Can Crowdchecking Curb Misinformation? Evidence from Community Notes

Huaxia Rui

Joint work with **Yang Gao@UIUC** and **Maggie Zhang@UVA**

Simon Business School, The University of Rochester

October 3, 2025

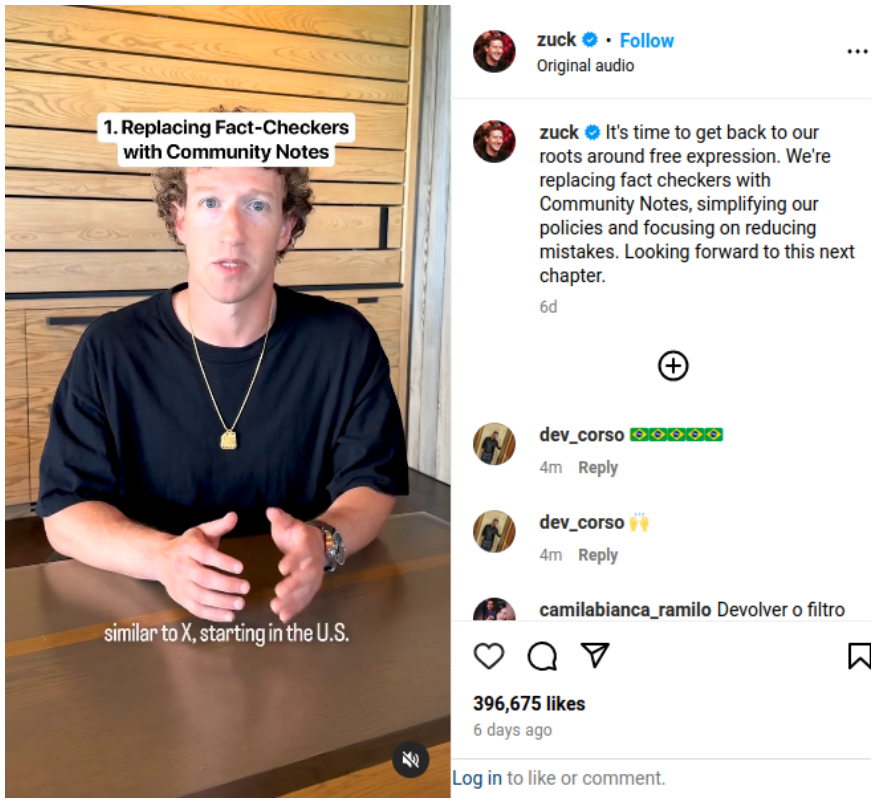
Misinformation ☞ ☜ Content Moderation

Misinformation refers to information that is false, inaccurate, or misleading but presented as accurate (Thorson 2016).

Content moderation (pre- or post-exposure)

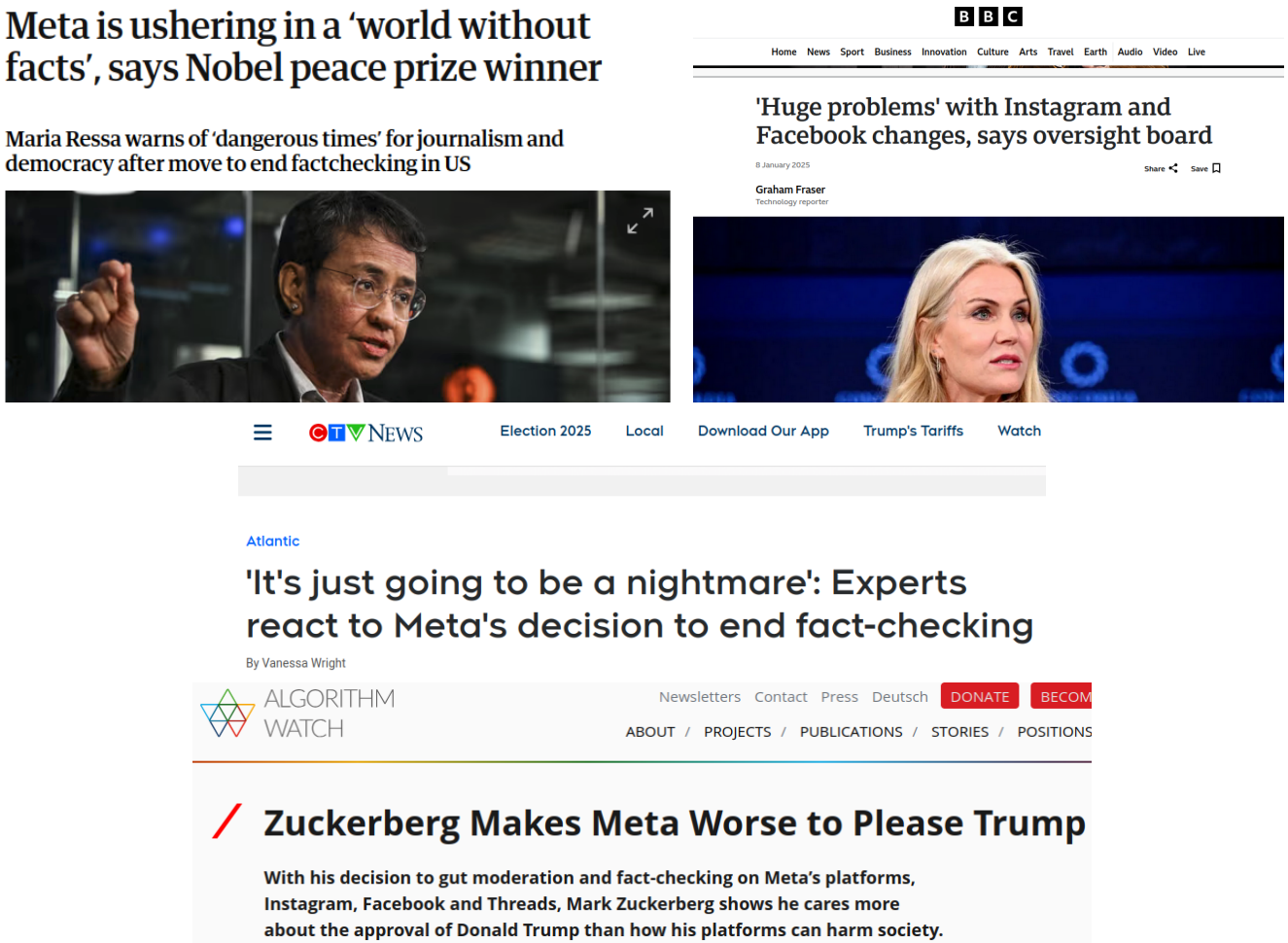
- speed
- arbiter?

Trust: Are fact-checkers impartial?



On Jan 7, 2025, Meta ended its independent 8-year-old fact-checking program which was launched in response to criticism over fake news during the 2016 US election.

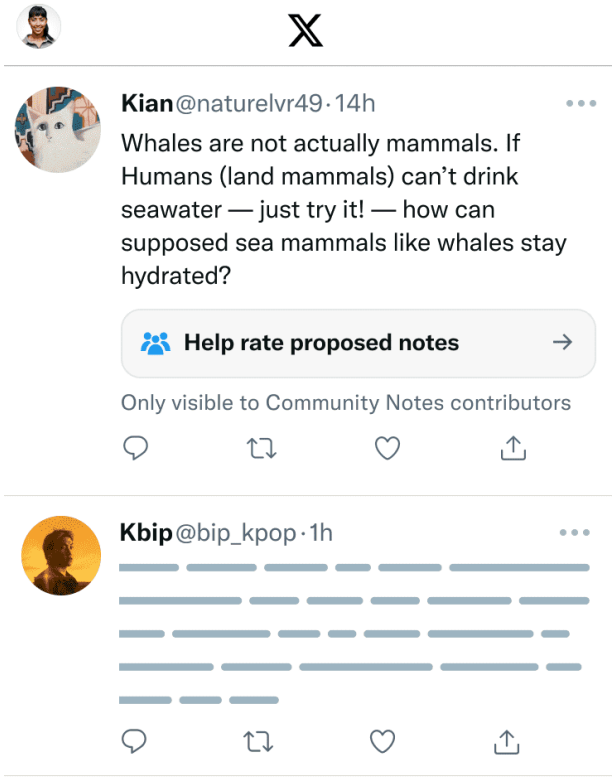
Fact-Checkers ⇝ Community Notes



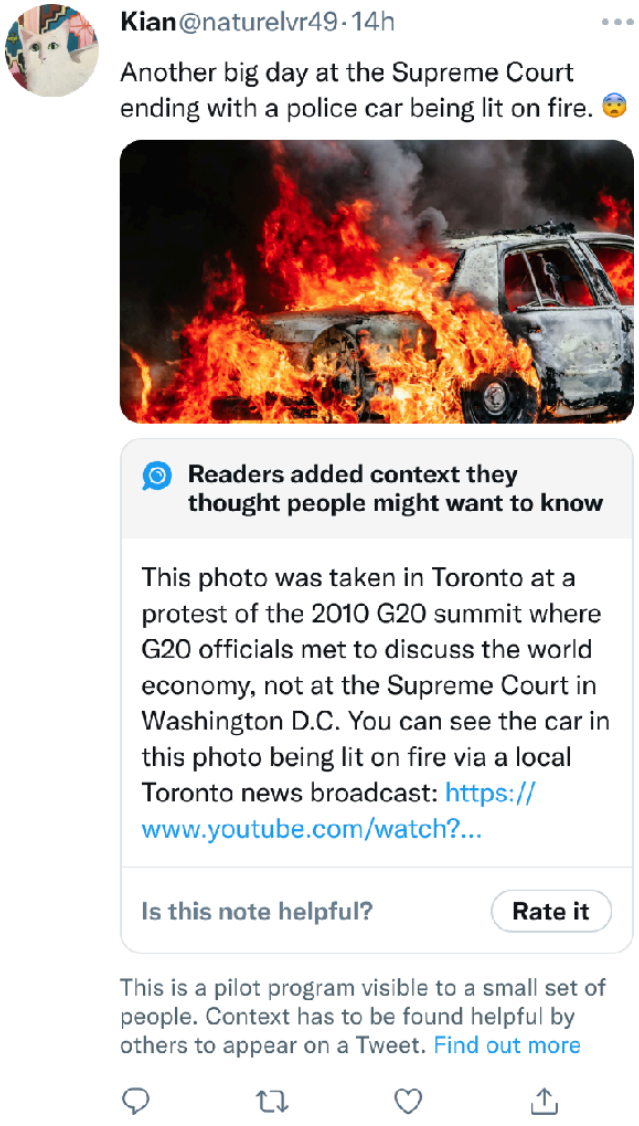
Twitter's Community Note was originally intended to complement, not replace, professional moderation (Yoel Roth, Twitter's head of Trust and Safety).

Note Status: To Display, or Not to Display?

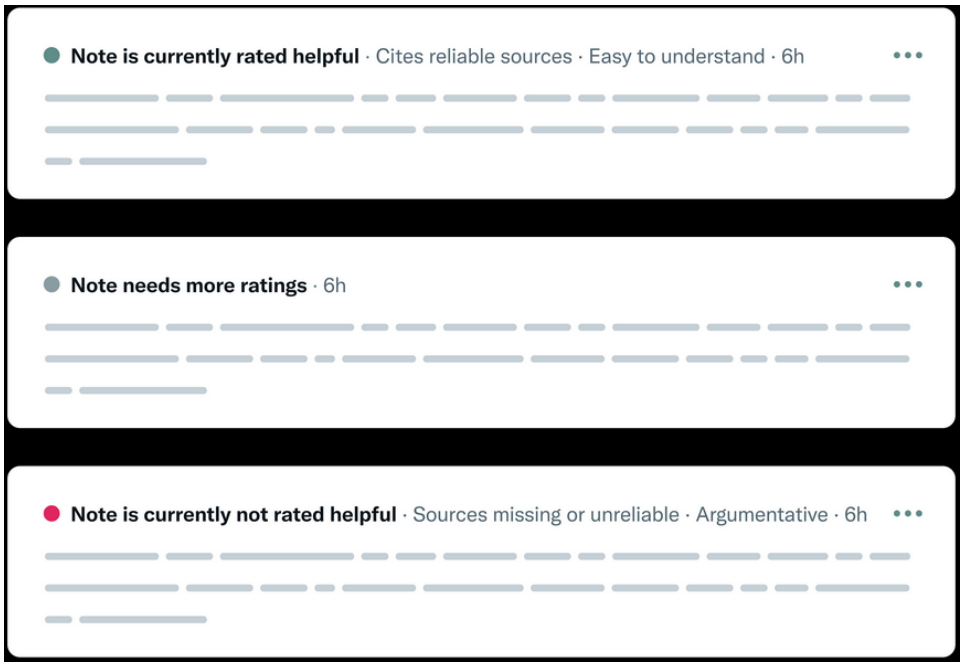
Community Notes, rolled out in 2021-2022, is a **crowdchecking** system where notes, containing corrective content, are submitted and rated by note contributors.



?



It All Boils Down to Rating



- All notes start with the **Needs More Ratings** status until they receive at least 5 total ratings.
- Notes with 5 or more ratings may be assigned a status of **Helpful** or **Not Helpful** according to an algorithm.

Literature

Dissemination and detection of misinformation on social media:

- Vosoughi et al. (2018), Mostagir & Siderius (2022), Allcott & Gentzkow (2017), ...

countermeasures against misinformation (audience-focused):

- Kim et al (2019) compared the effectiveness of different credibility rating (e.g., expert rating) on the spread of fake news.
- Moravec et al (2022) showed that reflective prompts can reduce belief in misinformation.

community notes (working papers, noter-focused):

- Borwankar et al. (2022) studied the impact of participating in the Community Notes program on note contributors' posting patterns.
- Shan et al. (2022) studied the effect of identity anonymization on the quantity and quality of notes.
- Shan & Qiu (2025) studied the role of peer recognition on various outcome variables.
- Zhou et al. (2025) found that authors labeled by Community Notes experience a temporary increase in audience engagement.

Research Question

RQ: Can publicly displaying community notes facilitate voluntary retraction?

Why study this RQ?

If it is valid, it cannot be retracted, any more than the dead can be brought to life. — Lincoln (1863)

Misinformation on social media causes real damage.

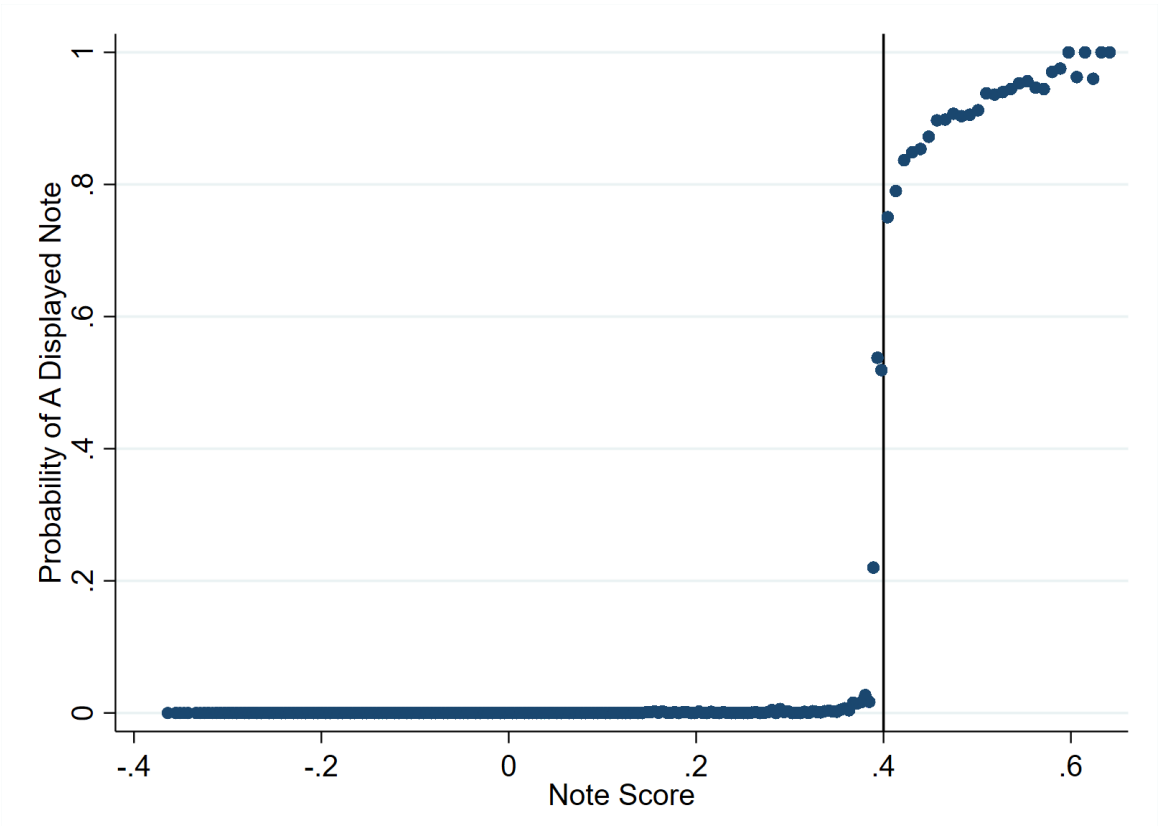
- Forcible content removal: censorship, polarization
- Voluntary retraction: a more civilized approach

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

- $-$: People are not (always) receptive to criticism, especially by peers lacking credential and their doubts about the validity of the criticisms, not to mention many other reasons (e.g., belief heterogeneity, pride).
- $+$: Displaying potentially corrective content alongside the original content triggers more critical thinking by the audience, causing **reputation concern** of the original content creator.

Regression Discontinuity (RD)

A note score (i.e., i_n) of at least 0.4 is required for a note to be categorized as **Helpful** which then allows a note to be displayed.



The bin size is 0.007 for note scores less than 0.4 and is 0.014 for note scores larger than 0.4.

Inertia mechanism: a note's score needs to drop below the threshold by more than 0.01 before the note loses **Helpful** status.

RD Assumption: Partial Manipulation (Lee 2008)

Let U be potential confounder (i.e., type) associated with each note (e.g., tweet, author, note, noter, raters).

- $X = \chi(U)$: observable running variable
- $D = \mathbf{1}_{X \geq 0.4}$: treatment status
- $Y_d = \phi_d(U)$: potential outcome for treatment $d \in \{0, 1\}$.

Let $F(x|u)$ be the conditional CDF of X given $U = u$.

$0 < F(0.4|u) < 1$ and $F(x|u)$ is continuously differentiable in x at $x = 0.4$, $\forall u$.

- Since the distribution $F(x|u)$ depends on u in a general way, individuals are allowed to manipulate the treatment probability, but not precisely.
- The probability of obtaining X just below and just above 0.4 are the same for each type (i.e., RCT around the threshold).

$$\lim_{x \to 0.4+} \mathbb{E}[Y|X = x] - \lim_{x \to 0.4-} \mathbb{E}[Y|X = x] = \mathbb{E}[Y_1 - Y_0|X = 0.4]$$

$$= \int (\phi_1(u) - \phi_0(u)) \frac{f(0.4|u)}{f(0.4)} G(du),$$

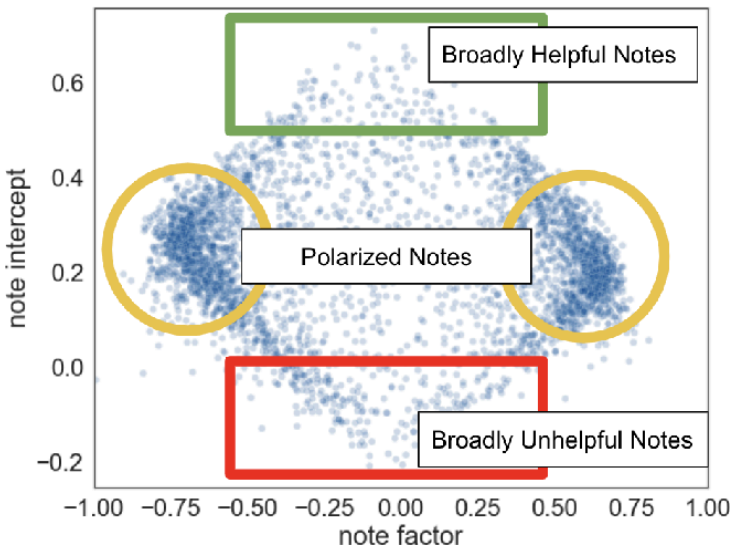
Bridging-Based Ranking

The matrix factorization (MF) algorithm predicts each rating as

$$\hat{r}_{un} = \mu + i_u + i_n + f_u \cdot f_n$$

where i_n, the note's intercept term, captures its helpfulness beyond users' viewpoints and leniency, and f_n is a 1-dimensional note factor. Model parameters are estimated using observed ratings $r_{un} \in \{0, 1, \text{null}\}$ by

$$\min_{i,f,u} \sum_{r_{un}} (r_{un} - \hat{r}_{un})^2 + \lambda_i(i_u^2 + i_n^2 + \mu^2) + \lambda_f(\|f_u\|^2 + \|f_n\|^2)$$

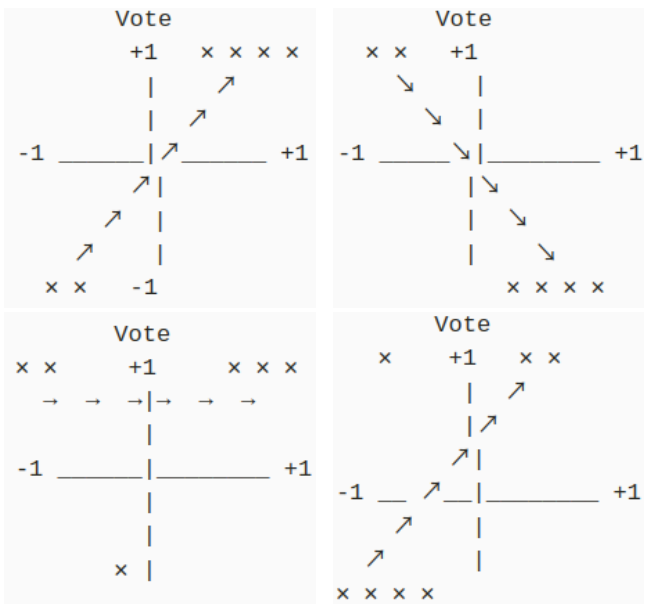


1. Filter out raters with < 10 ratings or notes with < 5 raters.
2. Fit MF to compute provisional scores.
3. Compute user helpfulness scores and filter out all ratings from users that didn't meet a criteria.
4. Fit MF to compute the final score and label.

Can i_n be Precisely Manipulated?

Unlike a simple averaging rule, the community note mechanism makes **precise** manipulation of note score around 0.4 difficult.

1. Bidirectional nature of manipulation and the computational cost
2. Coordinated manipulation is hard due to the bridging algorithm.



- **Eligibility** criteria to become a Community Notes contributor
- **Reputation** system
- **Bridging** algorithm — for a note to be shown on a post, it needs to be found helpful by people who have tended to disagree in their past ratings

X-axis is f_u. Credit: Jonathan Warden

Data

Twitter releases a public dataset of Community Notes every day. From June 11 to August 2, 2024, and from January 1 to February 28, 2025, we

1. downloaded the daily snapshots and conduct a day-to-day comparison to identify newly created notes;
2. extracted the tweet ID associated with each new note;
3. monitored the daily status of each noted tweet[1] including textual content, engagement metrics, author characteristics, and note display status.
   - Two compute nodes within an HPC cluster, each configured with 600 GB of RAM, one A100 GPU, and four CPU cores;
   - Note scores computed in about 10 hours for 14 GB of notes data daily.

[1] Twitter releases notes and rating with a 48-hour delay. Out of 106,297 noted tweets in the 2024 sample, 21,057 of them (about 19.8%) received notes during this hold-out period, among which 17,221 (about 81.8%) were retracted. For these 17,221 tweets, we do not have their status information and have obtained their display status using Twitter's scoring algorithm.

Summary Statistics

Variable	No. of Obs.	Mean	Std. Dev.	Definition
Outcome Measures				
$Retract$	264,600	0.21	0.41	Binary variable indicating whether a tweet is retracted
Treatment				
$NoteDisplayed$	264,600	0.13	0.33	Binary variable indicating whether a community note displays under a tweet
Tweet Characteristics				
$LogViews$	208,836	11.38	3.01	Log-transformed number of views a tweet receives
$LogLikes$	211,393	6.88	2.90	Log-transformed number of likes a tweet receives
$LogComments$	211,393	4.57	2.27	Log-transformed number of comments a tweet receives
$LogShares$	211,393	5.07	2.69	Log-transformed number of shares a tweet receives
$TweetTenure$	211,393	25.90	98.56	The number of days since the tweet was posted
User Characteristics				
$LogFollowers$	211,393	10.69	2.92	Log-transformed number of followers the user has
$LogFollowings$	211,393	6.50	2.25	Log-transformed number of accounts the user is following
$LogUpdates$	211,393	9.78	2.02	Log-transformed number of tweets ever published by the user
$LogMediaTweets$	211,393	7.95	2.14	Log-transformed number of tweets with media ever published by the user
$BlueCheck$	211,393	0.67	0.47	Binary variable indicating whether a user is verified
$UserTenureYear$	211,393	7.88	5.32	The number of years since a user joined Twitter

$\Pr(U \leq u|X = x)$ and $\Pr(Z \leq z|X = x)$ are continuous in x at $x = 0.4$.

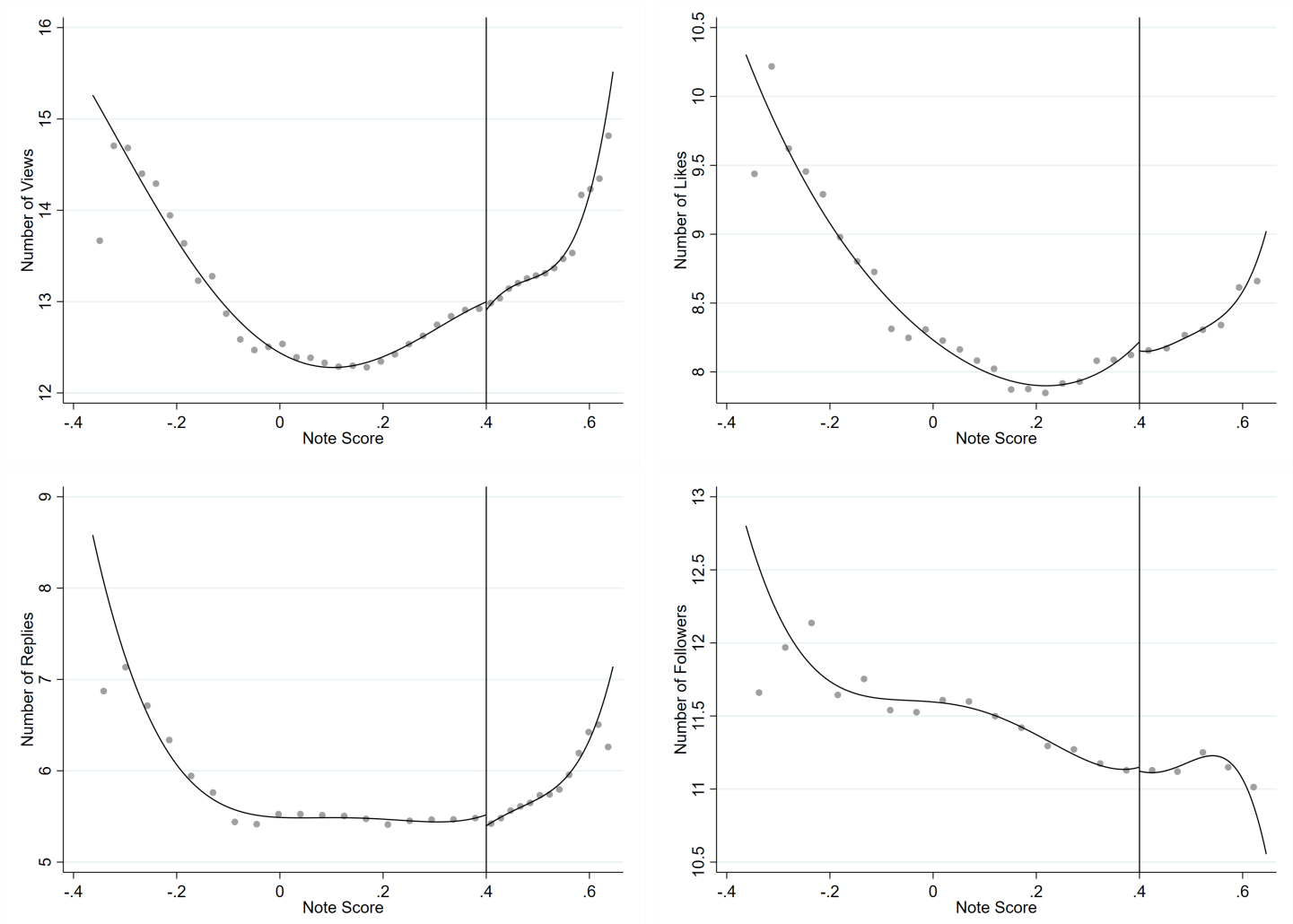


Figure: Covariate continuity checks: views (TL), likes (TR), comments (BL), and followers (BR)

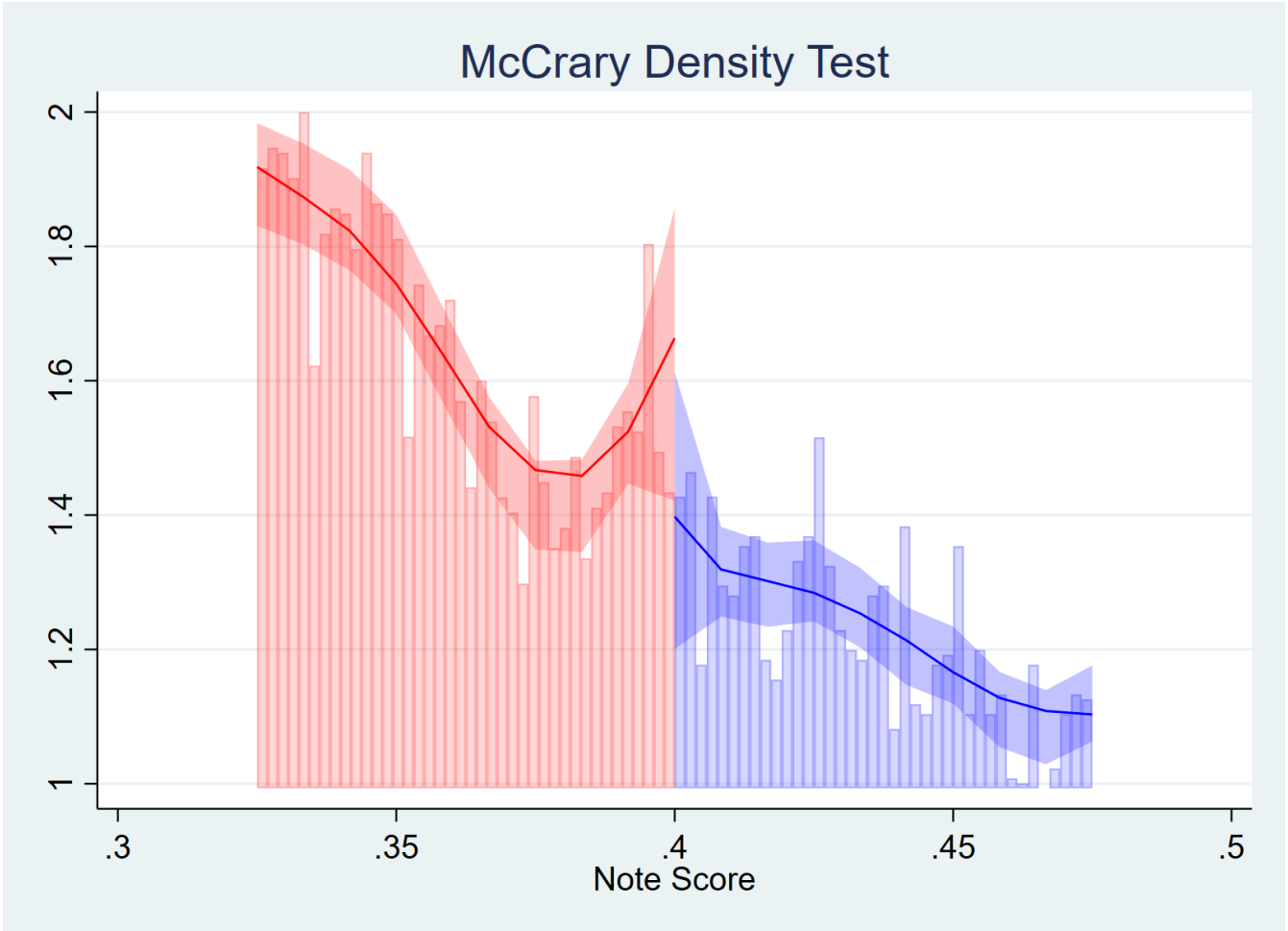
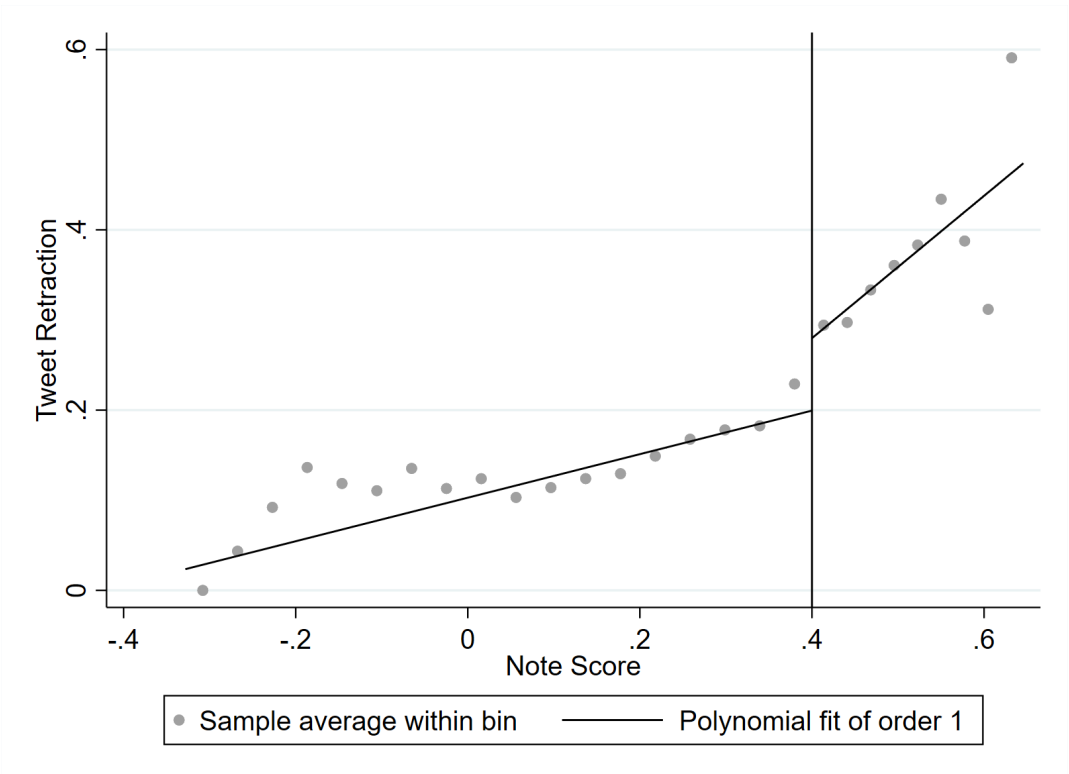


Figure: The t-statistic for discontinuity (by the Stata package *rddensity*) is -0.252 with a p-value of 0.801. Note that a running variable with a continuous density is neither necessary nor sufficient for identification (e.g., non-monotonic manipulation)

Model-Free Evidence



Examples of noted tweets that were subsequently retracted:

- *Suicide rates are through the roof since lockdowns.*
- *Hillary Clinton just endorsed a racist.*
- *NAFO (North Atlantic Fella Organization) is a terrorist organization actively engaging in hate speech to perpetuate the war between Russia & Ukraine.*

$$NoteDisplayed_i = \alpha_0 + \pi \cdot AboveThreshold_i + \sum_{k=1}^{K} \rho^k (NoteScore_i - Threshold)^k + AboveThreshold_i \cdot \sum_{k=1}^{K} \sigma^k (NoteScore_i - Threshold)^k + \eta_i$$

	$NoteDisplayed$	
	(1)	(2)
$AboveThreshold$	0.1531***	0.2651***
	(0.0122)	(0.0122)
Polynomial	Linear	Quadratic
Observations	21,806	52,677

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. The optimal bandwidth is calculated using the Calonico et al.(2014) procedure.

$$Retract_i = \alpha_1 + \beta \cdot \widehat{NoteDisplayed}_i + \sum_{k=1}^{K} \gamma^k (NoteScore_i - Threshold)^k$$
$$+ \widehat{NoteDisplayed}_i \cdot \sum_{k=1}^{K} \delta^k (NoteScore_i - Threshold)^k + \epsilon_i$$

	Retract	
	(1)	(2)
NoteDisplayed	0.3164***	0.1452***
	(0.0707)	(0.0389)
Polynomial	Linear	Quadratic
Observations	21,806	52,677

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are in parentheses. The optimal bandwidth is calculated using the Calonico et al.(2014) procedure.

Subsample 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

Subsample 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

Subsample 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

Subsample 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

Subsample RD Analysis by 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

Subsample RD Analysis by Tweet Age

	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

Subsample 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

Does Displaying Note Accelerate Retraction?

Table: 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	Int	t^2	Int	t^2
Observations	514,904	514,904	514,904	514,904
Log-Likelihood	-73062.538	-72724.836	-73068.660	-72728.605

Contributions to Literature

This study contributes to the literature on misinformation, particularly concerning countermeasures against misinformation.

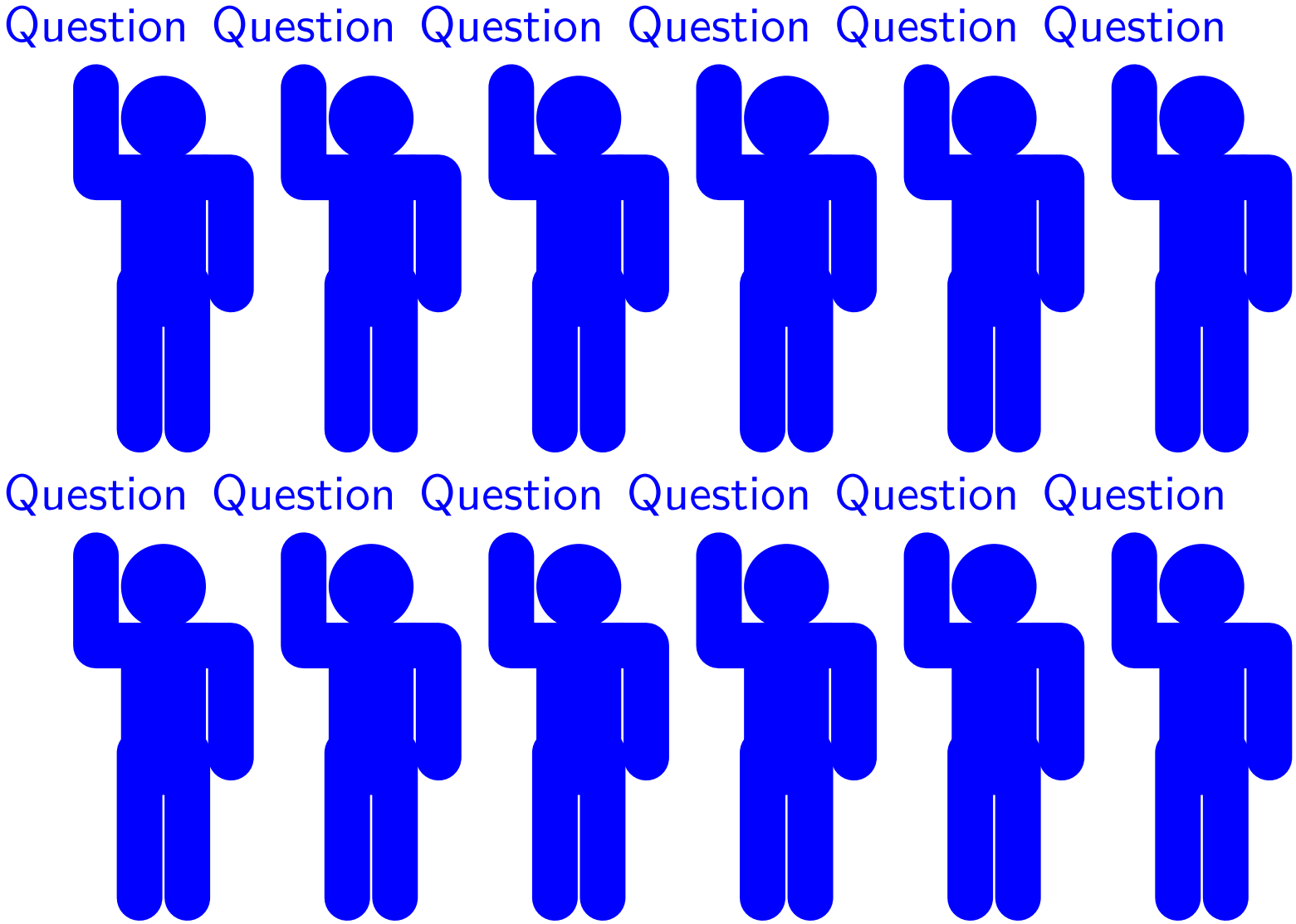
- We introduce a new angle to the misinformation literature by examining the effectiveness of crowdchecking on producers;
- We demonstrate the potential of **voluntary retraction** as an alternative to forcible removal of content;
- We uncover heterogeneous treatment effects suggesting reputation concern as the main mechanism.

Contributions to Practice

- **Individuals**: participate in crowdchecking much like one's jury duty.
- **Platforms**: like other crowd-based initiatives (e.g., Wikipedia), crowdchecking can be effective and is often less controversial and more scalable.
  - Transparency is crucial so that user trust will not erode.
  - Notify not just those who engaged with misinformation.
    For example, a platform can decide the proportion of users to notify based on note score.
- **Policymaker**: nudge or even mandate social media platforms to adopt transparent and effective crowdchecking systems.
  - Twitter: 2021
  - Facebook: 2025

Fact-Checkers $\rightsquigarrow$ Community Notes $\rightsquigarrow$ Fact-Checkers + Community Notes?

Thank You!


Question Question Question Question Question Question
Question Question Question Question Question Question