking can effectively lead to voluntary misinformation retraction.

Fourth, this study extends the growing IS literature on crowd-based misinformation interventions by deepening our understanding of crowdchecking as a distinct countermeasure. Unlike credibility rating systems that assess the *source* of content, Community Notes operates at the *post* level, enabling targeted, context-specific corrections. Furthermore, the algorithm behind Community Notes prioritizes cross-ideological agreement, surfacing only those notes deemed helpful by ideologically diverse contributors. This consensus-driven approach differentiates it from both expert-led and purely user-generated corrections, which may reflect homogenous viewpoints or top-down decision-making. Prior work on Community Notes has focused primarily on contributors—examining their motivations, quality of participation, or engagement patterns (e.g., Borwankar et al. 2022, Shan et al. 2022). In contrast, we focus on the recipients of corrections and provide the first large-scale empirical evidence that crowdchecking can influence misinformation authors' behavior. By doing so, we offer a more holistic view of how crowd-based interventions operate within social media ecosystems and contribute to misinformation governance.

8.2. Contributions to Practice

Insights from this study have important practical implications for various stakeholders. For individuals, especially those who may encounter misinformation on social media, it is important to participate in crowdchecking whenever they can and however they can, because it does make a difference. Much as the mandatory jury duty helps ensure everyone's right to a trial by jury in the United States, contributing one's knowledge to crowdchecking systems, whether by noting or rating, helps curb the spread of misinformation which eventually benefits everyone in a democratic society.

For social media platforms, the main takeaway from the current study is that crowdchecking is a viable approach to reduce the amount of misinformation on their platforms, and in most cases, is a less controversial and more scalable approach compared to the traditional approach of forcible content removal and professional fact-checking. We recommend all social media platforms to adopt a

crowdchecking system like Community Notes. To maintain effectiveness, it is crucial for platforms to ensure the transparency of these systems (e.g., by data sharing and open sourcing) so that user trust will not erode. Our research also offers more granular insights that could refine the implementation of crowdchecking systems. The mechanism test from our study indicates that reputational concerns and perceived social pressure drive the effectiveness of crowdchecking. Therefore, to enhance the effectiveness of crowdchecking, platforms should consider strategies that elevate reputational consequences and increase social pressure on misinformation spreaders. For instance, while Community Notes currently notify users who directly engaged with flagged posts, extending these notifications to include viewers and followers could not only curb the reach and spread of misinformation but also elevate reputational stakes, thereby potentially enhancing the effectiveness of crowdchecking. Additionally, our heterogeneity analysis shows that the efficacy of Community Notes diminishes for tweets that are older. This suggests a need for platforms to adapt the dynamics of their fact-checking systems, perhaps by focusing on newer content that tends to capture user attention more effectively due to the typical short attention spans on social media.

For policymakers, the effectiveness of crowdchecking serves as compelling evidence for supporting such initiatives. Given the significant societal harm posed by misinformation, particularly to younger generations, there is a critical need for regulatory support to reinforce these platform efforts. Policymakers can nudge or even mandate social media platforms to adopt transparent and effective crowdchecking systems and consider establishing standards for digital literacy that include understanding and participating in fact-checking processes. For instance, ongoing discussions in the US Senate regarding the role of major tech companies in curbing extremism and misinformation underscore the importance of legislative backing for platform accountability measures.[23] Such governmental actions not only reinforce the efforts to combat misinformation but also encourage platforms to assume more proactive roles in transparency and user education, ultimately leading to a more informed and discerning online community.

[23] For details, see https://www.congress.gov/event/117th-congress/house-event/111407

8.3. Limitations and Future Research

Like most empirical research, our study has limitations. First, the effectiveness of crowdchecking in curbing misinformation is evaluated only on a single social media platform (i.e., Twitter). Given the variety of social media platforms and the significant variations in their user bases, whether our findings generalize to other platforms, especially those in different countries, remains unclear. For example, major social media platforms such as YouTube, TikTok, and Facebook have recently started implementing crowdchecking systems similar to Community Notes to combat misinformation. Given the differences across platforms, it is important to investigate whether and to what extent crowdchecking is effective on other platforms.

Second, the internal validity of our empirical findings depends on the identification assumptions of the empirical analyses. Our research design relies on users' inability to precisely manipulate note scores around the threshold. While we are convinced by the validity of this assumption, we recognize the untestable nature of any causal assumption. We hope future research can find alternative identification strategies to further evaluate the insights revealed by the current study.

Third, while our findings suggest that receiving a public note increases tweet retraction, some tweet authors may respond through follow-up clarifications rather than retracting the original tweet. Although this behavior does not appear to be prevalent—otherwise, we would observe little or no treatment effect on retraction—it may occur among certain subgroups. Future research could examine this complementary strategy and explore the conditions under which users choose clarification over retraction, particularly as a function of their social standing, audience size, or content type.

Fourth, although leveraging data from two separate time periods (2024 and 2025) strengthens the external validity of our findings, the long-term impact of crowdchecking systems remains an open question. Over time, contextual factors—such as evolving user norms, shifting trust in platforms, or changes in perceived legitimacy of the notes—may influence how users respond to crowd-sourced corrections. Understanding how these dynamics unfold over a longer time horizon is important for evaluating the sustained effectiveness of crowdchecking in combating misinformation at scale.

Finally, although our study focuses on behavioral responses by content creators (i.e., tweet retractions), it does not directly assess belief change among the broader audience exposed to the misinformation. Crowdchecking systems, like other post-diffusion interventions, may face limitations in altering established beliefs. Future research could complement our analysis by investigating the perceptual and cognitive impacts of crowdchecking systems on audiences over time, including their ability to correct misbeliefs or, conversely, entrench them through motivated reasoning.

References

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2):211–236.

Allen MR, Desai N, Namazi A, Leas E, Dredze M, Smith DM, Ayers JW (2024) Characteristics of X (formerly Twitter) community notes addressing COVID-19 vaccine misinformation. *JAMA* 331(19):1670–1672.

Altay S, Hacquin AS, Mercier H (2022) Why do so few people share fake news? It hurts their reputation. *New Media & Society* 24(6):1303–1324.

Arai Y, Ichimura H (2016) Optimal bandwidth selection for the fuzzy regression discontinuity estimator. *Economics Letters* 141:103–106.

Arai Y, Ichimura H (2018) Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator. *Quantitative Economics* 9(1):441–482.

Arif A, Robinson JJ, Stanek SA, Fichet ES, Townsend P, Worku Z, Starbird K (2017) A closer look at the self-correcting crowd: Examining corrections in online rumors. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 155–168.

Baccarella CV, Wagner TF, Kietzmann JH, McCarthy IP (2018) Social media? It's serious! Understanding the dark side of social media. *European Management Journal* 36(4):431–438.

Borwankar S, Zheng J, Kannan K (2024) Unveiling the impact of privacy-preserving policies in crowd-based misinformation monitoring program. *Proceedings of the International Conference on Information Systems (ICIS)*.

Borwankar S, Zheng J, Kannan KN (2022) Democratization of misinformation monitoring: The impact of Twitter's Birdwatch program. *Available at SSRN 4236756* .

Broniatowski DA, Simons JR, Gu J, Jamison AM, Abroms LC (2023) The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. *Science Advances* 9(37):eadh2132.

Calonico S, Cattaneo MD, Farrell MH, Titiunik R (2017) rdrobust: Software for regression-discontinuity designs. *The Stata Journal* 17(2):372–404.

Cattaneo MD, Jansson M, Ma X (2018) Manipulation testing based on density discontinuity. *The Stata Journal* 18(1):234–261.

Chiou L, Tucker C (2018) Fake news and advertising on social media: A study of the anti-vaccination movement. Technical report, National Bureau of Economic Research.

Draws T, La Barbera D, Soprano M, Roitero K, Ceolin D, Checco A, Mizzaro S (2022) The effects of crowd worker biases in fact-checking tasks. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2114–2124.

Eisenegger M, Imhof K (2008) The true, the good and the beautiful: Reputation management in the media society. *Public relations research: European and international perspectives and innovations*, 125–146 (Springer).

Gao Y, Duan W, Rui H (2022) Does social media accelerate product recalls? Evidence from the pharmaceutical industry. *Information Systems Research* 33(3):954–977.

Garrett RK (2009) Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication* 14(2):265–285.

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on Twitter during the 2016 US presidential election. *Science* 363(6425):374–378.

Hahn J, Todd P, Van der Klaauw W (2001) Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69(1):201–209.

Hwang EH, Lee S (2025) A nudge to credible information as a countermeasure to misinformation: Evidence from Twitter. *Information Systems Research* 36(1):621–636.

Jiménez Durán R (2021) The economics of content moderation: Evidence from hate speech on Twitter. *Available at SSRN 4044098* .

Kim A, Dennis AR (2019) Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly* 43(3):1025–1039.

Kim A, Moravec PL, Dennis AR (2019) Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems* 36(3):931–968.

Kozyreva A, Herzog SM, Lewandowsky S, Hertwig R, Lorenz-Spreen P, Leiser M, Reifler J (2023) Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences* 120(7):e2210666120.

Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, et al. (2018) The science of fake news. *Science* 359(6380):1094–1096.

Lee DS (2008) Randomized experiments from non-random selection in US house elections. *Journal of Econometrics* 142(2):675–697.

Lee K, Ram S (2024) Explainable deep learning for false information identification: An argumentation theory approach. *Information Systems Research* 35(2):890–907.

Lewandowsky S, Ecker UK, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest* 13(3):106–131.

McCrary J (2008) Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142(2):698–714.

Moravec PL, Kim A, Dennis AR, Minas RK (2022) Do you really know if it's true? How asking users to rate stories affects belief in fake news on social media. *Information Systems Research* 33(3):887–907.

Mostagir M, Siderius J (2022) Learning in a post-truth world. *Management Science* 68(4):2860–2868.

Osmundsen M, Bor A, Vahlstrup PB, Bechmann A, Petersen MB (2021) Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review* 115(3):999–1015.

Panagopoulos C (2010) Affect, social pressure and prosocial motivation: Field experimental evidence of the mobilizing effects of pride, shame and publicizing voting behavior. *Political Behavior* 32:369–386.

Pennycook G, Rand DG (2019) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188:39–50.

Roozenbeek J, Schneider CR, Dryhurst S, Kerr J, Freeman AL, Recchia G, Van Der Bles AM, Van Der Linden S (2020) Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science* 7(10):201199.

Salge CAdL, Karahanna E, Thatcher JB (2022) Algorithmic processes of social alertness and social transmission: How bots disseminate information on Twitter. *MIS Quarterly* 46(1).

Scheufele DA, Krause NM (2019) Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116(16):7662–7669.

Schoenmueller V, Blanchard SJ, Johar GV (2024) Who shares fake news? Uncovering insights from social media users' post histories. *Journal of Marketing Research* 00222437241281873.

Shan G, Qiu L (2025) Examining unintended consequences of peer recognition on users' fact-checking contributions on social media: Evidence from a natural experiment. *Available at SSRN 5088565* .

Shan G, Wattal S, Thatcher JB (2022) How does anonymizing crowdsourced users' identity affect fact-checking on social media platforms? A regression discontinuity analysis. *ICIS 2022 Proceedings* 13.

Singer JD, Willett JB (1993) It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics* 18(2):155–195.

Talwar S, Dhir A, Kaur P, Zafar N, Alrasheedy M (2019) Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services* 51:72–82.

Tang H, Lenzini G, Greiff S, Rohles B, Sergeeva A (2024) "Who knows? Maybe it really works": Analysing users' perceptions of health misinformation on social media. *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 1499–1517.

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science* 359(6380):1146–1151.

Wei X, Zhang Z, Zhang M, Chen W, Dajun Zeng D (2022) Combining crowd and machine intelligence to detect false news on social media. *MIS Quarterly* 46(2).

Zhou Y, Hou J, Gao Y, Chen PY (2025) How does crowdsourced fact-checking approach tackling misinformation affect audience engagement? Evidence from Twitter's community notes program. *Proceedings of the 58th Hawaii International Conference on System Sciences*, 6549–6556.

Table 1 Regression Discontinuity Analysis

	Retract		*NoteDisplayed*	
	(1)	(2)	(3)	(4)
NoteDisplayed	0.3164***	0.1452***		
	(0.0707)	(0.0389)		
AboveThreshold			0.1531***	0.2651***
			(0.0122)	(0.0112)
Polynomial	Linear	Quadratic	Linear	Quadratic
Left Bandwidth	0.035	0.122	0.035	0.122
Right Bandwidth	0.043	0.075	0.043	0.075
Observations	21,806	52,677	21,806	52,677

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 2 Subsample Fuzzy RD Analysis by Engagement Level

	High		Low	
	Retract	*NoteDisplayed*	*Retract*	*NoteDisplayed*
	(1)	(2)	(3)	(4)
NoteDisplayed	0.1028***		0.0320*	
	(0.0382)		(0.0193)	
AboveThreshold		0.2271***		0.4416***
		(0.0170)		(0.0128)
Polynomial	Linear	Linear	Linear	Linear
Left Bandwidth	0.057	0.057	0.120	0.120
Right Bandwidth	0.049	0.049	0.047	0.047
Observations	11,393	11,393	18,631	18,631

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 3 Subsample Fuzzy RD Analysis by Number of Views

	High		Low	
	Retract	*NoteDisplayed*	*Retract*	*NoteDisplayed*
	(1)	(2)	(3)	(4)
NoteDisplayed	0.1203***		0.0211	
	(0.0393)		(0.0165)	
AboveThreshold		0.2220***		0.5387***
		(0.0163)		(0.0133)
Polynomial	Linear	Linear	Linear	Linear
Left Bandwidth	0.058	0.058	0.176	0.176
Right Bandwidth	0.047	0.047	0.042	0.042
Observations	12,612	12,612	24,885	24,885

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 4 Subsample Fuzzy RD Analysis by Number of Followers

	High		Low	
	Retract	*NoteDisplayed*	*Retract*	*NoteDisplayed*
	(1)	(2)	(3)	(4)
NoteDisplayed	0.0773***		0.0341	
	(0.0236)		(0.0291)	
AboveThreshold		0.3508***		0.3090***
		(0.0156)		(0.0147)
Polynomial	Linear	Linear	Linear	Linear
Left Bandwidth	0.092	0.092	0.078	0.078
Right Bandwidth	0.044	0.044	0.049	0.049
Observations	13,978	13,978	14,351	14,351

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 5 Subsample Fuzzy RD Analysis by Blue Checkmarks

	Yes		No	
	Retract	*NoteDisplayed*	*Retract*	*NoteDisplayed*
	(1)	(2)	(3)	(4)
NoteDisplayed	0.0733***		-0.0139	
	(0.0199)		(0.0409)	
AboveThreshold		0.3639***		0.3123***
		(0.0129)		(0.0204)
Polynomial	Linear	Linear	Linear	Linear
Left Bandwidth	0.101	0.101	0.078	0.078
Right Bandwidth	0.039	0.039	0.051	0.051
Observations	21,736	21,736	7,837	7,837

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 6 Subsample Fuzzy RD Analysis by Direct Message

	Enabled		Disabled	
	Retract	*NoteDisplayed*	*Retract*	*NoteDisplayed*
	(1)	(2)	(3)	(4)
NoteDisplayed	0.1060***		-0.0028	
	(0.0250)		(0.0225)	
AboveThreshold		0.3234***		0.3781***
		(0.0135)		(0.0163)
Polynomial	Linear	Linear	Linear	Linear
Left Bandwidth	0.077	0.077	0.115	0.115
Right Bandwidth	0.061	0.061	0.041	0.041
Observations	16,595	16,595	15,177	15,177

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 7 Subsample Fuzzy RD Analysis by Misinformation Egregiousness

	High		Low	
	Retract	*NoteDisplayed*	*Retract*	*NoteDisplayed*
	(1)	(2)	(3)	(4)
NoteDisplayed	0.0580*		-0.0204	
	(0.0311)		(0.0491)	
AboveThreshold		0.3287***		0.5034***
		(0.0174)		(0.0333)
Polynomial	Linear	Linear	Linear	Linear
Left Bandwidth	0.087	0.087	0.169	0.169
Right Bandwidth	0.050	0.050	0.049	0.049
Observations	11,144	11,144	4,425	4,425

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 8 Subsample Fuzzy RD Analysis by Tweet Tenure

	Old		New	
	Retract	*NoteDisplayed*	*Retract*	*NoteDisplayed*
	(1)	(2)	(3)	(4)
NoteDisplayed	-0.0059		0.1902*	
	(0.0196)		(0.1051)	
AboveThreshold		0.2794***		0.0904***
		(0.0180)		(0.0140)
Polynomial	Linear	Linear	Linear	Linear
Left Bandwidth	0.134	0.134	0.047	0.047
Right Bandwidth	0.066	0.066	0.040	0.040
Observations	20,069	20,069	12,220	12,220

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 9 Subsample Fuzzy RD Analysis by User Tenure

	Old		New	
	Retract	*NoteDisplayed*	*Retract*	*NoteDisplayed*
	(1)	(2)	(3)	(4)
NoteDisplayed	-0.0049		0.1546***	
	(0.0195)		(0.0471)	
AboveThreshold		0.3177***		0.2214***
		(0.0165)		(0.0163)
Polynomial	Linear	Linear	Linear	Linear
Left Bandwidth	0.089	0.089	0.048	0.048
Right Bandwidth	0.045	0.045	0.047	0.047
Observations	13,179	13,179	11,353	11,353

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. "Bandwidth" reports the optimal bandwidth. "Observations" reports the number of observations in the optimal bandwidth sample.

Table 10 Discrete-time Survival Analysis

	Cloglog		*Logit*	
	(1)	(2)	(3)	(4)
NoteDispalyed	0.3080***	0.3062***	0.3124***	0.3114***
	(0.0241)	(0.0260)	(0.0245)	(0.0265)
Tweet Controls	Yes	Yes	Yes	Yes
User Controls	Yes	Yes	Yes	Yes
Baseline Hazard	lnt	t^2	lnt	t^2
Observations	514,904	514,904	514,904	514,904
Log-Likelihood	-73062.538	-72724.836	-73068.660	-72728.605

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors in parentheses are clustered at the tweet level.