Online Anonymity under Stigmatization: Evidence from Geotagging Mandate in China

Mingzi Niu[*], Yiqian Wang[†] and Ziang Xiao[‡§]

Abstract

Social media platforms increasingly adopt user-geographic disclosure policies to combat bots and misinformation. Yet their effects on user participation and expressed opinions remain understudied. We develop a signaling model incorporating identity stigma when individuals express stigmatized views. In equilibrium, both the demographic composition of active users and prevailing opinions shift. Leveraging recent Chinese social media regulations and advances in large language models, we find that geographic disclosure, in a short-run, reduces participation among non-mainland users and those holding marginalized views, driving measurable opinion shifts against foreign entities. This outcome stems from heightened geographical stigma triggered by disclosure. Our findings highlight a critical trade-off: while IP disclosure may mitigate automated manipulation and misinformation, it compromises the representativeness of social media data for evaluating public opinion.

[*]The Chinese University of Hong Kong Business School. mingziniu@cuhk.edu.hk

[†]Digital Economy Lab, University of International Business and Economics. wangyiqian@uibe.edu.cn

[‡]Department of Computer Science, Johns Hopkins University. ziang.xiao@jhu.edu

[§]We are grateful to Prabhat Barnwal, Chris Ahlin, Kyoo Il Kim, Karen Macours, Yiming Cao, Soren Anderson, Jon Eugia, Anirudh Tagat, Imran Rasul, and Joris Mueller for valuable comments and suggestions. We also thank participants at the MSU Applied Economics Seminar, the BREAD Conference, and the Econometric Society World Congress for helpful feedback. **The authors declare that they have received no financial support for this research and have no financial relationships or other conflicts of interest relevant to this paper.** All errors are our own. All remaining errors are our own.

1 Introduction

Anonymity fosters freedom of expression, promotes diversity of opinions, and facilitates intergroup communication on social media platforms. However, this exact feature also diminishes the accountability of information dissemination, creating an environment where misinformation and hate speech can proliferate unchecked.

Meanwhile, the preservation of opinion diversity and intergroup communication on these platforms has emerged as a growing concern for policymakers. Burgeoning evidence shows that the government weighs social media sentiments and opinions when designing policies[1]. Therefore, it is increasingly critical to understand how anonymity shapes group participation and the diversity of opinions on social media platforms.

This paper sheds light on this debate by leveraging a unique identity-disclosure policy experiment in China. In early 2022, the Chinese government mandated one of China's largest social media platforms to geotag users – disclosing the users' geographic location on the platform. The policy is unique because it is the first government-mandated policy to regulate social media anonymity[2]. While the policy primarily focuses on deterring bots and foreign influences, it has also sparked rich discussions about its potential impact on civil participation (Yang & Xu, 2025).

To understand the potential effects of IP disclosure on user participation and opinions, we study a signaling game. This game considers individuals from two groups — the local group and the foreign group — who decide whether to support, criticize, or remain silent about a given post advocating a policy change. Groups learn the value of the policy change from different sources of information that may exhibit opposite media biases, with local media leaning towards the status quo and foreign media towards policy change. The media bias gap may reflect political censorship, ideological divides, or fundamental conflicts of interest across groups.

Crucially, we assume this media bias gap breeds and intensifies intergroup xenophobia. Individuals benefit from expressing their views, given their updated beliefs about the value of the policy change, while expecting the cost of social sanctions for their comments due to xenophobia. The social sanc-

[1]For example, Buntaine et al. (2024) used a social media experiment to show that advocating on social media increases the government's regulation of polluting factories in China.

[2]On November 23, 2025, X began geotagging its users at the country level.

tion cost associated with a comment varies with the media bias gap across groups and the identity disclosure policy implemented on the platform. We then characterize the equilibrium under two policies—(i) the social media is anonymous, and (ii) the group identities of participants are revealed—and examine how participation on social media changes with the intergroup media bias gap and the identity disclosure policy.

The model yields two sets of testable predictions. The first set of predictions, which we call the *Labeling Effects*, states that the identity disclosure suppresses the participation of the minority group. As a result, the opinions favored by the majority group will be more prevalent on social media. The second set of predictions, which we call the *Group Difference Effect*, states that the intergroup gap in media bias amplifies the Labeling Effect. Specifically, the larger the media bias gap, the more the identity disclosure suppresses the participation of the minority group.

We empirically test the predictions by compiling a unique dataset of 2.3 million comments posted on 18 thousand viral posts on Weibo, a major social media platform in China. These posts were from 509 influential posters who, on average, have 1.8 million followers. These posts and comments were created a few months before and after the policy implementation in April 2022. We developed a large language models powered pipeline to analyze sentiments and opinions in posts and comments, including users' overall attitudes and political leanings.

Preliminary data exploration reveals that the data aligns well with our model assumptions. First, we identify *non-mainland* users as the minority group who accounts for only 5% of all comments, who expressed substantially lower anti-foreign comments in political posts at the baseline. Second, posters' geographic identity predicts the political attitudes expressed in their posts, consistent with our assumption that geographic location serves as a proxy for media bias. Third, the proportion of geographically-targeted stigma directed at non-mainland users increased substantially following the disclosure policy, aligning with our model's characterization of xenophobic sanctions as a key factor influencing commenting decisions.

To identify the policy effects, we use a regression discontinuity in time (RDiT) design that exploits the sharp introduction of profile geotagging on March 18, 2022. The central identification assumption is that, within a narrow window around the policy change, the platform's underlying information environment evolves smoothly, so any discontinuous change in participation or expressed opinion can be attributed to identity disclosure rather than

concurrent shifts in content generation. To address this, we directly test whether post characteristics and post-level sentiment exhibit discontinuities at the cutoff. We focus on the profile-geotagging phase rather than the later comment-geotagging phase because the latter may already reflect spillovers or anticipation effects generated by the earlier reform.

Our preferred specification compares outcomes within a 40-day bandwidth around the policy cutoff, corresponding to the interval between profile geotagging and the later extension of geotagging to comments. Following Gelman and Imbens (2019), we use linear time trends on either side of the cutoff in the main specification and report quadratic trends as a robustness check. The regressions additionally control for post's contents, sentiments, exposure characteristics, as well as poster's characteristics. We further include day-of-week, post-topic, and poster fixed effects to absorb cyclical patterns and persistent heterogeneity across content and posters, and we cluster standard errors at the post-date level to account for correlated commenting behavior within the same post on the same day.

Our first result demonstrates that geotagging substantially reduces participation by geographically identifiable minority users. Using our preferred RDiT specification, we estimate that profile geotagging decreases the share of comments from non-mainland users by 0.86 percentage points, equivalent to a 19% decline relative to the pre-policy baseline. In contrast, we find no statistically or economically significant effect on non-mainland participation in second-order engagement (i.e., replies to comments), suggesting that the policy primarily suppresses initial expression rather than downstream interactions.

Our second result shows that geotagging substantially increases anti-foreign opinions among political posts. We evaluate the opinion through the sentiments detected from the comments. Geotagging in the profile increases anti-foreign sentiments by around 14.7%. On the other hand, the effect are large but less precisely estimated for non-political posts, with a point estimate of 18.8%.

To disentangle whether the opinion shift stems from existing commenters changing their views or from compositional changes in who comments, we decompose the effect into two components: the shift in opinions among users who comment both before and after geotagging, and the opinions expressed by new commenters who only participate after geotagging. We find that the increase in anti-foreign sentiment is primarily driven by the population shift in commenters, with new commenters expressing more anti-foreign sen-

timents than those who commented before geotagging. This suggests that the policy's impact on opinions is largely due to changes in the composition of participants rather than shifts in the views of existing commenters. Specifically, we find the new commenters are 18.4% more likely to express anti-foreign opinions than the existing commenters.

Our third result supports the Group Difference Effect. We find that IP disclosure reduces non-mainland participation more sharply among posters with larger media gaps. To measure the media gap, we benchmark non-mainland posters' posts and calculate the mean sentiment toward locations frequently mentioned in posts of specific topics. We approximate the media gap using Euclidean distance between the mean sentiments of mainland posters and the non-mainland poster benchmark. Examining two topics—the Russia-Ukraine War and COVID-19—we find that geotagging's suppressive effect on non-mainland participation is substantially larger for mainland posters with greater media bias gaps, confirming that intergroup media bias amplifies the Labeling Effect.

Finally, we investigate alternative mechanisms for the observed decline, including (i) strategic content moderation by original posters anticipating identity-based scrutiny, (ii) censorship, and (iii) bot deterance. Suggestive evidence does not support these explanations. Ruling out these channels strengthens our conclusion that the decline in non-mainland participation is primarily driven by increased social sanction costs from identity disclosure, consistent with the Labeling Effect and Group Difference Effect predicted by our model.

In summary, our study highlights a trade-off between user anonymity and user partcipation on the social media platform. We show that the shifting of the platform's opinion is not only driven by the decline of the minority group but also the compositional change of majority group. On the other hand, we are yet to show that curtailing the anonymity of users necessarily biased the representativeness of opinions on social media, as the marginalized opinions, such as pro-foreign opinions in our context, may be amplified on the platform before the policy change. Nevertheless, our finding suggests that anonymity reduction, such as IP disclosure, will substantially shift the user and opinion composition. Such a shift may distort the government's learning of public opinion through social media, which may further distort the government's policy design and implementation.

Overall, our study contributes to three strands of literature. First, our study contributes to the literature on the economics of intergroup interaction.

Many influential studies explored the economic return and behavioral change of integrating groups previously segregated by identities such as wealth level (Rao, 2019) and gender (Dahl et al., 2021) [3]. We contribute to this literature by examining the effects of segregating two groups based on location, political attitudes, and information acquisition methods.

Second, our study contributes to the economics of identity and social image. In a related study, Braghieri et al. (2022) experimentally showed that the concern about social image causes the publicly expressed view to deviate from the private view. In another experimental paper, Bursztyn et al. (2020) showed that the payment required for forgoing a socially recognized identity varies with whether the action can be observed. We add to the literature by showing the importance of covering identity to sustain intergroup communication.

Third, our study contributes to the growing literature on the economics of social media. Many recent important studies find social media has real-life effects, such as on mental health and political attitudes or outcomes [4]. Close to our context, Qin, Stromberg, and Wu (2017) and King, Pan, and Roberts (2013) study the role of Chinese social media in facilitating government surveillance. In terms of anonymity, Ederer, Goldsmith-Pinkham, and Jensen (2023) and Wu (2020) study the downside of anonymity on breeding harmful information even from highly-educated groups.

2 Institutional Background and Context

2.1 IP-based Geographic Location Disclosure

On March 1, 2015, the Cyberspace Administration of China (CAC) issued guidelines to all Chinese social media platforms, known as the Administration of Internet User ID. These guidelines prohibited user IDs containing keywords linked to sensitive topics, including national security, hate speech against minority groups, pornography, violence, and harassment. Despite granting

[3] Also other important identities such as ethnicities: Bazzi et al. (2019); race: Schindler and Westcott (2021); Corno et al. (2022), and social hierarchy: Lowe (2021)

[4] For mental health, see: Allcott et al. (2020); Mosquera et al. (2020); Braghieri, Levy, and Makarin, (2022)). For political attitudes or outcomes, see Allcott and Gentzkow (2017); Levy (2021); Burnsztyn et al. (2019); Enikolopov, Makarin, and Petrova (2020); Gorodnichenko, Pham, and Talavera (2021)

users considerable freedom regarding profile naming, the new Administration mandated users to verify their real ID before continuing to use the platform.

Over time, the Administration underwent several amendments. Related to this study, in October 2021, the CAC proposed an amendment demanding social media platforms disclose users' IP-based geographic locations. While the primary motive for this change wasn't specified, some alleged that the government believed disclosing IP locations could help curb misinformation fabricated by foreign bots. For instance, credibility would be challenged if someone posted about their COVID-19 lockdown experiences in Shanghai when their IP indicated they were based in a foreign country.

Sina Weibo, one of China's leading social media platforms, was the first to test the new guideline. From March 4, 2022, the platform started to experiment with the IP revelation function in posts related to the Russia-Ukraine War. Any post containing the keywords "Russia(俄罗斯)" or "Ukraine(乌克兰)" would have its comments tagged with the commenter's IP-based geographic location. This location is accurate up to the province within China or the country for users outside China.

The primary goal of this action was to reduce misinformation about the Russia-Ukraine War. The user's geographic location was a natural credibility indicator for the unidentified source. However, this strategy had limitations. For instance, a post can evade the keyword filter if it does not include the words "Russia" or "Ukraine." Conversely, users could manipulate the system by adding keywords to unrelated posts and triggering the IP revelation. The method cannot prevent users with virtual private network (VPN) where their IP addresses can be manipulated, although the government frequently bans such tools.

Profile geotagging Despite the controversy raised due to its implementation, Weibo expanded its identity-revelation efforts. Starting March 18, 2022, the platform attaches the user's IP-based geographic location to their profile. Figure E.2a shows a standard Weibo profile showcasing IP location. This IP location is determined by the latest IP address the user uses. Being inactive for over a month erases the geotag.

Comment geotagging From midnight on April 28, 2022, the platform started to attach the IP location to all new comments, regardless of whether the post was written before or after April 28. Figure E.2b illustrates this policy's sharp implementation. Two comments replied consecutively to a post, with the latter being tagged for replying just half an hour past midnight.

2.2 Mainland User vs. Non-Mainland User

China-based social media platforms are predominantly used by mainland Chinese users while serving a smaller portion of non-mainland Chinese users. Despite the shared language, anecdotal observations reveal a lingering distrust between these two groups. This division potentially arises from two factors. First, the two groups live in distinct environments with increasingly polarized political views. Most of the non-mainland comments in the data originate from North America and European countries, where the political systems differ largely from mainland China. Second, the groups receive news and information from different media sources. The government poses strong control on the media platforms' access to mainland users. The same news could be slanted in different ways. Mainland Chinese users' access to global information platforms such as Google, Facebook, and X (Twitter) has been restricted for over a decade.

This distrust manifests most explicitly in discussions of controversial topics. Before revealing the location identity, dissenters were frequently stigmatized as "foreign agents." For instance, frequently observed in COVID-19 posts made in early 2022, commenters who advocated for the termination of the zero-COVID policy were often stigmatized as non-mainlanders. These criticisms usually talk about foreign governments intentionally endangering Chinese public health.

3 Conceptual Framework

To illustrate the central mechanism through which the platform's identity disclosure policy shapes participation and opinion representation, and to guide our empirical analysis, we begin with a theoretical framework.

3.1 Model Setup

We consider a social media platform on which local and foreign users discuss a policy. Media slant may differ across groups; for example, local media tends to promote supportive propaganda, while foreign media is more likely to emphasize criticism. Under these media influences and the platform's identity-revealing policy, users decide whether to comment on the platform and, if they do, whether to support or oppose the policy.

Players. There is a unit mass of users $i \in [0,1]$. Each user belongs to either the local group (L) or the foreign group (F): $g \in \{L, F\}$. We denote by $\lambda \in (0,1)$ the fraction of local users.

Actions. Each user i needs to decide whether and how to comment on a policy: $d_i \in \{-1, 0, 1\}$. Here, $d_i = 0$, 1, -1 denotes that the user i does not comment, posts support, and posts criticism, respectively.

Information. Let $V \in \mathbb{R}$ be the unknown objective value of the policy. Crucially, we assume that groups hold systematically different beliefs about V due to media bias. In particular, local users are exposed to more favorable media coverage and therefore tend to form more optimistic views of the policy than foreign users:

$$v_i \mid i \in L \sim \mathcal{N}(\mu_L, \sigma^2),$$
$$v_i \mid i \in F \sim \mathcal{N}(\mu_F, \sigma^2),$$

where $\mu_L \geq \mu_F$, and $\sigma^2 > 0$ denotes the variance in opinions within each group. Here, $\mu_g - V$ captures the media bias that affects group g.

Define the gap in media slants between groups as

$$\Delta = \mu_L - \mu_F \geq 0.$$

This gap may arise from political censorship, ideological divisions, or underlying conflicts of interest across groups. Each user observes this intergroup gap Δ but does not know the exact magnitude of their own group's bias, and thus cannot fully debias their opinion.

We analyze the following two identity-revealing policies, $\mathcal{P} \in \{\emptyset, G\}$, on the social media platform.

I. *Anonymous Policy* ($\mathcal{P} = \emptyset$): group identities of commenters are hidden.

II. *Geotagging Policy* ($\mathcal{P} = G$): group identities of commenters are revealed.

Payoffs. For any user $i \in [0,1]$, we normalize the payoff from not commenting ($d_i = 0$) as zero, and specify the payoff from commenting $d_i = \pm 1$ under the identity-revealing policy $\mathcal{P} \in \{\emptyset, G\}$ as

$$U_i^{\mathcal{P}}(d_i) = I_i(d_i) + S_i^{\mathcal{P}}(d_i).$$

Here, $I_i(d_i) \in \mathbb{R}$ denotes the intrinsic value of decision d_i for user i; $S_i^{\mathcal{P}}(d_i) \in \mathbb{R}$ denotes the associated social sanction under the policy $\mathcal{P}$ for user i.

The intrinsic value of commenting, $d_i = \pm 1$, is given by

$$I_i(d_i) = \beta v_i d_i - \gamma, \tag{1}$$

where $\beta > 0$ and $\gamma \geq 0$. The first term represents the expressive benefit, proportional to the policy value perceived by user i.[5] Therefore, users prefer to post support ($d_i = 1$) when they perceive the policy as beneficial ($v_i > 0$), and to post criticism ($d_i = -1$) when they perceive it as harmful ($v_i < 0$). The second term, $\gamma \geq 0$, represents the operation cost of commenting (e.g., the effort involved in composing and posting a comment).

The *xenophobia sanction* of commenting, $d_i = \pm 1$, penalizes foreign users who oppose the policy but rewards those who support it, which is given by

$$S_i^{\mathcal{P}}(d_i) = \psi(\Delta) \Pr(i \in F | d_i, \mathcal{P}) d_i, \tag{2}$$

where $\psi(\cdot) \geq 0$. We assume that the intergroup media bias gap Δ intensifies xenophobia sanctions, and thus $\psi(0) = 0$ and $\psi'(\cdot) > 0$.

3.2 Analysis

3.2.1 Benchmark: No Media Bias ($\Delta = 0$)

We first consider a benchmark in which there is no media bias; that is, $\mu_L = \mu_F = V$. Hence, $\Delta = 0$, which implies there are no xenophobia sanctions against foreign commenters, and the two groups, $g \in \{L, F\}$, share the same distribution of opinions of the policy:

$$v_i \sim \mathcal{N}(V, \sigma^2).$$

Without xenophobia sanctions, users' commenting decisions are independent of their group identities (and thus the identity-revealing policies on the platform). User i posts support if and only if $\beta v_i - \gamma > 0$; that is,

$$v_i > \bar{v} := \frac{\gamma}{\beta} \geq 0. \tag{3}$$

[5] The expressive benefit captures the *warm glow* (see, e.g., Andreoni, 1995; Bursztyn et al., 2023) from expressing the right view—that is, the satisfaction individuals derive from doing their own part.

Likewise, user i posts criticism if and only if $\beta v_i - \gamma > 0$; that is,

$$v_i < -\bar{v} = -\frac{\gamma}{\beta}. \tag{4}$$

Since $v_i \sim \mathcal{N}(V, \sigma^2)$, we can derive the mass of supports as

$$\Pr(d_i = 1) = 1 - \Phi\left(\frac{\bar{v} - V}{\sigma}\right) = \Phi\left(\frac{-\bar{v} + V}{\sigma}\right), \tag{5}$$

where $\Phi(\cdot)$ denotes the Cumulative Distribution Function for the standard normal distribution, and the mass of negative commenters is given by

$$\Pr(d_i = -1) = \Phi\left(\frac{-\bar{v} - V}{\sigma}\right). \tag{6}$$

Note that when $V = 0$, $\Pr(d_i = 1) = \Pr(d_i = -1)$; that is, positive and negative comments are of equal size. In this case, the balanced distribution of opinions on social media accurately reflects the policy's null value. As the policy value V increases, the share of support ($d_i = 1$) rises, while the share of criticism ($d_i = -1$) declines. Therefore, the prevalence of opinions on social media indicates the true value of the policy; in this sense, information aggregation serves to reveal the truth.

3.2.2 Under Media Bias ($\Delta > 0$)

Media bias distorts users' commenting behavior on social media through two channels. First, it shifts the distribution of opinions across groups, leading the local group to hold more favorable views of the policy than the foreign group. Second, it activates xenophobic sanctions, which in turn shift the thresholds for posting support or criticism. The first channel operates independently of whether group identity is disclosed, whereas the second critically depends on the identity-revealing policy.

To be precise, under media bias, the thresholds for posting support or criticism for users of group g under the policy $\mathcal{P}$, (3) and (4), now become

$$\bar{v}^g_{\mathcal{P}} := \bar{v} - \frac{\psi(\Delta)}{\beta} \Pr(i \in F | d_i = 1, \mathcal{P}), \tag{7}$$

$$\underline{v}^g_{\mathcal{P}} := -\bar{v} - \frac{\psi(\Delta)}{\beta} \Pr(i \in F | d_i = -1, \mathcal{P}), \tag{8}$$

where $\bar{v}$ is given by (3). Thus, holding fixed the distribution of opinions across groups, media bias lowers the thresholds for both support and criticism. Consequently, relative to the benchmark case, the share of support rises while that of criticism declines.

3.2.3 Anonymous Policy ($\mathcal{P} = \emptyset$)

Consider the scenario when the group identity is hidden. The social sanction of xenophobia operates solely based on the content of comments, independent of the user's true group identity. Therefore, the thresholds for commenting are identical across groups; we can thus drop the superscript of group g in the expression of commenting thresholds (7) and (8).

Formally, any user of i group g holds an opinion $v_i \sim \mathcal{N}(\mu_g, \sigma^2)$, posts support ($d_i = 1$) if and only if their opinion is sufficiently high, $v_i \geq \bar{v}_\emptyset$, and posts criticism ($d_i = -1$) if and only if their opinion is sufficiently low, $v_i \leq \underline{v}_\emptyset$. Hence, within the group g, support occurs with probability $1 - \Phi(\frac{\bar{v}_\emptyset - \mu_g}{\sigma}) = \Phi(\frac{\mu_g - \bar{v}_\emptyset}{\sigma})$ and criticism occurs with probability $\Phi(\frac{\underline{v}_\emptyset - \mu_g}{\sigma})$. Therefore, by Bayes' rule,

$$\Pr(i \in F | d_i = 1, \mathcal{P} = \emptyset) = \frac{(1-\lambda)\Phi(\frac{\mu_F - \bar{v}_\emptyset}{\sigma})}{(1-\lambda)\Phi(\frac{\mu_F - \bar{v}_\emptyset}{\sigma}) + \lambda\Phi(\frac{\mu_L - \bar{v}_\emptyset}{\sigma})}, \tag{9}$$

$$\Pr(i \in F | d_i = -1, \mathcal{P} = \emptyset) = \frac{(1-\lambda)\Phi(\frac{\underline{v}_\emptyset - \mu_F}{\sigma})}{(1-\lambda)\Phi(\frac{\underline{v}_\emptyset - \mu_F}{\sigma}) + \lambda\Phi(\frac{\underline{v}_\emptyset - \mu_L}{\sigma})}. \tag{10}$$

For any group composition $\lambda \in (0, 1)$, equations (9) and (10) , together with and $\mu_L \geq \mu_F$, imply that

$$0 < \Pr(i \in F | d_i = 1) \leq \Pr(i \in F | d_i = -1) < 1, \tag{11}$$

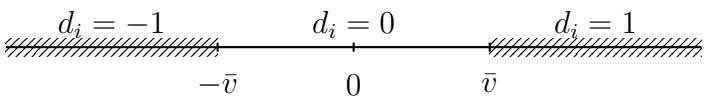
and the second inequality holds strictly if the media slant gap $\Delta > 0$. By (7) and (8), this in turn implies $\underline{v}_\emptyset \leq \bar{v}_\emptyset$. Intuitively, because the foreign group is more biased against the policy than the local group, as oppose to support, criticism is more indicative of a foreign user.

The following proposition characterizes the equilibrium under media bias when the platform adopts the Anonymous Policy

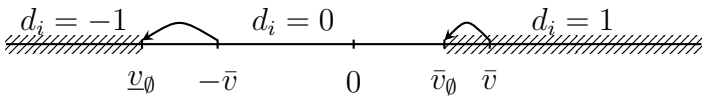
Proposition 1 (Equilibrium under Anonymous Policy)**.** *For any $\lambda \in (0, 1)$, there exists an equilibrium characterized by two cutoffs $\underline{v}_\emptyset \leq \bar{v}_\emptyset$ such that users with sufficiently low opinions $v_i < \underline{v}_\emptyset$ post criticism ($d_i = -1$), users*

with intermediate opinions $v_i \in (\underline{v}_\emptyset, \bar{v}_\emptyset)$ remain silent ($d_i = 0$), and users with sufficiently high opinions $v_i > \bar{v}_\emptyset$ post support ($d_i = 1$).[6] *Additionally, there exist $\underline{\lambda}, \bar{\lambda} \in (0,1)$ such that, if $\lambda > \bar{\lambda}$ or $\lambda < \underline{\lambda}$, then the equilibrium cutoffs $(\underline{v}_\emptyset, \bar{v}_\emptyset)$ are unique.*[7]

Figure 1: Commenting Regions



(a) Benchmark: No Media Bias



(b) Under Anonymous Policy

3.2.4 Geotagging Policy ($\mathcal{P} = G$)

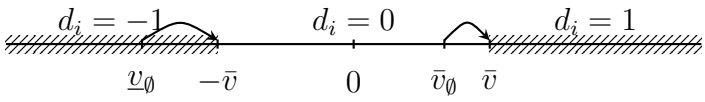
Now we turn to the case when the group identity is revealed when commenting. In this case, the social sanction or reward is directly applied to foreign commenters. By (7) and (8), we derive (i) a local user $i \in L$'s commenting choices are characterized by the cutoffs $(-\bar{v}, \bar{v})$ where $\bar{v}$ is given by (3), as in the benchmark case without xenophobia sanction; and (ii) a foreign user $i \in F$'s commenting choices are characterized by the cutoffs $(-\bar{v} - \frac{\psi(\Delta)}{\beta}, \bar{v} - \frac{\psi(\Delta)}{\beta})$.

Proposition 2 (Equilibrium under Geotagging Policy)**.** *For any $\lambda \in (0,1)$, there exists a unique equilibrium characterized by two pairs of cutoffs, $(\underline{v}_L, \bar{v}_L)$ and $(\underline{v}_F, \bar{v}_F)$ such that for each group $g \in \{L, F\}$: users with sufficiently low opinions $v_i < \underline{v}_g$ post criticism ($d_i = -1$), users with intermediate opinions $v_i \in (\underline{v}_g, \bar{v}_g)$ remain silent ($d_i = 0$), and users with sufficiently high opinions $v_i > \bar{v}_g$ post support ($d_i = 1$), where $\bar{v}_L = -\underline{v}_L = \frac{\gamma}{\beta}, \bar{v}_F = \frac{\gamma - \psi(\Delta)}{\beta}, \underline{v}_F = -\frac{\gamma + \psi(\Delta)}{\beta}$.*

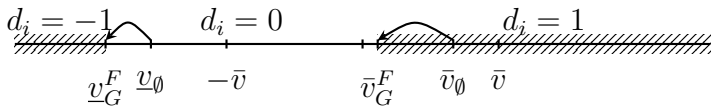
[6] $\underline{v}_\emptyset = \bar{v}_\emptyset$ if and only if $\gamma = \Delta = 0$.

[7] All proofs are omitted in the main text and can be found in the appendix.

Figure 2: Commenting Regions Under Geotagging Policy.



(a) Local Group



(b) Foreign Group

3.3 Predictions

To summarize, we learn from Section 3.2 that:

(i) If the platform adopts the anonymous policy, then the local and foreign groups share the same commenting cutoffs:

$$\bar{v}_\emptyset = \bar{v} - \frac{\psi(\Delta)}{\beta} \Pr(i \in F | d_i = 1, \emptyset),$$
$$\underline{v}_\emptyset = -\bar{v} - \frac{\psi(\Delta)}{\beta} \Pr(i \in F | d_i = -1, \emptyset),$$

where $\bar{v} = \gamma/\beta \geq 0$, and $0 < \Pr(i \in F | d_i = 1, \emptyset) \leq \Pr(i \in F | d_i = -1, \emptyset) < 1$, with the conditional probabilities defined by (9) and (10). In particular, in the absence of media bias ($\Delta = 0$), the commenting cutoffs reduce to $\bar{v}$ and $-\bar{v}$.

(ii) If the platform adopts the geotagging policy, then the commenting cutoffs for the local group are:

$$\bar{v}_G^L = \bar{v}, \qquad \underline{v}_G^L = -\bar{v};$$

whereas the commenting cutoffs for the foreign group are:

$$\bar{v}_G^F = \bar{v} - \frac{\psi(\Delta)}{\beta}, \qquad \underline{v}_G^F = -\bar{v} - \frac{\psi(\Delta)}{\beta}.$$

Drawing on these results, we can generate the following predictions.

Prediction 1 (Labeling Effects)**.** *Suppose* $\mu_F \leq -\psi(\Delta)/\beta$ *and* $\lambda \in (0,1)$ *is sufficiently high. When the group identity becomes revealed, the model delivers the following predictions:*

1. *Participation shift: participation by foreign users on the social media decreases relative to the participation by local users;*

2. *Opinion shift: the share of criticism decreases, while the share of support increases;*

The labeling effects described above are driven by the change in commenting thresholds for foreign users when the platform switches from an anonymous policy to a geotagging policy. Specifically, the threshold shift is approximately $\psi(\Delta)/\beta$, which is intensified by the intergroup gap in media slant Δ. In particular, if two groups share an identical media slant (i.e., $\Delta = 0$), then the commenting thresholds reduce to $\bar{v}$ and $-\bar{v}$, replicating the benchmark case. In this case, commenting behavior is independent of users' group identity and of whether the platform reveals identity.

Prediction 2 (Group Difference)**.** *The labeling effects are amplified by the intergroup gap in media slant* Δ*, and it disappears when two groups share an identical media slant (i.e.,* $\Delta = 0$*).*

4 Data

4.1 Data collection and classification

We discuss data collection and classification in detail in Section C. In brief, to construct the sample, we developed a topic-oriented data collection procedure because no comprehensive roster of influential Weibo accounts exists. We first identified major news topics during the sample period and derived 30 related keywords capturing the most salient discussions on Chinese social media. Using these keywords, we retrieved posts that attracted substantial engagement and compiled a pool of active posters. After excluding government- or media-affiliated accounts, celebrity and fan accounts, and deactivated users, we sampled from the remaining influential posters and then scraped all posts they published during the sample period. For each month's most-liked posts, we further collected comments and associated metadata.

The resulting dataset contains 17,748 posts from 509 posters and 2.3 million comments.

We complement the scraped data with LLM-based text analysis to measure the content of posts and comments at scale. Table D.2 summarizes the primary models that we employ for each task. For posts, we classify topics, identify key entities, and code the poster's attitude toward those entities. For comments, we code whether they express approval or disapproval of the post, whether they contain anti-foreign sentiment, and, for replies, whether they employ geographically targeted stereotyping. To improve coding quality in this high-context setting, we use a multi-stage pipeline that combines claim extraction, slang normalization, and codebook-based prompting, with prompts iteratively refined against expert-coded samples and final labels determined by majority vote across repeated model runs.

4.2 Data Summary

Table 1 shows comprehensive summary statistics for posts (Panel A), posters (Panel B), and comments (Panel C).

4.2.1 Posters

We sampled 509 posters who are predominantly influential on the platform. As shown in Figure E.5a, on average, these posters own 1.8 million followers, with the most popular poster followed by a staggering 24.8 million followers. As shown in Figure E.5b, approximately 18% of these posters are non-mainland users. The majority of non-mainland posters are based in the United States, followed by Canada, Australia, and European countries.

4.2.2 Posts

In total, we collected 17,748 Weibo posts posted between January 1, 2022, and August 31, 2022. The most popular post, a rumor about two Chinese comedians, received nearly one million likes, over 300 thousand reposts, and over 80 thousand comments. Figure E.3 illustrates the key attributes of these posts.

Table D.3 summarizes the distribution of post categories. Since many posts span multiple categories, the proportions sum to more than 100%. Politics (36.2%) and social news (52.2%) dominate our dataset, reflecting

Table 1: Summary Statistics of Posters, Posts, and Comments

	Mean	SD	Min	p25	p75	Max
Panel B. Poster			**N = 509**			
# of followers (10,000)	181.4	262.6	0.7	23.9	218.9	2,475.1
# of likes	7089.3	15,397	31.7	656.6	6342.1	150826
# of reposts	608.8	1168.5	0.1	55.6	558.8	9533.1
# of comments	619.3	1454.9	14.2	100.5	568.4	17690.7
Panel A. Post			**N = 17,748**			
# of likes	7221.6	28752.2	0	236	3348.5	944,024
# of reposts	619.7	3372.2	0	14	279	310,457
# of comments	635.9	2309.9	1	61	462	85,617
Panel C. Comment			**N=2,279,997**			
# of likes	11.97	354.4	0	0	1	127,683
# of first-stage comments likes	17.32	455.82	0	0	1	127,683

Notes: This table summarizes the main characteristics of Posts, Posters, and Comments in our data. The post statistics show the distribution of the number of likes, reposts, and comments received by the posts. The poster statistics show the distribution of the *average* number of likes, reposts, and comments of her posts. The comment statistics show the distribution of the number of likes, reposts, and comments received by the comments. The first-stage comments are the comments that replied directly to the post, not the comments that replied to some other comments.

both our keyword selection strategy and the prevalence of politically and socially oriented content among influential posters on Chinese social media. Entertainment posts comprise 24.9%, knowledge posts comprise 11%, and advertisements and sports posts are minimal at 3.7% and 2.9%, respectively, validating our effective exclusion of fan accounts, which typically focus on celebrities and promotional content. Additionally, COVID-19 and Russia-Ukraine War posts account for 20.78% and 8.28% of all posts, respectively.

The LLM also extracted key locations from each post and identified whether any were geographically relevant. Overall, a majority of posts (65%) involve China. Posts concerning North American countries follow at approximately 16%, with the United States being the primary focus. Approximately 8% of posts involve European countries or the Russia-Ukraine conflict. Hong Kong or Taiwan-related and East Asia-related posts each account for roughly 5%.

Figure E.7 shows substantial geographic variation in sentiment. On average, posts express positive sentiment toward China and European countries, with positive attitudes toward Europe driven largely by tourism-related content. The USA and Russia receive similarly negative sentiment, while Ukraine receives significantly more negative sentiment. The Russia-Ukraine sentiment disparity aligns with anecdotal evidence that Chinese social media users generally favor Russia over Ukraine[8]

4.2.3 Comments

We scraped a total of 2.3 million comments made to these posts. The number of likes received by comments presents a distribution that shows a high skewness to the right, averaging 11.4 likes, with 75% having no likes. Notably, the most liked comment garnered an astonishing 127,000 likes.

We find non-mainland comments only account for 4.22% of all comments in our data. Figure E.10a shows the geographic distribution of non-mainland comments. The number of comments in each country positively correlates with the number of Chinese immigrants in each country. The US leads the count with over 16,000 comments, surpassing the comment volume from several Chinese provinces. Other countries with significant comment counts also host large Chinese immigrant populations, including Canada, Australia, Japan, various European nations, and countries in South Asia. Notably, while African and South American countries contribute relatively few comments, these comments are distributed widely across both continents.

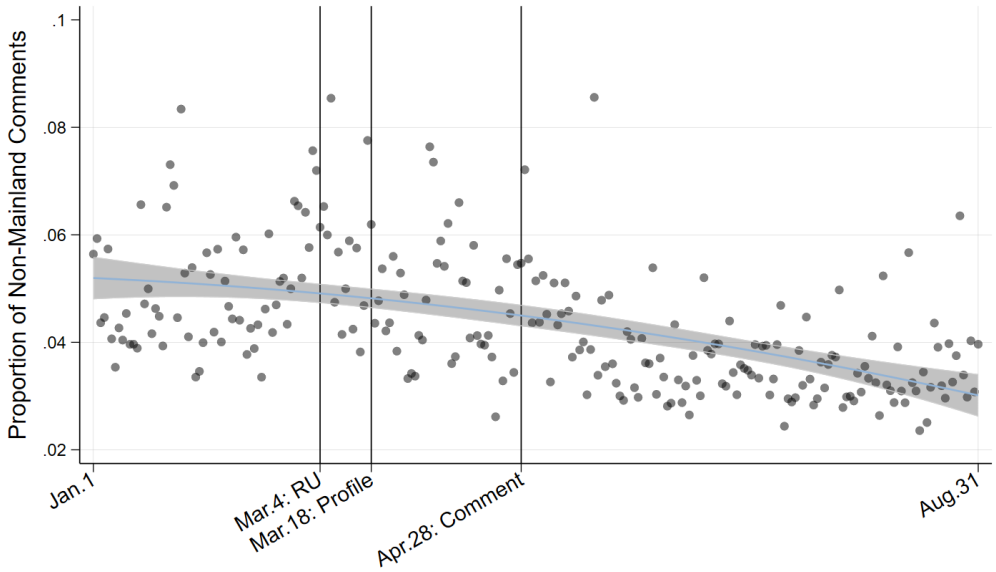
Figure E.10b illustrates the distribution of comments from mainland users. As further depicted in Figure E.11, we find a noticeable correlation between the number of comments and the population of provinces and municipal cities. However, Beijing and Shanghai stand out as exceptions. Despite having medium-sized populations compared to other provinces and municipalities, they rank second and third in the number of comments. Only Guangdong province, the most populated province in China, has more comments than these two cities.

Figure 3 illustrates the main motivation of the paper. Each dot represents the daily share of comments made by non-mainland users, measuring the relative participation of non-mainland users. The figure reveals an overall

[8]One caveat: this figure excludes poster fixed effects. Anti-Western posters may disproportionately generate anti-Western content, substantially reducing mean sentiment toward the USA and Ukraine.

downward trend beginning around Mid-March, 2022, when Weibo first tested comment geotagging on Russia-Ukraine–related posts.

Figure 3: Daily Share of Non-Mainland Comments, January 1 – August 31, 2022



Notes: The dots represent the daily share of comments made by non-mainland users. The commenter's geographic origin is identified using the IP location tag displayed in the user's profile or in the comment. The fitted line shows the ordinary least squares trend, with the shaded region representing the 95 percent confidence interval for the mean prediction.

4.3 Empirical Evidence for Theoretical Assumptions

In the theoretical framework, we make the following basic assumptions:

Assumption 1. *Media sources from different groups slant information in different directions.*

Assumption 2. *Individuals from different groups hold different baseline beliefs.*

Assumption 3. *Xenophobic sanctions exist against minority groups.*

In this section, we provide empirical evidence to support these assumptions and to validate the relevance of our theoretical framework in the context under study.

4.3.1 Poster location-based sentiment heterogeneity

Using location-based sentiment data, we provide direct empirical support for *Assumption 1* by showing that posters—each of whom can be viewed as an individual media source—exhibit systematic differences in the sentiment that can be predicted by their geographic location.

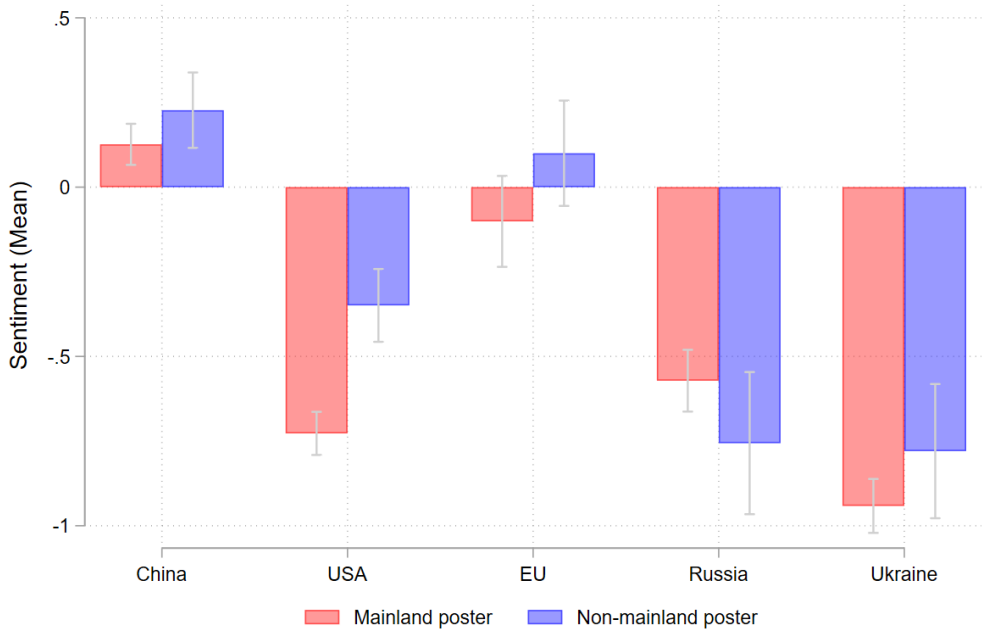
Figure 4 shows that both mainland and non-mainland posters express significantly negative sentiment toward the United States, Russia, and Ukraine, relative to the mean sentiment across all posts. However, compared with non-mainland posters, mainland posters exhibit (i) significantly stronger negative sentiment toward the United States, (ii) weakly stronger negative sentiment toward Ukraine, and (iii) weakly weaker negative sentiment toward Russia. In addition, sentiment toward European countries is marginally negative among mainland posters but marginally positive among non-mainland posters.

This intergroup gap in poster sentiment persists even when conditioning on the topic of the posts. When discussing similar topics, the sentiment gap between the two groups is reflected in their attitudes toward specific locations. In discussions of COVID-19, mainland and non-mainland posters exhibit substantial differences in their sentiment toward Beijing and the United States, with mainland posters expressing more negative sentiment toward both locations, as illustrated in Figure E.8. In discussions of the Russia–Ukraine conflict, mainland posters express significantly more positive sentiment toward Russia, whereas non-mainland posters express significantly more positive sentiment toward Ukraine, as shown in Figure E.9.

4.3.2 Commenter location-based sentiment heterogeneity

As suggestive evidence for *Assumption 2*, Table 3 documents significant differences in the sentiment of comments between mainland and non-mainland users. Overall, non-mainland users display more subdued sentiment. Their comments are 8% less negative, roughly equally positive, and 10% more neutral compared with those of mainland users. In political posts, non-mainland

Figure 4: Poster Location-Based Sentiment Heterogeneity



Notes: This figure displays average sentiment toward key locations mentioned in the posts, separated by mainland and non-mainland posters. Sentiment is measured using the LLM-based coding procedure described in Section C, on a 5-point Likert scale ranging from -2 (most negative) to $+2$ (most positive). Error bars represent 95% confidence intervals around mean sentiment estimates.

users also exhibit lower anti-Western sentiment. Only 44% of their responses to international political posts convey anti-Western sentiment, compared with 52% among mainland users. This disparity underscores systematic differences in the political attitudes between the two groups.

4.3.3 Geographical discrimination

Figure 5 provides evidence for *Assumption 3* by showing that non-mainland users indeed receive greater geographical stigma with their group identity disclosed. Specifically, we examine replies responding to other comments before (left panel) and after (right panel) the platform introduced geotagging

Table 2: Mainland vs. Non-mainland Comments

	Mainland			Non-mainland			
	N	mean	sd	N	mean	sd	diff
A. General Characteristics							
# of likes	2,183,756	11.97	356.03	96241	11.76	316.39	-0.210
Comment length	2,183,756	22.473	25.71	96241	24.918	28.06	2.445***
B. Sentiments of first-order comments							
Disapprove	889,817	0.29	0.45	33,349	0.27	0.45	0.012***
Neutral	889,817	0.35	0.47	33,349	0.35	0.48	-0.002
Approve	889,817	0.36	0.48	33,349	0.37	0.48	-0.009***
Anti-foreign	321,381	0.35	0.48	18,075	0.28	0.45	0.072***

In Panel A, the comment length is calculated by the STATA *ustrlen* command, which assigns a length of 1 to each Chinese/English character and punctuation. The length of English comments is inflated due to this nature but non-Chinese comments are extremely rare in our data. In Panel B, we report the sentiments of first-order comments, which are the comments that are replied directly to the posts.
All the difference is calculated by regressing the characteristic on a non-mainland dummy without adjusting the standard error.

in user profiles, and compare the proportion of geo-stigma in disapproval comments directed at mainland and non-mainland commenters.

Comparing the bars within each panel, we observe that both groups experience geo-stigma, with non-mainland users facing stronger stigma. Prior to the IP-disclosure policy, the difference in received geo-stigma between the two groups is statistically insignificant. After the policy is implemented, however, the difference becomes statistically significant. In particular, comparing across panels shows that geo-stigma slightly decreases for mainland commenters but increases substantially for non-mainland commenters.

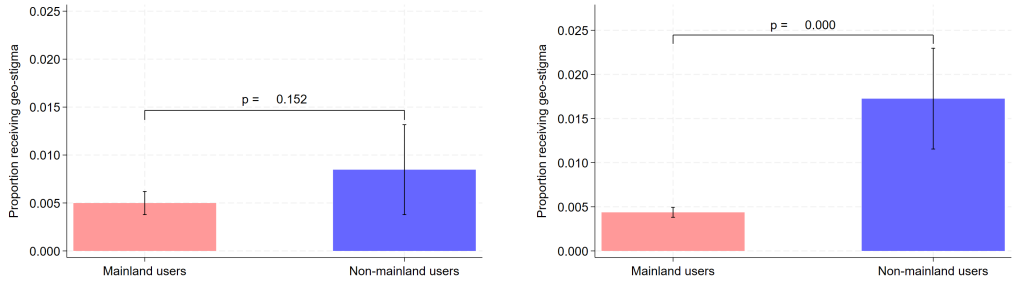
5 Empirical Strategy

5.1 Regression Discontinuity in Time Design

We estimate the impacts of IP disclosure on user participation and opinion expressions using a Regression Discontinuity in Time (RDiT) design.

The identification assumption is that the information and sentiment environments on the platform are identical in a narrow time window around the policy implementation. Therefore, changes in users' participation and opinions are caused by identity revelation rather than other channels, such as posters altering the content, narrative, or sentiment of the posts to cater

Figure 5: Geo-stigma Received by Mainland vs. Non-mainland Comments when being Disapproved



(a) Pre IP-disclosure (b) Post IP-disclosure

Notes: These two figures show the comparisons between the geo-stigma replies received by the unfavorable comments between mainland commenters and non-mainland commenters. Error bars represent 95% confidence intervals around mean sentiment estimates.

to the tastes of a specific group of users. We test for discontinuities in characteristics and sentiments of the posts on the policy implementation day.

Specifically, we estimate a RDiT equation specified as Equation 12:

$$y_{ijpt} = \beta_1 Geotag_t + \beta_2 f^a(Time_t) + \beta_3 Geotag_t \times f^b(Time_t) + \gamma X_{jpt} + DoW_t + PostType_p + Poster_p + \epsilon_{ijpt}. \quad (12)$$

Subscript i stands for the i-th commenter, j stands for the j-th post, p stands for the p poster, and t stands for the date. y_{ijpt} indicates if the comment i's geographic location is somewhere outside of mainland China (Non-mainland). $Geotag_t$ indicates whether the profile geotagging policy has been implemented at date t. We pursue an arguably cleaner identification by focusing on analyzing the profile-geotagging phase. This is because any estimated effect of the comment-geotagging policy (post-April 28, 2022) may be contaminated by the potential spillover and anticipation resulting from the profile-geotagging phase. Therefore, $Geotag_t$ equals one if the comment i was made on or after Mar 18, 2022.

To address these concerns, we restrict the bandwidth to 40 days—the policy interval between profile and comment geotagging. We group two days as a one-time interval to smooth the trend, as the variation in the ratio of the commenter's group across days is large in the sample period. In addition,

we estimate a "donut" RD specification by dropping one week before and after the policy's implementation to address any potential anticipation and the instantaneous effect.

The variable $Time_t$ is the date centered around the first day of geotagging in the user's profile. $f^a(Time_t)$ and $Geotag_t \times f^b(Time_t)$ control the time trends before and after the policy implementation. In the preferred specification, f^a and f^b are linear time trends recommended by Gelman and Imbens (2019). We also report the quadratic time trends to evaluate the robustness of the results. We use a uniform kernel to put equal weights on the observations within the time window. Using a uniform kernel mitigates the potential bias in favor of the short-run treatment effect.

In a set of control variables X_{jpt}, we include the post's number of likes, reposts, comments, its time span[9], and the poster's number of followers to control for the post's diffusion level. Further, we include binary indicators for (1) whether the post mentions any non–mainland China entities, and (2) whether the poster resides somewhere outside of mainland China to address the potential heterogeneity in attracting specific groups of users. For example, posts containing non-mainland China entities or written by non-mainland China posters are more likely to attract comments from non-mainland China users. Furthermore, we control location sentiments, a set of categorical variables that assign positive, negative, or neutral sentiment to locations such as China, Beijing, the USA, etc. If a post does not explicitly include or discuss a location, we set all values to neutral.

We include Day-of-Week fixed effects in all specifications to rule out cyclical effects within the narrow time window. We also include post-topic fixed effects to rule out any time-invariant heterogeneity within each combination of topics. For example, a post discussing Putin's geopolitical strategy will be classified as a *Political* post, but a post discussing Putin's childhood experience will be classified as a *Political*×*Knowledge* post. The potential sentimental differences towards Putin may attract different groups of commenters. Furthermore, we include the posters fixed effects to control the potential heterogeneity across posters. These heterogeneity includes the differences in posters' sentimental bias, political inclination, ability to diffuse the post, and any time-invariant poster-level characteristics that could affect the user participations.

[9]Measured as the number of days between the posting date and the date of the most recent comment.

Lastly, we cluster the standard errors at the post-date level to address the unobserved correlation in commenters' engagement time. Specifically, commenters who comment on the same post on the same day may share similar characteristics and sentiments. For example, they may be the follower of the poster and therefore see the post at the same period of time. Henceforce, they may also share the same sentiments towards the post. Clustering at the post-date level allows for arbitrary correlation in the error term within each post-date cluster, while assuming independence across clusters.

6 Empirical Findings

6.1 *Participation shift*: Relative decline of Non-mainland users' participation

In general, we find that geotagging in profiles leads to a substantial decline in participation among non-mainland users. Table 3 reports the estimated impacts of geotagging users' profiles on the participation of non-mainland users. These estimates represent the local average treatment effect of the policy, as obtained by comparing the average non-mainland participation 40 days before and after the policy implementation. Columns (1)–(3) report estimates based on the first-order comments, which are the comments that replied directly towards the post. Column (1) reports the linear specification without controls and fixed effects, which yields an estimated reduction of 19.4%. Column (2) report the estimation from the prefered specification, which adds controls, fixed effects, and drop one week of observations before and after the policy implementation date. It obtains a similar estimate of 19.1% with higher precision.

Column (3) report the estimation from a specification that adopt the quadratic time trend with all other components the same as Column (2). The estimated reduction inflates to 28.6%, which suggests that the higher-order polynomial may be sensitive to the variation within a narrow time bandwidth. Therefore, we consider the linear specification with controls and fixed effects as our main result.

Columns (4) and (5) report the estimates based on the reply towards the comments. These are the comments but are replied towards the first-order comments and used to evaluate the interaction between commenters. We do not find either specification yields significant and meaningful estimates, which

suggests that the policy does not significantly affect the interaction between commenters. One possible explanation is that these replies are folded in the comment section and are less visible to the public, so the effect of geotagging is less salient.

Table 3: Geotagging in Profile on Non-mainland Participation

	Comment's IP Location is Non-Mainland China				
	1st Order Comments			2nd Order Comments	
	(1)	(2)	(3)	(4)	(5)
IP in Profile	-0.00928*	-0.00861***	-0.0139*	0.00435	-0.00319
	(0.00506)	(0.00329)	(0.00789)	(0.0790)	(0.0129)
Observations	386,042	317,459	317,459	212,706	212,706
Pre-policy Ratio	0.0477	0.0486	0.0486	0.0634	0.0634
Controls	✗	✓	✓	✓	✓
DoW FEs	✗	✓	✓	✓	✓
Post-type FEs	✗	✓	✓	✓	✓
Poster FEs	✗	✓	✓	✓	✓
Linear Trend	✓	✓	✗	✓	✗
Donut Window	✗	✓	✓	✓	✓
Quadratic Trend	✗	✗	✓	✗	✓

Notes: This table reports the estimation results from an RDiT using the non-mainland comment dummy as the outcome. The bandwidth is set at 40 days, the number of days before the rolling-in of the next policy phase. The regression includes the set of controls defined in the previous section, as well as day-of-week and post-type fixed effects. Standard errors are clustered at the post-date level. Linear and Quadratic stand for the functional form of the time trends before and after the cutoff date. Donut stands for whether we dropped a week of observations before and after the cutoff date.

6.2 *Opinion shift*: Anti-foreign sentiments

Second, we show that profile geotagging exacerbates anti-foreign sentiments. We measure these sentiments using indicators of anti-foreign opinions expressed in comments. Column (1) of Table 4 presents the main result for political posts, which aligns with the model prediction. Specifically, the anti-foreign sentiment in comments increases substantially by 14.7% (0.0643/0.4372), with the estimate significant at the 10% level.

To decompose the drivers of this opinion shift, Equation 13 shows that the change in mean opinion, $\Delta Opinion$, on the platform consists of attitudinal changes among repeaters (users commenting both before and after the policy) and population shifts due to the policy.

$$
\begin{aligned}
\Delta Opinion &= \overbrace{O_1}^{\text{mean opinion after policy}} - \overbrace{O_0}^{\text{mean opinion before policy}} \\
&= \sum_{i \in P_1} \omega_{i,1} o_{i,1} - \sum_{i \in P_0} \omega_{i,0} o_{i,0} \\
&= \sum_{i \in P_1} \omega_{i,1} o_{i,1} - \sum_{i \in P_1} \omega_{i,1} o_{i,0} + \sum_{i \in P_1} \omega_{i,1} o_{i,0} - \sum_{i \in P_0} \omega_{i,0} o_{i,0} \\
&= \underbrace{\sum_{i \in P_1} \omega_{i,1} (o_{i,1} - o_{i,0})}_{\text{attitudinal change}} + \underbrace{\sum_{i \in P_1} \omega_{i,1} o_{i,0} - \sum_{i \in P_0} \omega_{i,0} o_{i,0}}_{\text{population shift}},
\end{aligned}
\tag{13}
$$

where $\omega_{i,t}$ for $t \in \{0, 1\}$ denotes the weight of individual i's opinion at time t, and $o_{i,t}$ for $t \in \{0, 1\}$ denotes individual i's opinion at time t.

In practice, we measure anti-foreign sentiments in each comment and assign equal weight to each observation for simplicity. We classify commenters into three types: (1) those who comment only before the policy (Exiters); (2) those who comment only after the policy (New-comers); (3) those who comment both before and after the policy (Repeaters). We then interact the main specification with indicators for each commenter type to separately identify the policy effects across groups. The baseline Geotagging coefficient represents the difference in anti-foreign sentiment between new-comers and exiters, reflecting the composition of users attracted to the platform by the policy.

According to Table 4 column (2), new-comers express substantially stronger anti-foreign sentiments. The Geotagging coefficient (β_1) is 0.079, indicating that new-comers are 18.4 percentage points more likely to express anti-foreign sentiments than exiters. In contrast, although repeaters express 8.2 percentage points more anti-foreign sentiment than exiters at baseline, the negative estimate on the Geotagging $\times$ Repeater interaction (β_2) offsets this baseline difference. A joint hypothesis test fails to reject the null that repeaters' opinions remain different from the exiters. This finding suggests that population composition shifts, rather than attitude change among existing users, drives the observed increase in anti-foreign opinions.

Table 4: Shift of Anti-foreign Sentiments in Comments Made to Political Posts

	Comment contains anti-foreign sentiment	
	(1)	(2)
β_1: Geotagging	0.0643*	0.0780**
	(0.0382)	(0.0396)
β_2: Repeater		0.0349*
		(0.0187)
β_3: Geotagging × Repeater		-0.0427
		(0.0277)
Observations	55,757	55,757
Pre-policy ratio	0.4372	0.4250
Joint Hypothesis Tests		
Repeaters: pre vs. post-geotagging		
$\beta_1 + \beta_3$		0.3996

Notes: This table presents RD estimates of the impact of profile geotagging on the anti-foreign sentiments expressed in the comments. The outcome variable is a dummy variable that indicates if the comment contains anti-foreign sentiment. The sample is restricted to the posts that contains at least one foreign entity. The specification contains a linear time trend and drops a one-week donut. We control post's basic metadata and sentiments toward specific locations. We also control the Day-of-Week, post type, and poster fixed effects. The standard errors are clustered at the post-date level.

Table D.4 shows the estimation for non-political posts. We find a similar pattern of opinion shift, but the estimates are noisier and less significant. The Geotagging coefficient is 0.054, equivalent to a 18.8% increase in anti-foreign sentiments. However, the estimate is not statistically significant at conventional levels. The interaction term between Geotagging and Repeater is negative but also insignificant, and the joint hypothesis test fails to reject the null that repeaters' opinions remain different from the exiters. Therefore, the patterns observed in political posts are less clear in non-political posts, but the point estimates are consistent with the model's prediction that population composition shifts drive the increase in anti-foreign sentiments.

6.3 *Group difference*

Group Difference Effect (Prediction 2) suggests that the labeling effect is amplified when posts exhibit larger media gaps between groups. We test this prediction by constructing a measure of sentiment bias and estimating the RD specification with interactions between this measure and the treatment variable.

Specifically, we use the sentiment gap measured by Equation 14 to evaluate group differences. We define the sentiment gap as the difference in posters' attitudes toward different locations. For example, when comparing Russia-Ukraine war posts, a poster with anti-foreign views may negatively describe Ukraine as a puppet state, while a pro-foreign poster may positively describe Ukrainian soldiers as heroes. Thus, the sentiment gap toward a specific location in event-related posts reflects the media gap across different groups' narratives.

We set the average sentiment toward location o expressed by non-mainland posters as a benchmark, denoted $\bar{\mu}^o_{NM}$. The sentiment gap on location o between post p and the non-mainland benchmark is then $\Delta^{o,p} = \mu^{o,p} - \bar{\mu}^o_{NM}$. We aggregate these differences via Euclidean distance across locations to create an overall media gap measure:

$$\Delta^p = \sqrt{\sum_o (\mu^{p,o} - \bar{\mu}^o_{NM})^2} \tag{14}$$

We focus on two case studies: Russia-Ukraine War and COVID-19 posts, where sentiment gaps between mainland and non-mainland posters' narratives are most pronounced. For Russia-Ukraine War posts, we select Russia, Ukraine, and USA as reference locations. For COVID-19 posts, we select China, USA, Beijing, and Shanghai. A poster's sentiment toward each location is measured as the average sentiment score across their posts mentioning that location, using the Likert scale from Section 4.2.2. We standardize Δ^p so that a one-unit increase represents one standard deviation of the sentiment gap.

Figure 6 presents the marginal effects of profile geotagging at different sentiment gap levels. These coefficients are derived from the linear-donut specification of Equation 12, interacting the treatment and time trend variables with poster-level sentiment gaps. We estimate effects only for mainland posters, treating non-mainland sentiment as the baseline. Figures 6a and 6b

support Prediction 1, showing that geotagging effects are negatively correlated with sentiment gaps—mainland posters with larger sentiment gaps experience the greatest reduction in non-mainland participation.

Specifically, when mainland posters' sentiment gaps fall below the mean, we find that the policy weakly increases relative non-mainland engagement. Conversely, reduction emerges among posters with above-average sentiment gaps. In Russia-Ukraine War posts, this reduction becomes marginally significant for posters exceeding the sentiment gap threshold. COVID-19 posts show substantial reductions but with noisier estimates. Overall, the negative correlation between policy effects and sentiment gaps aligns with our model predictions.

7 Alternative Channels

Next, we discuss three alternative channels that could have led to the decline in non-mainland users on the platform.

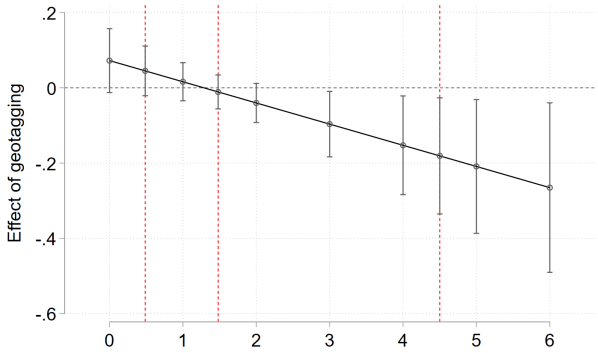
7.1 Strategic content moderation

Disclosing users' IP locations provides posters with information about the geographical distribution of their audience, enabling them to adjust content to align with follower preferences. For instance, a poster discovering that most followers are non-mainland Chinese might increase pro-foreign sentiment to retain engagement, or conversely, reduce it to appeal to a predominantly mainland audience. To assess whether such strategic adjustments occurred, we estimate specifications based on Equation 12 using post-date level characteristics as outcomes.

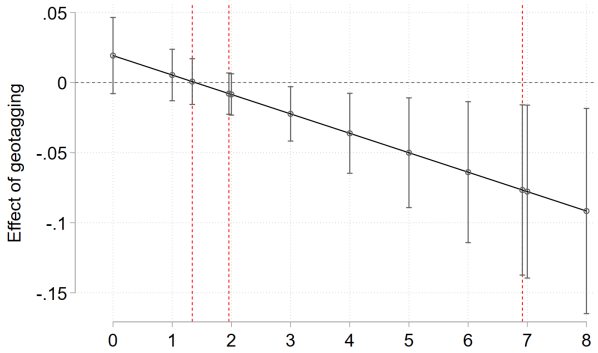
Figure E.15 shows that posters do not systematically change post categories, location mentions, or sentiments toward locations during the sample period[10]. Specifically, we measure the daily proportion of posts with certain characteristics. Panel (a) illustrates the policy's impact on post categories, Panel (b) on location mentions, and Panel (c) on sentiments toward locations. We find no statistically significant systematic changes in these features following policy implementation, suggesting that the policy is unlikely to alter the platform's information environment in the short term. This content

[10]This does not preclude the possibility that posters alter content or narratives over longer horizons.

Figure 6: Geotagging with Media Gap Heterogeneity



(a) Media Gap on Russia, Ukraine, China, and USA Covered in Russia-Ukraine War Posts



(b) Media Gap on China, USA, Beijing, and Shanghai Covered in COVID-19 Posts

Notes: The estimates plotted in both figures are generated by interacting the treatment and trend variables from Equation 12 with the media gap. The media gap in Figure (a) is the Euclidean distance between the mean sentiments expressed from posts written by non-mainland posters, and the sentiments expressed by the post written by a mainland poster in all COVID-19 posts towards the USA, Russia, and Ukraine. The media gap in Figure (b) is the Euclidean distance between the mean sentiments expressed from posts written by non-mainland posters, and the sentiments expressed by the post written by a mainland poster in all COVID-19 posts towards the USA, China, Beijing, and Shanghai. Locations are chosen based on the relative frequency. The figure plots the 90%-level confidence intervals.

continuity indicates that systematic changes in commenter behavior result

from commenters' direct responses to the policy, rather than posters strategically adjusting content to maximize engagement based on newly available geographical information.

7.2 Censorship

The reduction in non-mainland commenters could also result from targeted censorship[11]. Since non-mainland commenters tend to express pro-foreign sentiments, their identities' disclosure might trigger stronger censorship. Measuring censorship directly is notoriously difficult; the standard approach requires repeatedly scraping posts to detect censorship (King et al., 2013). However, we use historical data where censorship has already been applied.

To address this concern, we exploit a platform loophole to indirectly measure censorship. Sina Weibo employs *shadow censoring*, wherein censored comments become invisible but do not reduce the displayed comment count[12]. We can thus estimate shadow censoring by comparing displayed comment counts against our scraped comments.

Figure E.17 presents the policy's impact on shadow censoring across two implementation phases. We focus on posts with fewer than 500 comments, as our data repository only captured the latest 500 comments. Positive (negative) values indicate increased (decreased) shadow censoring following the policy. Figure E.17a shows that profile geotagging modestly affects shadow censoring levels. By contrast, Figure E.17b reveals substantially stronger shadow censoring after comment geotagging. This comparison suggests that while censorship may drive the decline during the comment geotagging phase, it is unlikely the primary mechanism during the profile geotagging phase.

Two limitations merit mention. First, our measure captures only within-post censoring. Detecting post-level censoring would require repeated real-time scraping. However, within-post and post-level censoring likely correlate strongly, so our measure should predict post-level censoring. We postulate that post-level censoring did not systematically vary due to the policy.

Second, comment count discrepancies may stem from other sources. For instance, Weibo occasionally inflates counts for paid promotions, but there is no clear reason to adjust promotion strategies based on geotagging. As

[11] The implicit platform or government coercion generated by the IP disclosure is unlikely a main source of self-censorship, as Weibo requires real IDs for account registration and continuation since 2011.

[12] In contrast, the comment count decreases when posters delete comments.

both panels of Figure E.17 show, entertainment and advertising posts—where promotions typically occur—exhibit no significant censoring effects from the policy.

7.3 Bot elimination

Another potential channel is that identity disclosure may deter foreign bots and encourage the deletion of non-mainland accounts following policy implementation. To assess this mechanism, we employ a pre-trained algorithm to predict the likelihood that first-order commenters are bots based on account characteristics[13].

Table E.1 reports estimates from Equation 12 using predicted bot scores and account suspension or deactivation as outcomes. Columns (1) and (3) show that geotagging produces no statistically significant changes in either measure. The policy coefficients are small and precisely estimated near zero for both bot scores and account deactivations, indicating that the policy did not eliminate automated accounts or increase account attrition immediately after implementation.

Columns (2) and (4) examine heterogeneous effects for non-mainland users. Notably, column (2) reveals that non-mainland commenters are actually less likely to be classified as bots before policy implementation. The interaction coefficients for both outcomes are insignificant, further suggesting that the decline in non-mainland commenters does not stem from bot elimination or increased enforcement-related account attrition following the policy.

8 Conclusion

This paper examines how reducing anonymity reshapes participation and opinion expression on social media. Using Sina Weibo's IP disclosure policy as a quasi-natural experiment, we demonstrate that identity disclosure significantly reduces participation among geographically identifiable minority groups and alters the composition of expressed opinions. The observed increase in anti-foreign sentiment stems primarily from compositional changes in the user base rather than shifts in individual expressions. This effect

[13]The bot detector achieves 97% accuracy and is available at: https://github.com/mengxiao2000/pybotfinder.

is more pronounced where underlying intergroup media-bias gaps are larger, supporting our framework: identity disclosure increases the social cost of participation for minority users when cross-group disagreement is more salient. Our findings reveal a critical consequence of anonymity regulation: identity disclosure not only increases transparency but also restructures the participant set and the distribution of visible viewpoints.

We highlight two important extensions to this study. First, our evidence does not establish whether anonymity curtailment makes discourse less representative of public opinion overall. We document compositional shifts but cannot determine whether these move the platform toward or away from population-level representativeness without additional evidence. Nevertheless, for policymakers relying on social media signals, compositional shifts carry important implications: they reshape perceptions of who participates, which views are visible, and how opinion is distributed. Broadly, our findings suggest that modest identity disclosure can substantially reconfigure intergroup communication in digital public spheres.

Second, our evidence does not capture whether identity disclosure has persistent long-run consequences. Over time, posters, platforms, and agencies may strategically respond to the geographical information revealed through commenters' geotags, altering content, shifting sentiment, and changing censorship patterns. Identifying long-run effects would require structural modeling involving multiple strategic actors. Additionally, complementary policies designed to combat misinformation and bots through IP disclosure will likely further affect the representativeness of social media data.the long-run consequences of anonymity curtailment. Future research should explore how identity disclosure interacts with other platform policies and user behaviors to shape the evolution of online discourse and its representativeness over time.

References

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–676.

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

Andreoni, J. (1995). Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics*, *110*(1), 1–21.

Bazzi, S., Gaduh, A., Rothenberg, A. D., & Wong, M. (2019). Unity in diversity? how intergroup contact can foster nation building. *American Economic Review*, *109*(11), 3978–4025.

Braghieri, L., Levy, R., & Makarin, A. (2022). Social media and mental health. *American Economic Review*, *112*(11), 3660–3693.

Buntaine, M. T., Greenstone, M., He, G., Liu, M., Wang, S., & Zhang, B. (2024). Does the squeaky wheel get more grease? the direct and indirect effects of citizen participation on environmental governance in china. *American Economic Review*, *114*(3), 815–850.

Bursztyn, L., Callen, M., Ferman, B., Gulzar, S., Hasanain, A., & Yuchtman, N. (2020). Political identity: Experimental evidence on anti-americanism in pakistan. *Journal of the European Economic Association*, *18*(5), 2532–2560.

Bursztyn, L., Egorov, G., Enikolopov, R., & Petrova, M. (2019). *Social media and xenophobia: Evidence from russia* (tech. rep.). National Bureau of Economic Research.

Bursztyn, L., Egorov, G., Haaland, I., Rao, A., & Roth, C. (2023). Justifying dissent. *The Quarterly Journal of Economics*, *138*(3), 1403–1451.

Corno, L., La Ferrara, E., & Burns, J. (2022). Interaction, stereotypes, and performance: Evidence from south africa. *American Economic Review*, *112*(12), 3848–3875.

Dahl, G. B., Kotsadam, A., & Rooth, D.-O. (2021). Does integration change gender attitudes? the effect of randomly assigning women to traditionally male teams. *The Quarterly journal of economics*, *136*(2), 987–1030.

Ederer, F., Pinkham-Goldsmith, P., & Jensen, K. (2023). Anonymity and identity online.

Enikolopov, R., Makarin, A., & Petrova, M. (2020). Social media and protest participation: Evidence from russia. *Econometrica*, *88*(4), 1479–1514.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447–456.

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from# brexit and# uselection. *European Economic Review*, *136*, 103772.

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in china allows government criticism but silences collective expression. *American political science Review*, *107*(2), 326–343.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, *111*(3), 831–870.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review*, *111*(6), 1807–1844.

Mosquera, R., Odunowo, M., McNamara, T., Guo, X., & Petrie, R. (2020). The economic effects of facebook. *Experimental Economics*, *23*, 575–602.

Qin, B., Strömberg, D., & Wu, Y. (2017). Why does china allow freer social media? protests versus surveillance and propaganda. *Journal of Economic Perspectives*, *31*(1), 117–140.

Rao, G. (2019). Familiarity does not breed contempt: Generosity, discrimination, and diversity in delhi schools. *American Economic Review*, *109*(3), 774–809.

Schindler, D., & Westcott, M. (2021). Shocking racial attitudes: Black gis in europe. *The Review of Economic Studies*, *88*(1), 489–520.

Wu, A. H. (2020). Gender bias among professionals: An identity-based interpretation. *Review of Economics and Statistics*, *102*(5), 867–880.

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, 75–78.

Yang, L. Y., & Xu, Y. (2025). User location disclosure fails to deter overseas criticism but amplifies regional divisions on chinese social media. *arXiv preprint arXiv:2507.03238*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

Appendices

A Omitted Proofs

A.1 Proof of *Proposition 1*

By (7), (8) and (11), $\underline{v}_\emptyset < \bar{v}_\emptyset$ if $\underline{v}_\emptyset$ and $\bar{v}_\emptyset$ exist (and the inequality holds strictly if the media slant gap $\Delta > 0$). We show the existence and uniqueness of $\underline{v}_\emptyset$; the existence and uniqueness of $\bar{v}_\emptyset$ can be established by symmetry, which we ignore in the proof.

Define $F(v) = f_1(v) - f_2(v)$ where

$$f_1(v) = v + \frac{\gamma}{\beta}, \tag{A.1}$$

$$f_2(v) = -\frac{\psi(\Delta)}{\beta}\frac{1-\lambda}{(1-\lambda)+\lambda\hat{\Phi}(v)}, \tag{A.2}$$

and $\hat{\Phi}(v) = \Phi(\frac{v-\mu_L}{\sigma})/\Phi(\frac{v-\mu_F}{\sigma}) \in (0,1]$. Note that $-\frac{\psi(\Delta)}{\beta} \leq f_2(v) \leq -\frac{\psi(\Delta)}{\beta}(1-\lambda)$ for any $v \in \mathbb{R}$. Therefore, $\lim_{v\to-\infty} F(v) = -\infty, \lim_{v\to\infty} F(v) = \infty$. Since $F(\cdot)$ is continuous, by *Intermediate Value Theorem*, there exists an $\underline{v}_\emptyset \in \mathbb{R}$ such that $F(\underline{v}_\emptyset) = 0$—and thus, (8) and (10) hold simultaneously.

Furthermore, since for any $v \in \mathbb{R}$,

$$\lim_{\lambda\to1} F'(v) = 1 - \lim_{\lambda\to1}\frac{\psi(\Delta)}{\beta}\frac{(1-\lambda)\lambda\hat{\Phi}'(v)}{[(1-\lambda)+\lambda\hat{\Phi}(v)]^2} = 1 > 0,$$

$$\lim_{\lambda\to0} F'(v) = 1 - \lim_{\lambda\to0}\frac{\psi(\Delta)}{\beta}\frac{(1-\lambda)\lambda\hat{\Phi}'(v)}{[(1-\lambda)+\lambda\hat{\Phi}(v)]^2} = 1 > 0,$$

there exists $(\underline{\lambda}, \bar{\lambda}) \in (0,1)^2$ such that for any $\lambda \in (0, \underline{\lambda}] \cup [\bar{\lambda}, 1)$, $F'(v) > 0$, and thus the solution to $F(v) = 0$ is uniquely given by $\underline{v}_\emptyset = F^{-1}(0)$.

A.2 Proof of Prediction 1

Suppose that xenophobia sanctions are sufficiently high such that $\psi(\Delta) \geq \beta(V - \mu_F)$. Relative to the benchmark case, the distribution of foreign users' opinion shifts left by $V - \mu_F$, while the commenting cutoffs shift left simultaneously by $\psi(\Delta)/\beta$. Since $\psi(\Delta)/\beta \geq V - \mu_F$, the cutoff shift weakly

exceeds the shift in the opinion distribution. As a result, among foreign users, the probability of posting criticism decreases, whereas the probability of posting support increases. Moreover, the reduction in negative commenting outweighs the increase in positive commenting, resulting in a net decline in participation by the foreign users on the platform.

Further, suppose the local group significantly dominates the foreign group in size. By (9) and (10), we have

$$\lim_{\lambda \to 1} \Pr(i \in F | d_i = 1, \mathcal{P} = \emptyset) = \lim_{\lambda \to 1} \Pr(i \in F | d_i = -1, \mathcal{P} = \emptyset) = 0.$$

By (7) and (8), it follows that

$$\lim_{\lambda \to 1} \bar{v}_{\emptyset} = \bar{v}, \qquad \lim_{\lambda \to 1} \underline{v}_{\emptyset} = -\bar{v}.$$

Therefore, when the share of local users, $\lambda \in (0, 1)$, is sufficiently high, all users' commenting choices under the anonymous policy resemble the first-best benchmark case without media bias.

Under the geotagging policy, however, the commenting cutoffs for the dominant local group remain approximately unchanged, while those for the minority foreign group shift leftward.

B VPN Usage in China

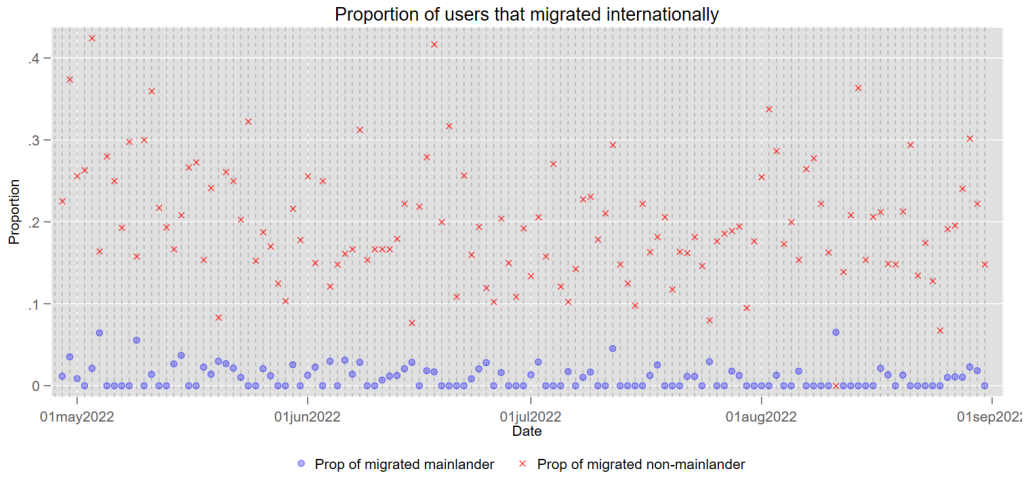
Due to the expansion of the Great Firewall project, through which the Chinese government barred mainland Chinese access to a large set of foreign platforms, including YouTube, Facebook, X (formerly Twitter), and many other highly visited websites, Chinese netizens anecdotally increased their use of VPN to bypass the restriction. In the following discussion, we argue that the usage or manipulation of VPN should not influence our key results.

First, most contemporary VPN applications have configured with *Dynamic Split Tunneling* in default, automating users to selectively route certain types of traffic through the VPN while letting other traffic go directly to the internet. In our case, visiting domestic websites such as Sina Weibo will route the traffic directly to the website without first changing the IP address. Therefore, the IP location exhibited on Weibo should be mostly correct unless the users intentionally terminate the split tunneling and disguise the trace entirely through VPN.

Second, we take a small sample of random commenters who made the comment between April 28 and August 31, 2022. This sample includes 12,580 mainland users and 8,154 non-mainland users. The location of these commenters is identified based on the tag in their comments. We obtain a second source of their IP locations by scraping their personal Weibo profile page in early 2023. As the IP location tagged in the profile page used the most frequently used IP address in the previous month, it is less likely to be influenced by the VPN. Figure B.1 shows the proportion of comments with an IP location in the profile that differs internationally from the IP location in the comment. In general, we find 17.2% of commenters have an IP location in the comment that differs internationally from their profiles. Among these, only 0.93% are mainlanders according to comments but non-mainlanders according to the profile, but 20.07% are non-mainlanders according to comments but mainlanders according to the profile. This comparison suggests that many mainland users used VPN and were identified as non-mainlanders when they posted the comment. On the other hand, the proportion of non-mainland users who used VPN to be identified as mainlanders is much less significant. Physical migration is unlikely to drive the result as China still strictly restricted international travel when we obtained this data set.

In addition, Figure B.1 provides more suggestive evidence against the potential of VPN manipulation as a driving force. We find the proportion of migration does not substantially change over time. Had there been a grow-

Figure B.1: "International Migration" of IP Location



ing hostility against the non-mainlanders, earlier users, who were influenced longer by the policy, were more likely to disguise or return as a mainlander. However, we observed neither earlier non-mainland commenters were more likely to migrate than later non-mainland commenters, nor were later mainland commenters more likely to migrate than earlier mainland commenters.

C Data Collection and Classification

C.1 Topic-oriented process

Ideally, we would randomly select posters or posts from a comprehensive roster of influential accounts. However, no such roster exists. Instead, we developed a topic-oriented selection process. First, we collected important news during our sample period and extracted relevant keywords. As summarized in Table D.1, we derived 30 keywords representing the most influential topics on Chinese social media during this period. For example, "Human Trafficking" corresponds to a high-profile case in Jiangsu province, while "Taiwan" refers to Nancy Pelosi's visit in August 2022.

Next, we retrieved Weibo posts containing these terms and collected all posts that received more than 100 likes during the sample period. We then obtained a list of 7,499 posters who authored these posts. We excluded approximately 2,000 posters that (1) were currently associated with government entities or news agencies, (2) operated celebrity or fan accounts, or (3) had deactivated accounts, to minimize potential platform or government censorship while maximizing content diversity. Due to the computational cost, we ultimately sampled 10% of the remaining posters, yielding a final set of 509 posters. This selection procedure identified a set of influential posters on the platform whose content was widely circulated and commented on by users with diverse demographics.

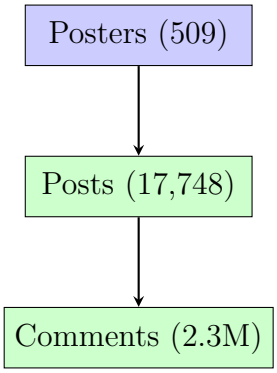
Following the selection of posters, we scraped all posts written by these posters during the sample period. For each month's top 5 most-liked posts, we scraped comments in their default order[14]. We collected both post and comment metadata, including the number of likes, reposts, comments, and geographic location when available.

As illustrated in Figure C.1, we obtained 17,748 posts from 509 posters and 2.3 million comments on these posts. We collected essential metadata indicating the popularity of posts and comments, such as the number of likes, reposts, and comments received. We also obtained geographic location information for posters and commenters when available[15].

[14]Weibo typically displays up to 500 comments, ordered by popularity and timestamp

[15]For comments posted before the comment-level disclosure system launched on April 28, 2022, we extracted geotags from user profiles to obtain IP location data. This method resulted in approximately 30% of pre-April 28 commenters lacking IP location information, as these users may have deactivated their accounts or become inactive by the time of our

Figure C.1: Data Structure



C.2 Text Analysis

Posts: Topics, Key entity and Attitude

We extracted three types of information from the posts. First, we obtained the topics of the posts, which are broadly categorized into politics, social news, economy, entertainment, daily, sports, knowledge, and advertisement. Each post can cover multiple topics. For example, a post on the COVID lockdown could expand politics, social news, and daily. In preparation for case studies, we also further classify whether the post is specific to COVID-19 and the Russia-Ukraine War.

Second, we extracted the key entity from the posts. The key entities are defined by the key person, item, location, or events that happened in the posts. For example, the key entities for a post about the Russia-Ukraine War could be Russia, Ukraine, and War.

We then label the poster's attitude towards each of the key entities on a 5-point Likert Scale from -2 (NEGATIVE) to 2 (POSITIVE) with 0 as NEUTRAL. For example, when a poster describes Russia as the invader in a post, we may label the poster's sentiment towards Russia in this post as NEGATIVE; on the other hand, if the poster described Russia's military action as resistance towards NATO's expansion, we may label the poster's sentiment towards Russia as NEUTRAL (0), LIKELY POSITIVE (1), or POSITIVE (2), depending on the context.

To sum up, we extract three sets of information from the posts: (1) Topics

---

data collection

of the post; (2) A vector of the post's key objects; (3) A vector of the poster's sentiment towards each key object.

Sentiments of comments

We extracted three types of sentiments from the comments. First, we evaluated whether the comment expressed approval or disapproval of the post. For example, suppose a poster describes Russia as the invader, but a comment addresses the tension between Russia's national safety and the expansion of NATO. In that case, we may label the comment with -1 as disapproval.

Second, specific to the interest of this paper, we evaluated if the comment expressed xenophobic sentiments. To enhance the relevance, we only evaluated these sentiments for comments on posts related to politics, economy, and social news, with at least one foreign entity identified as the key object. We measure the anti-foreign using the negative sentiment towards Western countries, ideologies, and institutions.

Last of all, we categorize the comment replies. Comment replies are generally considered part of comments, but were explicitly made to a comment of the post. We label the sentiment of these replies and specifically explore whether the comment replies contain geographical stigma or stereotyping to express their dissent to the comment. This strategy may become more common on social media when the commentator's location is publicly revealed. For example, a foreign commenter who advocated for COVID-19 reopening in China might encounter geographical stereotyping as most Chinese supported the lockdown at the beginning of 2022.

To sum up, we obtained three sets of information from the comments: (1) Approval or disapproval towards the post; (2) Xenophobic or anti-Western sentiments; (3) Geographical stereotyping.

Text Analysis with Large Language Models

Given the scale of the data, we used large language models (LLMs) for textual analysis. LLMs are neural language models pre-trained on massive text corpora, which equips them with broad linguistic and world knowledge for tasks that require natural language understanding (Xiao et al., 2023; Ziems et al., 2023). Unlike traditional machine learning–based approaches that require substantial training datasets, LLMs can be steered with natural-language instructions and a small number of demonstrations.

In our task, comment- and reply-level tasks where meaning often depends on informal language and the relationship to the parent post requires strong contextual understanding, which is limited in today's LLMs. To obtain high-quality results, we adopted a multi-stage pipeline: (1) extracting key claims from each comment or reply using an LLM, (2) disambiguating internet-specific political slang with a custom dictionary, and (3) assigning labels using a codebook-based prompting approach.

For claim extraction, we used GPT-4o to identify the key claims in each comment or reply and appended them to the original text, distilling lengthy or ambiguous content into explicit propositional statements for the downstream coding model. For slang disambiguation, we curated a political slang dictionary mapping internet-specific political language on Chinese social media, where euphemisms, homophones, and neologisms are commonly used to circumvent content moderation or signal in-group membership, to standard-language equivalents. Together, these preprocessing steps reduced the inferential burden on the coding models and improved label consistency.

For codebook-based coding, we adapted the approach in Xiao et al. (2023), combining LLMs with pre-developed codebooks to assign each post or comment a pre-defined code. Each codebook specifies a name, description, and examples for every code; we included five examples per code as few-shot demonstrations. We randomly sampled 1.3–20% of the data for each task and had expert coders assign labels, then iteratively refined the prompt until it reached adequate inter-rater agreement with experts. The remaining data were coded with the finalized prompt. To minimize stochasticity, we set the model temperature[16] to 0, repeated the process three times, and determined final labels by majority vote. We used ChatGLM-9b for post-level analyses and GPT-4o and GPT-5 for comment- and reply-level analyses requiring deeper contextual reasoning.

[16]Model temperature controls output randomness, ranging from 0 to 1; lower values produce more deterministic output.

D Tables

Table D.1: Keywords about China in Between January 1 and August 31, 2022

	Keywords
Technology	New Energy, Space Travel, BYD, Tesla, Data Leak, Microchip
Economy	Exchange Rate, GDP, Mortgage, Job Market, Stock Market
Entertainment	Winter Olympic, Short Videos, User-Generated Media
Social Affairs	COVID-19, Population, Birth Rate, Human Trafficking, Scam, Plagiarism, Air-crash Gaokao, PKU, Study Abroad
Politics	Corruption, Taiwan, Hong Kong, Ukraine, Xinjiang, Territory

Notes: The keywords are translated from Chinese. BYD represents the biggest electronic car manufacturer in China; Gaokao represents the college entrance exam of China; PKU represents Peking University.

Table D.2: Summary of LLM Classifications

level	label	model
post	category	ChatGLM-9b
post	location	ChatGLM-9b
post	sentiment towards location	ChatGLM-9b
comment	anti-western	GPT-4o
reply	agree/disagree	GPT-5
reply	approach of agree/disagree	GPT-5

Table D.3: Topics of Posts

Topic	Daily	Politics	Social	Econ	Entertain	Sports	Knowledge	Ad
N	4.879	6,426	9,268	1,437	4,417	506	1,944	664
%	27.49	36.2	52.2	8.1	24.9	2.9	11.0	3.7

Notes: Post topics are identified separately by ChatGLM-9b. Therefore, many posts overlap several topics. For example, 5,007 posts are identified as politics and social affairs, and 868 are identified as social affairs and economic issues.

Table D.4: Shift of Anti-foreign Sentiments in Comments Made to Non-Political Posts

	Comment contains anti-foreign sentiment	
	(1)	(2)
β_1: Geotagging	0.0537	0.0502
	(0.0328)	(0.0345)
β_2: Repeater		-0.00540
		(0.0268)
β_3:Geotagging × Repeater		-0.0427
		(0.0277)
Observations	44,592	44,592
Pre-policy Ratio	0.2852	0.2758
Joint Hypothesis Tests		
Repeaters: pre vs. post-geotagging		
$\beta_1 + \beta_3$		0.2209

Notes: This table presents RD estimates of the impact of profile geotagging on the anti-foreign sentiments expressed in the comments. The outcome variable is a dummy variable that indicates if the comment contains anti-foreign sentiment. The sample is restricted to the posts that contains at least one foreign entity. The specification contains a linear time trend and drops a one-week donut. We control post's basic metadata and sentiments toward specific locations. We also control the Day-of-Week, post type, and poster fixed effects. The standard errors are clustered at the post-date level.

E Figures

Figure E.1: The prompt structure incorporates meta-prompt tuning and in-context learning. Content in [] was replaced by codebook and data when performing the deductive coding task. When analyzing Chinese data, we translated the prompt into Chinese.

Text Analysis Prompt Structure

Imagine you are an expert in qualitative analysis. Your task is to code the text into one code. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Coding Steps:

1. Read the text thoroughly to understand its content and context.
2. Review the codebook to understand the codes and their definitions. The codebook serves as a guide for coding the text consistently and systematically.
3. Assign codes from the codebook to input text. This process involves matching the content of the text to the codes and their definitions in the codebook. It's essential to be consistent and systematic in this step to ensure the reliability of the analysis.
4. Assign the final label from one of the [CODE-1, CODE-2, ..., CODE-n]
5. Review the coded text and refine the coding if necessary.

Classify the [text] into one of the [CODE-1, CODE-2, ..., CODE-n] Based on the following codebook that includes the description of each code and a few examples. Note that the code is only for the [text], not the context, but the context should be considered when making the decision.

Codebook: CODE: [CODE-1]; DESCRIPTION: [CODE-1-DESCRIPTION]; EXAMPLES: [CODE-1-EXAMPLES]. CODE: [CODE-2]; DESCRIPTION: [CODE-2-DESCRIPTION]; EXAMPLES: [CODE-2-EXAMPLES]. ... CODE: [CODE-n]; DESCRIPTION: [CODE-n-DESCRIPTION]; EXAMPLES: [CODE-n-EXAMPLES].

Context: [CONTEXT]

Text: [TEXT]

Choose from the following candidates: [CODE-1, CODE-2, ..., CODE-n]

Figure E.2: Examples of IP Tag in Profile and in Comment

(a) In Profile

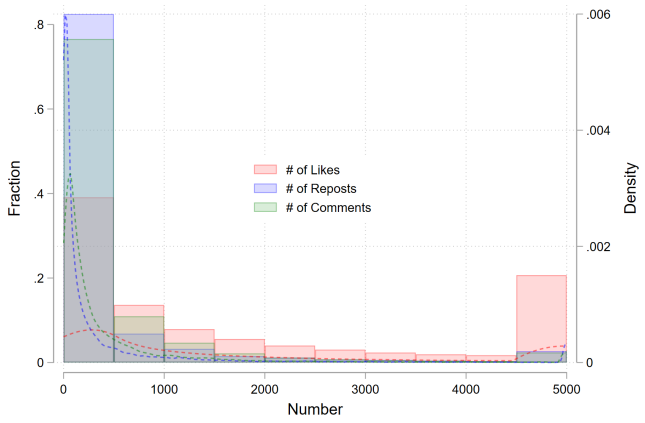


(b) In Comment



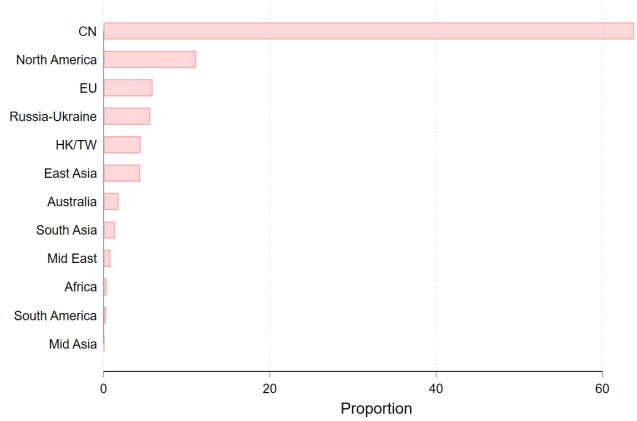
Notes: Figure (a) shows the profile geotagging in the red circle. The geotag shows the user is currently posting Weibo from Jiangsu province, China. The right panel of Figure (b) shows the comment geotagging. The post was posted close to the midnight of April 28, 2022. Comments made before April 28 were not geotagged, but comments made after April 28 were tagged immediately.

Figure E.3: The Distribution of # of Likes, Reposts, and Comments



Note: This figure summarizes the distribution of # of Likes, Reposts, and Comments received by the posts. The total number is capped at 5000, whereas the average number of Likes is 7213, leading to the distribution of Likes stacked at the right tail.

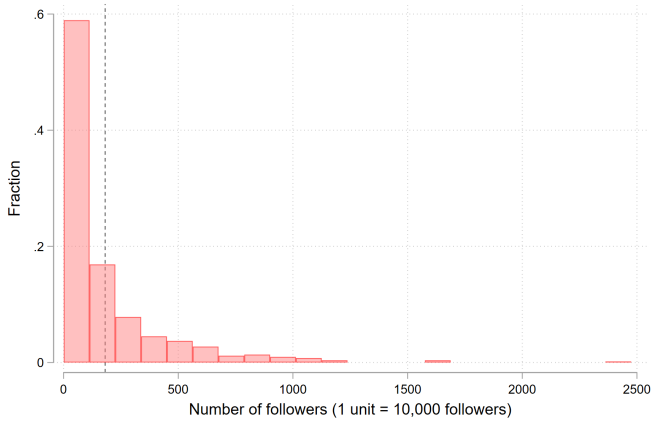
Figure E.4: The Distribution of Countries/Regions Mentioned in the Post



Note: This figure summarizes the distribution of countries and/or regions mentioned in the posts. For example, around 16% posts mentioned countries from North America in the posts.

Figure E.5: Poster's Basic Statistics

(a) The Distribution of # of Followers



(b) The Poster's IP-Location

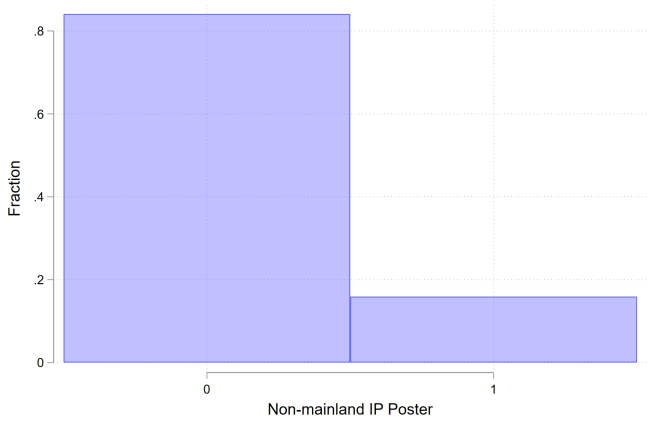
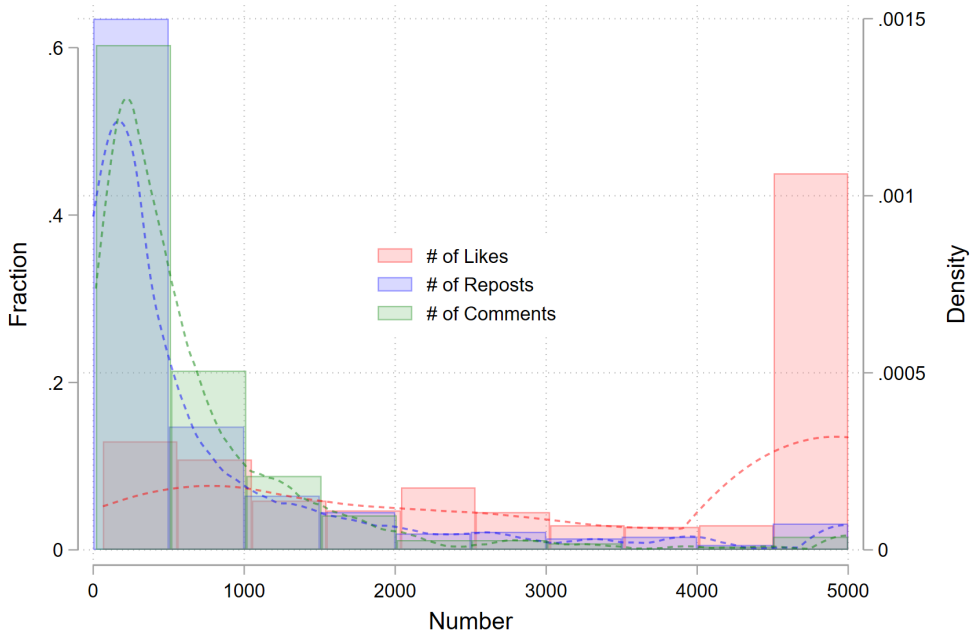


Figure E.6: The Distribution of # of Likes, Reposts, and Comments



Note: This figure summarizes the distribution of average # of Likes, Reposts, and Comments received by posters' posts. The total number is capped at 5000, whereas the average number of Likes is 9050, leading to the distribution of Likes stacked at the right tail.

Figure E.7: Mean Sentiments Toward Locations Mentioned in the Posts

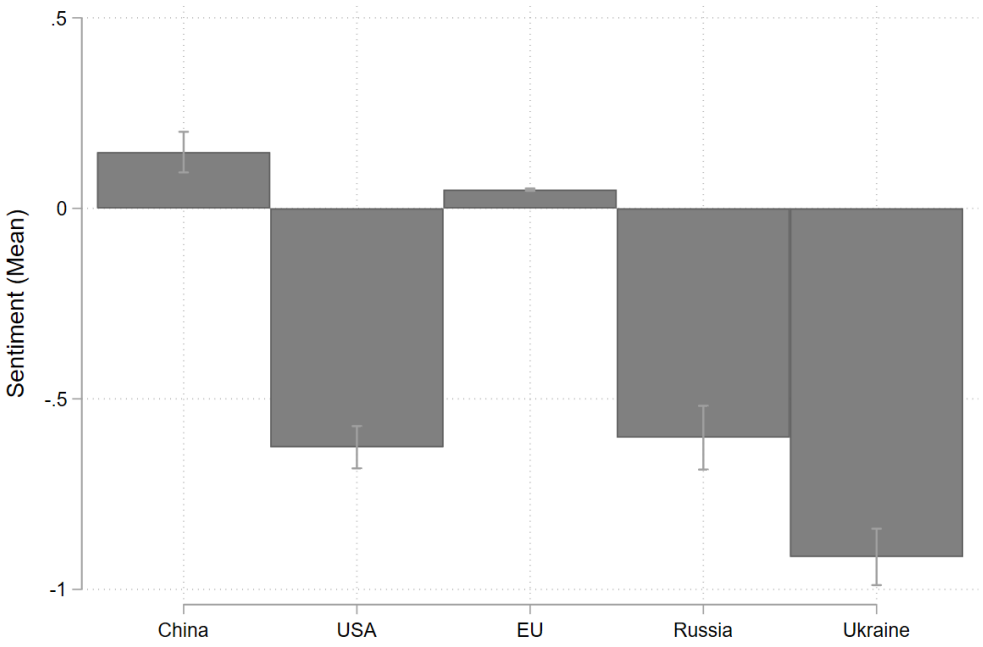
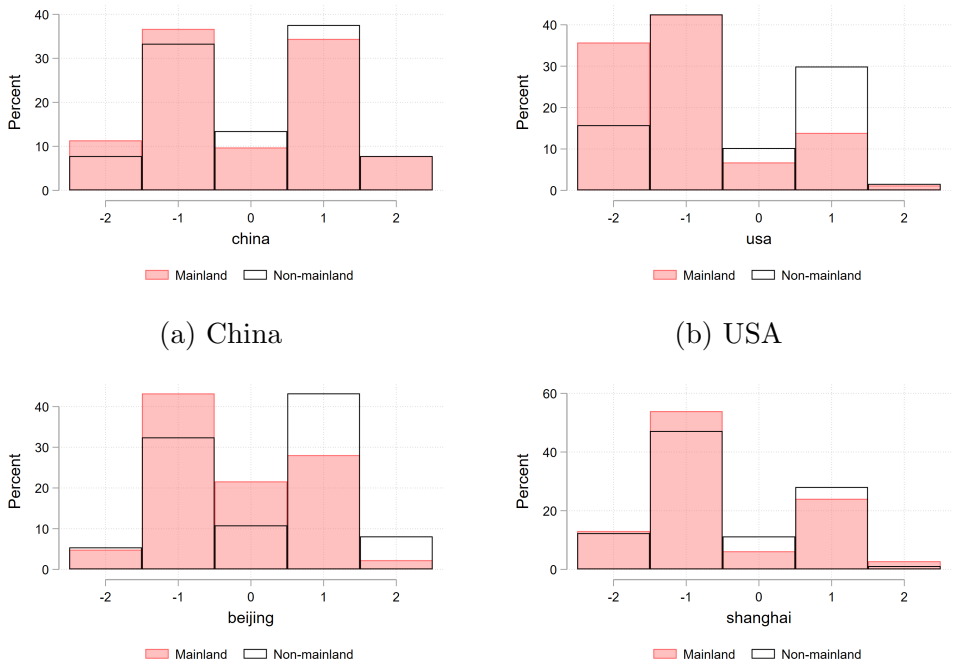


Figure E.8: Sentimental Differences in COVID-19 Posts

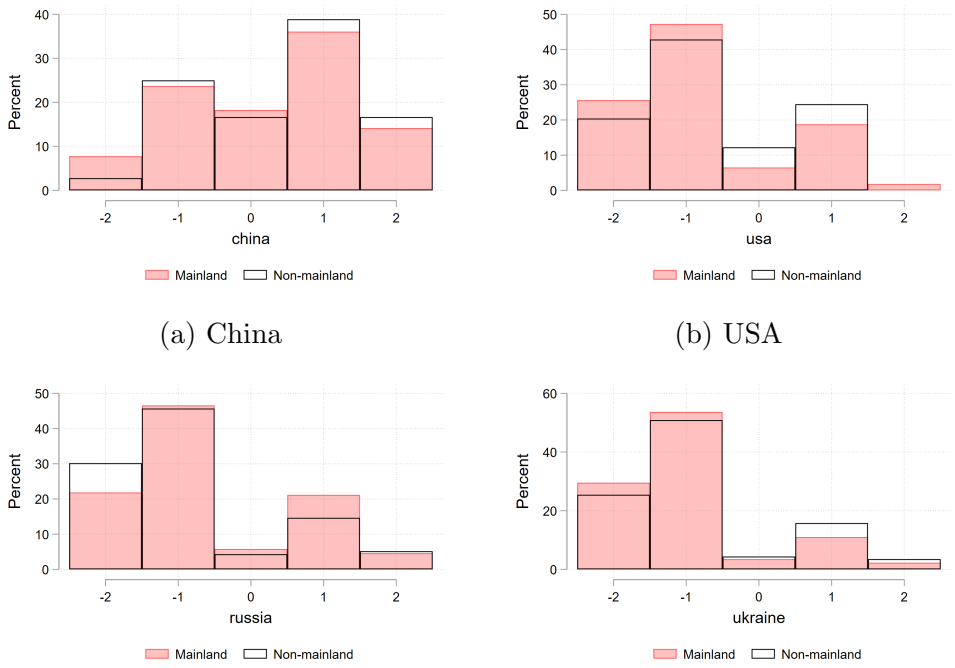


(a) China (b) USA

(c) Beijing (d) Shanghai

Notes: All four figures plot the proportion of sentiments, ranging from -2 (very negative) to 2 (very positive), towards four different locations. The LLM first identifies the location from the post and then evaluates the poster's sentiment towards the location based on the post's content.

Figure E.9: Sentimental Differences in Russia-Ukraine War Posts



(a) China

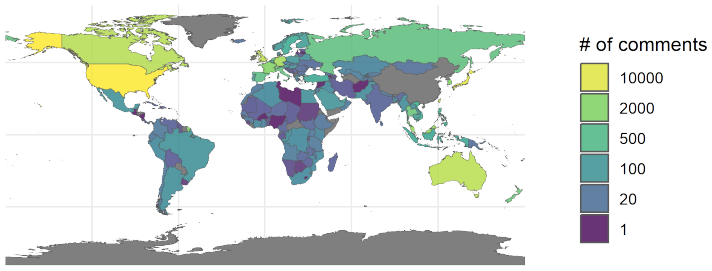
(b) USA

(c) Russia

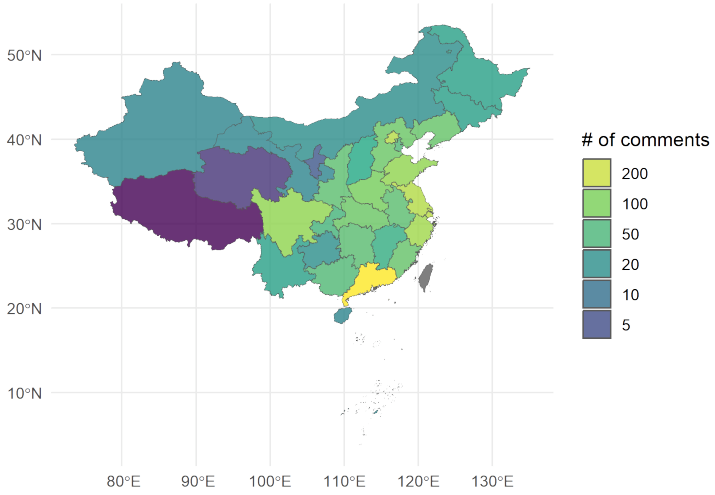
(d) Ukraine

Notes: All four figures plot the proportion of sentiments, ranging from -2 (very negative) to 2 (very positive), towards four different locations. The LLM first identifies the location from the post and then evaluates the poster's sentiment towards the location based on the post's content.

Figure E.10: The Number of Comments Made According to the IP Location



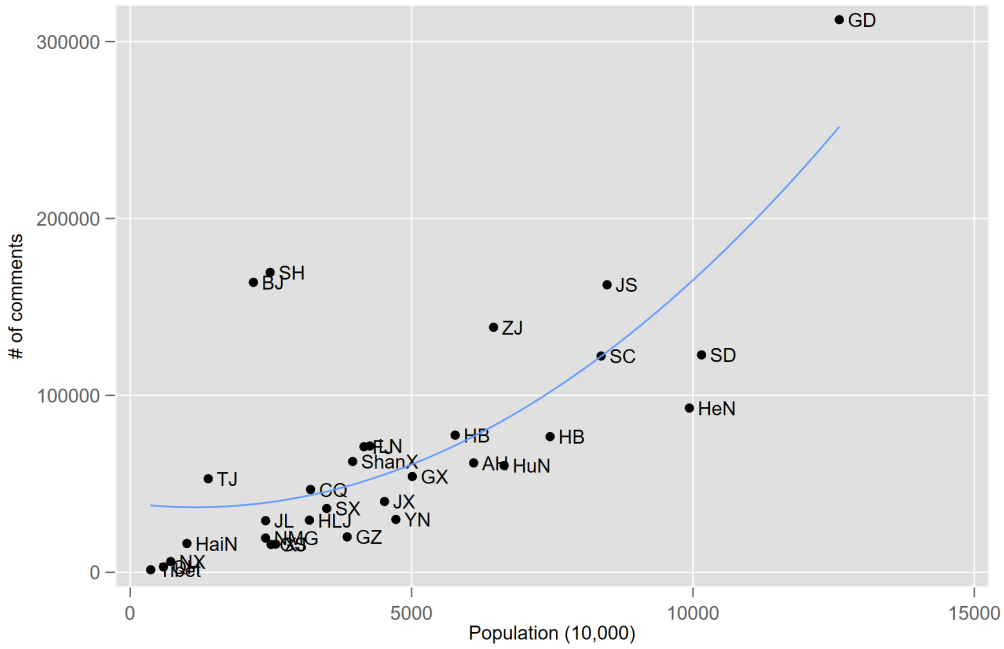
(a) Non-mainland



(b) Mainland

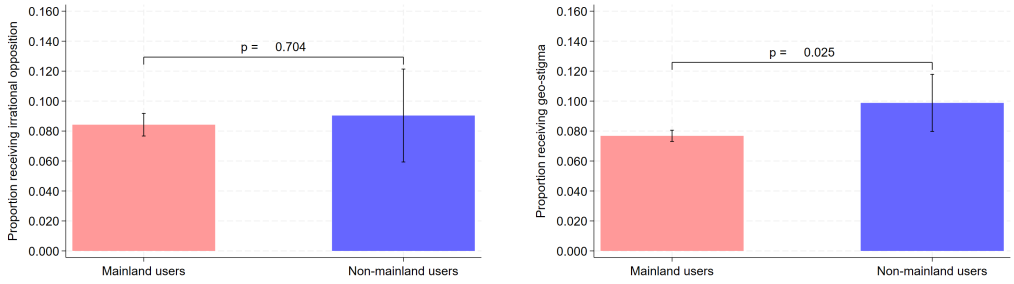
Note: One unit of comment equals to 10,000 of comments.

Figure E.11: Number of Comments vs. Population



Note: This figure shows the number of comments from each Chinese province or municipality is highly correlated with the true population of 2022. The only two exceptions are Beijing and Shanghai, which are noted as BJ and SH on the middle left of the plot. These two exceptions suggest a substantially higher proportion of social media users in major Chinese cities.

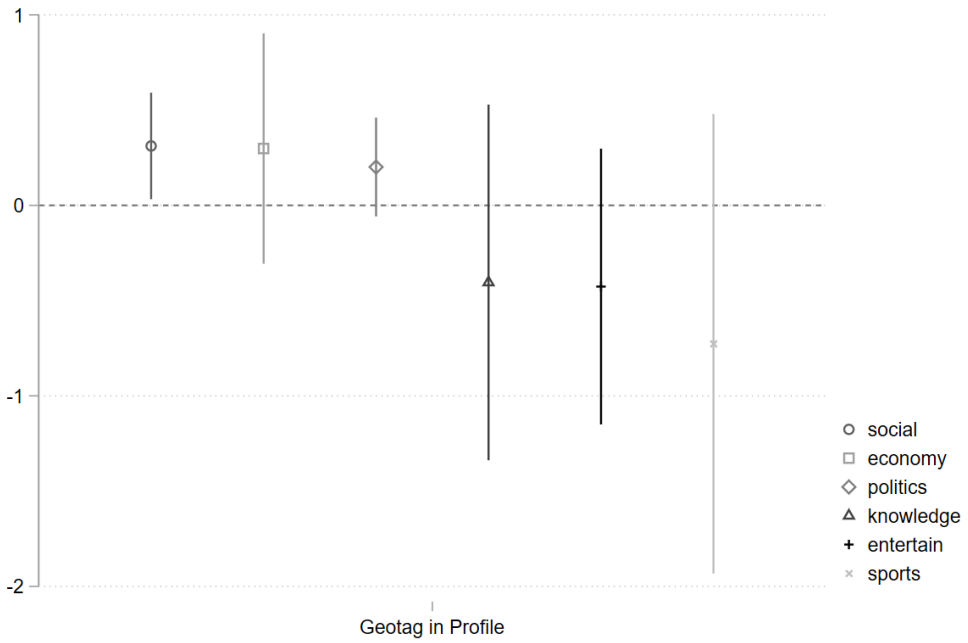
Figure E.12: Irrational Opposition Received by Mainland vs. Non-mainland Disapproval Comments



(a) Pre IP-disclosure (b) Post IP-disclosure

Notes: These two figures show the comparisons between the irrational oppositional replies received by the unfavorable comments between mainland commenters and non-mainland commenters. A reply is classified as irrational opposition if it contains geo-stigma, emotional opposition, or uncivilized vocabularies. Error bars represent 95% confidence intervals around mean sentiment estimates.

Figure E.13: Policy Effects on Anti-Western Sentiments in Different Topics



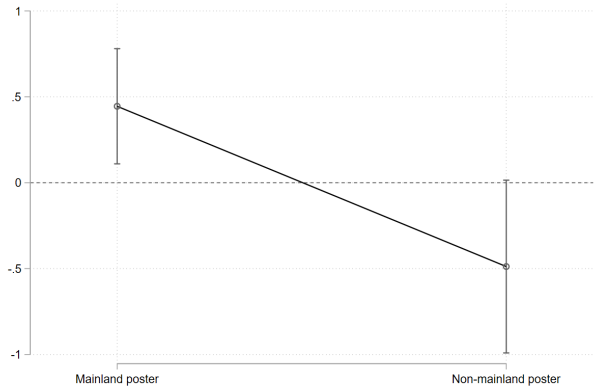
Notes: The coefficients are obtained via anti-Western donut RD regression with quadratic time trends and sample restricted to certain topics.

Table E.1: Policy's Impacts on Predicted Bot Scores and Account Deactivation

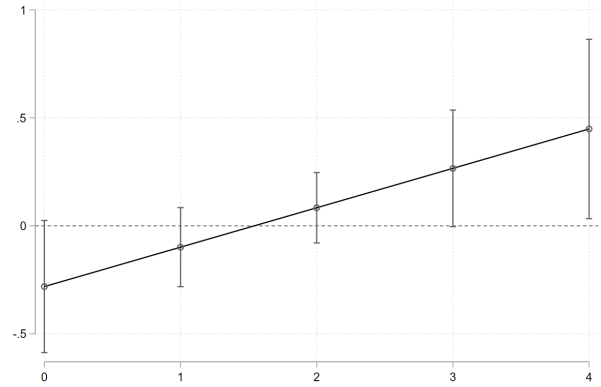
	Bot Score		Account deactivation	
	(1)	(2)	(3)	(4)
β_1: Geotagging	-0.00222	-0.00333	0.000406	0.00168
	(0.00281)	(0.00292)	(0.00293)	(0.00310)
β_2: Non-mainland Comment (NM)		-0.0405***		0.00554
		(0.00810)		(0.0195)
β_3: Geotagging × NM		0.0186		-0.0249
		(0.0120)		(0.0207)
$\beta_1 + \beta_3 = 0 \quad (p - value)$:		0.185		0.242
Observations	384,772	384,772	530,167	530,167
Pre-policy Ratio	0.2388	0.2398	0.0607	0.0605
Controls	✓	✓	✓	✓
Day-of-Week FEs	✓	✓	✓	✓
Poster-Type FEs	✓	✓	✓	✓
Linear	✓	✓	✓	✓
Donut	✓	✓	✓	✓

Notes: This table reports the effects of policy on the commenter's bot-score and whether the account has been banned or deactivated. The bandwidth is set at 40 days, the number of days before the rolling-in of the next policy phase. The regression includes the set of controls defined in the previous section, as well as day-of-week and post-type fixed effects. The estimate uses the linear time trend. Standard errors are clustered at the post-date level. Donut stands for whether we dropped a week of observations before and after the cutoff date.

Figure E.14: Revisiting Poster Location Heterogeneity and Sentimental Gaps in Anti-Western Sentiments



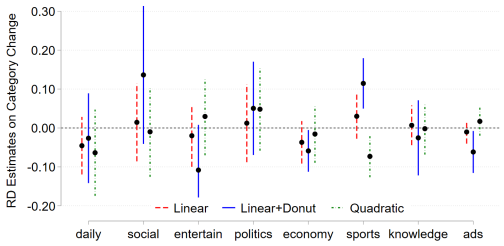
(a) Poster Location Heterogeneity



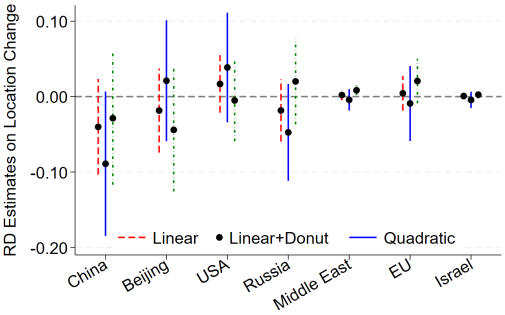
(b) Sentimental Gap Heterogeneity

Notes: The left figure is the marginal effects of poster's location on anti-Western sentiments. The coefficients are obtained via anti-Western RD regression with the treatment dummy and quadratic time trends interacted with the poster's location dummy. The right figure is the geotagging's impact on anti-Western sentiments correspond to each sentimental gap. The sentimental gap is calculated by the Euclidean distance on Russia between the post's sentiment and non-mainland poster's posts' mean sentiment. Then the sentimental gap measure is interacted with the treatment dummy and quadratic time trends in the anti-Western RD regression.

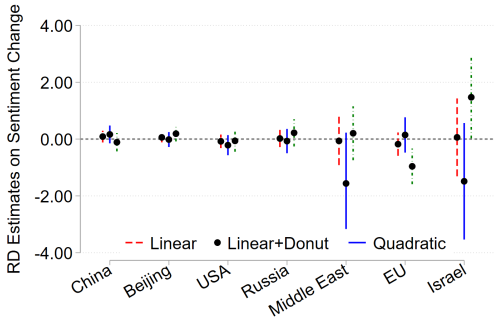
Figure E.15: RD Estimates of Change in the Post's Features



(a) RD Estimates of Change in the Post's Category



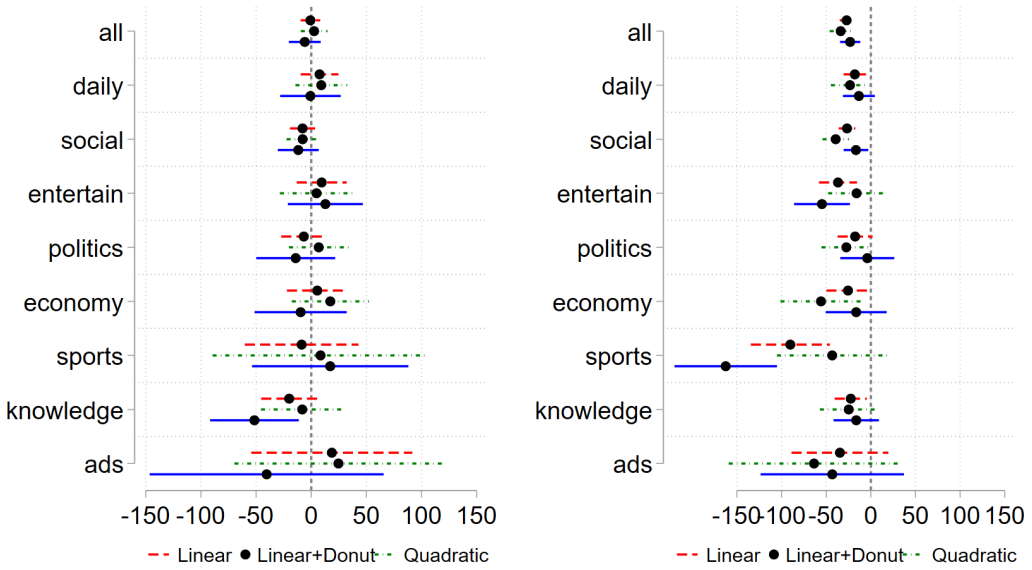
(b) RD Estimates of Change in the Post's Location Mentions



(c) RD Estimates of Change in the Post's Location Sentiments

Notes: These figures present RD estimates of the impact of profile geotagging on post categories and on sentiments toward specific locations, obtained from a post-date level specification of Equation 12. The red dashed lines show estimates using a linear time trend; the blue solid lines use a quadratic time trend; and the green dashed lines use a linear trend after excluding one week of observations on either side of the policy cutoff. All regressions control for post-level metadata specified in Section 5. Confidence intervals are based on robust standard errors.

Figure E.17: Policy's Impact on Shadow Banning, Separated by Post's Category



(a) Geotag in User Profile (b) Geotag in User Comment

Notes: These figures present RD estimates of the impact of profile geotagging and comment geotagging on the level of shadow banning, obtained from a post-level specification of Equation 12. The outcome variable is the level of shadow-banning of each post. The red dashed lines show estimates using a linear time trend; the blue solid lines use a quadratic time trend; and the green dashed lines use a linear trend after excluding one week of observations on either side of the policy cutoff. Panel (a) plots RD estimates where the policy variation is the geotagging in users' profiles. Panel (b) plots RD estimates where the policy variation is the geotagging in users' comments. All regressions control for post-level metadata, including the number of likes, reposts, and comments; post length; the time span between posting and the latest comment; and whether the post references a non-mainland location. Confidence intervals are based on robust standard errors.