Political Effects of Regulating Online Anonymity

Mingzi Niu*, Yiqian Wang† and Ziang Xiao‡

Abstract

Anonymity fosters the freedom of speech and intergroup communication on social media platforms. Meanwhile, burgeoning misinformation from anonymous sources stirs public support for platforms to regulate anonymity. This paper sheds light on this policy debate. We develop a signaling model incorporating identity stigma when individuals support stigmatized views. Leveraging recent Chinese social media regulations and text analysis advancements, we document two online political effects that are consistent with the model predictions: the identity revelation policy crowds out the participation of politically marginalized groups and silences unfavorable opinions. These findings highlight the potential trade-offs when designing identity-regulation policies.

*The Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem. mingzi.niu@mail.huji.ac.il

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

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

1 Introduction

"The new freedom of expression brought by the Internet goes far beyond politics. People relate to each other in new ways, posing questions about how we should respond to people when all that we know about them is what we have learned through a medium that permits all kinds of anonymity and deception."
— Peter Singer, 2005, Visible Man: Ethics in a World Without Secrets

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 providers, creating an environment where misinformation and hate speech can proliferate unchecked.

On the other hand, the extent to which opinion diversity and intergroup communication should be preserved has recently become a concern for policymakers. Burgeoning evidence shows that the government weighs social media sentiments and opinions when designing policies[1]. A lightweight anonymity regulation may not be enough to eradicate unaccounted harmful information. In contrast, an intensive anonymity regulation may entirely change the demography of users and opinions, henceforth distorting the government's learning of public opinion through social media. Therefore, it becomes increasingly critical to understand how anonymity shapes intergroup discourse and the diversity of opinion on social media.

This paper sheds light on this debate by leveraging a unique identity-revelation policy experiment in China. In early 2022, the Chinese government mandated one of China's largest social media platforms to geotag users – by revealing the user's geographic location on the platform. The policy is unique because it is the first government-mandated policy to regulate social media anonymity with an unparalleled influence by all identity-revelation practices. For example, X (formerly Twitter) used a blue checkmark to authorize users, but the authorization is neither mandated by the government nor the platform. PTT, a Taiwan-based social media platform, reveals users' IP addresses in the posts and comments. Still, the revelation is unique to the platform.

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

On the contrary, the policy we studied in this paper credibly provides a quasi-experiment to causally understand how identity shapes intergroup discourse and diversity of opinions. Initially, the platform geotagged only users who commented on posts related to the Russia-Ukraine Conflicts, which was deemed to reduce misinformation. The policy soon developed into universal implementations. In the first phase, the platform geotagged users by adding the tag to the user profile, making users' geographic location visible to the public. In the second phase, the platform geotagged users in all new comments posted on the platforms after the implementation day of the last phase. Each phase change was unannounced.

We motivate the empirical findings using a theoretical framework where individuals of two groups — local group and foreign group — decide whether to support, criticize, or keep silent to a given post that advocates a policy change. The two groups learn the value of the policy change from different sources of information, which potentially exhibit opposite media biases with the local media leaning towards status quo whereas the foreign one 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 belief about the policy change value, while expecting the cost of social sanction for their comments due to xenophobia. The social sanction cost associated with a comment varies with the media bias gap across groups and the identity revelation policy implemented on social media. 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 revelation policy.

The model yields the following two testable predictions. First, revealed group identity promotes the majority group's participation on social media (both support and criticism) while hindering expression from the minority group. We call this effect of media policy on participation *Labeling Effect*. Second, the intergroup gap in media bias amplifies the Labeling Effect, which disappears if two groups share an identical media slant.

Aligned with the model setup, we investigate the platform's intergroup interaction between non-mainland and mainland users. Despite the shared language, anecdotes suggest a lingering mistrust between these two groups. This divergence potentially arises from the remarkable intergroup difference

in the political environments and media sources, with mainland users' access to foreign news agencies primarily restricted by the Chinese government's deployment of the Great Firewall (GFW) more than a decade ago.

We evaluate the participation and viewpoints of two groups by compiling a unique dataset of 2.3 million comments posted on 18 thousand viral posts on the platform. These posts were from 509 influential posters. These posters on average attract 1.8 million followers with the most influential one attracts 24.8 million followers. These posts and comments were created a few months before and after the policy implementation. We fine-tuned a large language model to detect the sentiments and viewpoints of the posts and comments, including their generic attitudes and political inclinations. Using these data, we evaluate the policy's impact on the group participation and viewpoints and the potential mechanisms that channel the effects.

We document four important facts from the data. First, non-mainland users are indeed the 'cyber-minority' as they make up only 5% of all comments. Second, the poster's geographic identity also predicts the political attitudes in their posts. Third, the presence of non-mainland users substantially reduced during the sample period, and the policy likely triggered the reduction. Fourth, anti-Western sentiments grow in the sample period, which is also likely triggered by the policy.

Next, we estimate the impact of policy on intergroup participation and diversity of viewpoints. The main econometric strategy exploits the sharp implementation of two phases of policy on the platform. Specifically, we estimate the policy effects through a regression discontinuity in time (RDiT) design that uses time as the running variable. We show that the continuity assumption is very likely permitted by showing that a rich set of characteristics of the posts are continuous at the implementation date, suggesting the information environment of the social media is unaffected by the policy at the implementation date. By comparing the features of comments made before and after the policy implementation cutoff, we estimate the policy effects on the relative presence of non-mainland users and political attitudes.

Our first finding is the geotagging policy substantially discourages the participation of marginalized groups. The local estimates derived from the RDiT estimation suggest the policy reduces the proportion of non-mainland user comments by 20.3%. In line with the model prediction, we find the policy decreases both approval and disapproval comments from non-mainlanders, even after controlling the post's mean approval rate.

We document an interesting phenomenon that we described as the *Attrac-*

tion Effect. We find the policy, in the short-run, disproportionately attracts non-mainland users to 'test' through commenting. Therefore, the location revelation in the comment increases the proportion of non-mainland user comments by 20.8% . The increment is robust to the choice of bandwidth, and the estimate peeks at 20-day bandwidth with an increment of 30.7%. Meanwhile, we also find that the estimate diminishes with the bandwidth: the estimate falls to 12.2% at 40-day bandwidth. We find the policy effect reduces to negative when we further expand the bandwidth. This phenomenon suggests we must be cautious in interpreting the short-run versus long-run effects obtained from an RDiT design.

We test the second prediction, which states that the intergroup gap in media bias amplifies the Labeling Effect, using two approaches. In the first approach, we show the policy effect is heterogeneous to the poster's geographic identity, as the poster's geographic identity predicts the political sentiments. We find the mainland posters experienced a 41.7% reduction in non-mainland commenters, which explained most of the reduction of non-mainland users. In the second approach, we propose a measure that evaluates the sentimental gap between two posts on the same topic. The sentimental gap approximates the media bias, and we find that the reduction of non-mainland users happens more severely in the posts that expressed sentiments further from the mean sentiments expressed in posts written by the non-mainland posters.

We test multiple channels that could lead to the reduction. First, according to the theoretical framework, geotagging policy will simultaneously enhance the out-group punishment. Indeed, we find stronger geographical stigma expressed in the comments replied to the non-mainlander's opposing comments. Second, the reduction could due to the changing content or sentiment environment as the posters' response to the policy change. However, RD estimates show continuity of these measures at the policy implementation date. This suggest that while posters may change the content and the narrative in the long-run equilibrium, they are less likely to respond rapidly to the policy change.

Third, we tentatively explore the possibility of censorship as one driving force. Even though we are not be able to measure which post and comment got censored, we exploit a plausible platform loophole that could measure the within-post censorship. We show that the strength of within-post censorship is also continuous at the policy implementation date, suggesting that censorship is unlikely to be the channel through which unfavorable non-mainlanders' comments got disproportionately censored.

Our second finding suggests that the policy further polarizes the opinions or sentiments on the platform. Specifically, we show that geotagging substantially increases anti-Western sentiment. Geotagging in the profile increases the anti-Western sentiments by around 34%, and the effects are salient in topics that usually address controversial discussions such as politics, social news, and the economy.

The crowding out of non-mainland users could partially explain the rise of anti-Western sentiments. We find significant heterogeneity in the poster's geographic identity and sentimental gaps in the policy's impact on anti-Western sentiments. Combining with the first findings, we show that the crowding out of non-mainland users positively correlates with the exacerbation of anti-Western sentiments. We hypothesize that reducing non-mainland users increases the anti-Western sentiments not only through amplifying the representative of mainland users but also leads to a norms shift on social media regarding anti-Western sentiments, making these sentiments more acceptable.

In general, 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, Kotsadam, and Rooth 2021) [2]. We contribute to this literature by looking at the effect of segregating two groups distinct by their location, political attitude, and methods of acquiring information.

Second, our study contributes to the economics of identity and social image. In a related study, Braghieri (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 [3].

[2]Also other important identities such as ethnicities: Bazzi et al., 2019; race: Schindler and Westcott, 2021; Corno, La Ferrara, Burns, 2022, and social hierarchy: Lowe 2021

[3]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

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 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 elsewhere.

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 checker for the unidentified source. However, this strategy had limitations.

(2017); Levy (2021); Burnsztyn et al. (2019); Enikolopov, Makarin, and Petrova (2020); Gorodnichenko, Pham, and Talavera (2021)

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 identity-revelation 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 D.2a shows a standard Weibo profile showcasing self-reported gender, birthday, and location, followed by the system-assigned IP location. This IP location is determined by the latest IP address the user uses. Being inactive for over a month erases the IP location.

Comment identity-revelation 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 D.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.

The CAC's amendment was approved in June 2022, ordering all China-based platforms to show user IP locations from August 1, 2022. While the display method wasn't exactly instructed, most platforms adopted Weibo's approach, tagging comments directly with the user's IP location.

2.1.1 Sina Weibo

Sina Weibo has dominated the Chinese social media market for decades, with a monthly active user count exceeding 300 million. In 2014, Sina Weibo was listed on NASDAQ, the first China-based social media platform to be listed outside China. While international registration is accessible with a verifiable phone number, the platform chiefly serves Chinese speakers residing in China.

Named after the Chinese homonym for "Microblogging," Sina Weibo has an interface and functionality similar to Twitter. Users can craft posts incorporating text, images, and videos and are free to follow, read, and comment on content. The news feed algorithm displays posts from followed users, recommended posts from other users, and advertisements. Engaging with content is intuitive, as users can like, share, comment on posts, and interact with other users' comments.

Sina Weibo allows large flexibility in customizing profile pages. Users can share details ranging from their occupations and educational backgrounds to birthdays. A unique feature facilitates experts in verifying their real-world expertise, granting them an official verification sticker atop their profile. Additionally, an IP-based location, presumably derived from the most frequently used IP in the preceding days, is forcibly attached to the user's profile unless the user has been inactive for months.

For content discovery, Weibo offers both keyword and hashtag searches. An advanced search function provides users with rich filters, from post types (text, images, videos) to the timeframe within which content was published.

Given its vast user base and rich data, academics, especially social scientists, frequently turn to Weibo as a prime research subject. Noteworthy studies published in prominent journals[4] have delved into the profound dynamics of media capture, censorship, and the Chinese government's influence over the platform. With the rapid advancements in text analysis technologies, further groundbreaking research is anticipated to spring from this platform.

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 mistrust 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 mistrust 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

[4]Such as King, Pan, and Roberts (2017) and Qin, Stromberg, and Wu (2017)

the zero-COVID policy were often stigmatized as non-mainlanders. These criticisms usually talk about foreign governments intentionally endangering Chinese public health.

3 Theoretical Framework

We start with a theoretical framework to motivate the empirical hypotheses. This model provides two important predictions that guide the empirical analysis.

3.1 Model Setup

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

Actions. Each agent i needs to decide whether and how to comment on a post that advocates a local policy change: $d_i \in \{-1, 0, 1\}$. Here, $d_i = 0$, 1, -1 denotes that the agent i does not comment, comments to support (like), and comments to criticize (dislike), respectively.

Information. Let $\omega \in \mathbb{R}$ be the unobserved objective value of the policy change to the local people. All agents share the same prior belief on the policy change value: $\omega \sim \mathcal{N}(\omega_0, \sigma_0^2)$. For a given group $g \in G$, each agent in that group $i \in g$ receives a private signal $s_i = \mu_g + \epsilon_i$, where ϵ_i is identically and independently drawn from the normal distribution $\mathcal{N}(0, \sigma_\epsilon^2)$. We assume that $\mu_L \leq \omega \leq \mu_F$ to capture the mainstream media slants, with local media biased towards the status quo and foreign media biased towards the policy change. Define the gap in media slants between groups as

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

The intergroup media slant gap potentially reflects political censorship, ideological divides, or fundamental conflicts of interest across groups. ω_0, σ_0^2, μ_L, μ_F, and σ_ϵ^2 are known parameters.

In the baseline model, we consider *naive* agents who correctly recognize the media bias of the opposing group but perceive signals from their own

group as "objective" and unbiased.[5] We analyze the following two information environments $\xi \in \{\emptyset, g\}$ of the social media.

I. *Anonymous Environment* ($\xi = \emptyset$): group identities of commenters are hidden.

II. *Geotagging Environment* ($\xi = g$): group identities of commenters are revealed.

Payoffs. For any agent $i \in [0, 1]$, we normalize the payoff from not commenting ($d_i = 0$) as zero, and specify the payoff from the commenting decision $d_i = \pm 1$ in the environment $\xi \in \{\emptyset, g\}$ as

$$U_i(d_i) = B_i(d_i) - C_\xi(d_i).$$

Here, $B_i(d_i) \in \mathbb{R}$ denotes the expressive benefit from the decision d_i,[6] and we assume it to be proportional to the expected value of the policy change perceived by the agent i:

$$B_i(d_i) = \beta E[\omega | s_i] d_i. \tag{1}$$

$C_\xi(d_i)$ denotes the expected xenophobia sanctions associated with the decision d_i. We assume that the media slant gap Δ intensifies xenophobia sanctions, and therefore

$$C_\xi(d_i) = \psi(\Delta) \left[\lambda \Pr(i \in F | d_i, \xi) + (1 - \lambda) \Pr(i \in L | d_i, \xi)\right] \geq 0, \tag{2}$$

where $\psi(0) = 0$, $\psi'(\cdot) > 0$, and $\xi \in \{\emptyset, g\}$ denotes the information environment.

[5]The assumption of naivete can be relaxed without changing our main results qualitatively. In the case of full sophistication, where all agents know the exact magnitude of their media bias $b_g = \mu_g - \omega$, they can correct their information accordingly $\tilde{s}_i = s_i - b_g = \omega + \epsilon_i$. Therefore, the full sophistication case is equivalent to learning from unbiased information sources—that is, mainstream media biases cannot distort agents' learning—and implies $\Delta = 0$. In the case of partial sophistication, where a positive mass of agents underestimate the magnitude of media bias of their own information source ($|\tilde{b}_g| < |b_g|$), they cannot fully recover an unbiased signal for themselves and thus remain to be affected by media slants. The degree of naivete only affects the magnitude of the "corrected" gap in media slants $\tilde{\Delta} = (\mu_F - \tilde{b}_F) - (\mu_L - \tilde{b}_L) \in (0, \Delta)$.

[6]The expressive benefit captures the warm glow (see, e.g., Andreoni, 1995; Bursztyn et al., 2023) from expressing the right view, which refers to a sense of satisfaction that people experience for "doing their part".

3.2 Analysis

We first derive the posterior belief on the policy change value for any agent $i \in g$. Upon receiving the signal $s_i \sim \mathcal{N}(\mu_g, \sigma_\epsilon^2)$, the agent i bayesian updates their belief and the posterior belief follows the normal distribution $\mathcal{N}(\mu_i, \sigma^2)$ with

$$\mu_i = \frac{\tau_0 \omega_0 + \tau_\epsilon s_i}{\tau_0 + \tau_\epsilon}, \tag{3}$$

where $\tau_0 = \sigma_0^{-2}$ and $\tau_\epsilon = \sigma_\epsilon^{-2}$ denote the precision of the prior belief and the observed signal, respectively.

Combining (1) and (3), we have

$$B_i(d_i) = \beta \frac{\tau_0 \omega_0 + \tau_\epsilon s_i}{\tau_0 + \tau_\epsilon} d_i.$$

Note that the social sanctions associated with each choice, $C_\xi(d)$, are fixed for any given information environment $\xi \in \{\emptyset, g\}$; they do not vary with the individual commenting decision of the agent i. Therefore, the agent i chooses $d_i = 1$ if and only if

$$\begin{aligned} \beta \frac{\tau_0 \omega_0 + \tau_\epsilon s_i}{\tau_0 + \tau_\epsilon} &\geq C_\xi(1), \\ \Rightarrow \quad s_i &\geq \bar{s}_\xi, \end{aligned}$$

where

$$\bar{s}_\xi = k C_\xi(1) - b, \tag{4}$$

$$k = \frac{\tau_0 + \tau_\epsilon}{\beta \tau_\epsilon} > 0, \tag{5}$$

$$b = \frac{\tau_0 \omega_0}{\tau_\epsilon}. \tag{6}$$

The agent chooses $d_i = -1$ if and only if

$$\begin{aligned} -\beta \frac{\tau_0 \omega_0 + \tau_\epsilon s_i}{\tau_0 + \tau_\epsilon} &\geq C_\xi(-1), \\ \Rightarrow \quad s_i &\leq \underline{s}_\xi, \end{aligned}$$

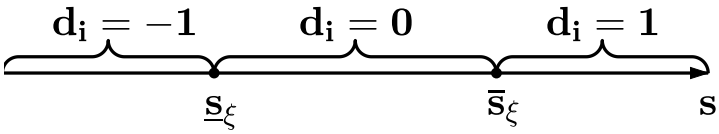
where

$$\underline{s}_\xi = -k C_\xi(-1) - b, \tag{7}$$

and k is given by (5) and b is given by (6).

Since $k > 0$ and $C_\xi(d) \geq 0$ for any $d = \pm 1$, $\xi \in \{\emptyset, g\}$, we have $\underline{s}_\xi \leq \bar{s}_\xi$ where the equality holds if and only if $\Delta = 0$. Therefore, the commenting decision for any agent $i \in [0, 1]$ exhibits a cutoff feature: they support the post ($d_i = 1$) when observing a sufficiently good signal $s_i \geq \bar{s}_\xi$, criticize the post ($d_i = -1$) when observing a sufficiently bad signal $s_i \leq \underline{s}_\xi$, and do not comment otherwise, i.e., $s_i \in (\underline{s}_\xi, \bar{s}_\xi)$.

Figure 1: Commenting Decisions Upon Signal Observation



Note also that the relative precision of signals, τ_ϵ/τ_0, controls the social media participation. With a more precise or concentrated signal, the threshold for expressing support goes lower whereas that for criticism goes higher. In other words, if agents have higher confidence in their private information, then participation on the social media, both commenting to support and criticize, will be promoted. Furthermore, a more precise signal implies a lower sensitivity of participation to the sanction costs (k). Therefore, if the agent is more convinced by their source of information, then they tend to express their views and are less affected by xenophobia.

3.2.1 Anonymous Environment ($\xi = \emptyset$)

Consider the scenario when the group identity is hidden when commenting. The social sanction of xenophobia operates based on the content of comments. In particular, if the two groups are of the same size, then a supportive comment is indicative of a foreign agent, and thus tends to trigger xenophobia sanction from the local group. Intuitively, this is because an agent gives a positive comment only when observing a sufficiently good signal, which is more likely to be obtained in a foreign group that is biased toward a policy change.

Formally, any agent of group g, $i \in g$, draws a signal $s_i \sim \mathcal{N}(\mu_g, \sigma_\epsilon^2)$, supports the post ($d_i = 1$) if and only if $s_i \geq \bar{s}_\emptyset$ and criticizes the post ($d_i =$

-1) if and only if $s_i \leq \underline{s}_\emptyset$. Hence, within the group g, support occurs with probability $1-\Phi(\frac{\bar{s}_\emptyset-\mu_g}{\sigma_\epsilon})$ and criticism occurs with probability $\Phi(\frac{\underline{s}_\emptyset-\mu_g}{\sigma_\epsilon})$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. Therefore, by Bayes' rule,

$$\begin{aligned}
&\Pr(g=F|d=1) = \frac{(1-\lambda)[1-\Phi(\frac{\bar{s}_\emptyset-\mu_F}{\sigma_\epsilon})]}{(1-\lambda)[1-\Phi(\frac{\bar{s}_\emptyset-\mu_F}{\sigma_\epsilon})]+\lambda[1-\Phi(\frac{\bar{s}_\emptyset-\mu_L}{\sigma_\epsilon})]},\\
&\Pr(g=L|d=1) = 1-\Pr(g=F|1),\\
&\Pr(g=F|d=-1) = \frac{(1-\lambda)\Phi(\frac{\underline{s}_\emptyset-\mu_F}{\sigma_\epsilon})}{(1-\lambda)\Phi(\frac{\underline{s}_\emptyset-\mu_F}{\sigma_\epsilon})+\lambda\Phi(\frac{\underline{s}_\emptyset-\mu_L}{\sigma_\epsilon})},\\
&\Pr(g=L|d=-1) = 1-\Pr(g=F|d=-1).
\end{aligned}$$

Clearly, under $\mu_F \geq \mu_L$, if $\lambda = 1/2$, then support is more likely to come from the foreign group and criticism is more likely to come from the local group, i.e., $\Pr(g=F|d=1) \geq 0.5 \geq \Pr(g=F|d=-1)$, confirming our aforementioned intuition.

Bringing these conditional probabilities to (2), we have, for any $i \in I$,

$$C_\emptyset(1) = \psi(\Delta)\lambda(1-\lambda)\frac{2-\Phi(\frac{\bar{s}_\emptyset-\mu_F}{\sigma_\epsilon})-\Phi(\frac{\bar{s}_\emptyset-\mu_L}{\sigma_\epsilon})}{(1-\lambda)[1-\Phi(\frac{\bar{s}_\emptyset-\mu_F}{\sigma_\epsilon})]+\lambda[1-\Phi(\frac{\bar{s}_\emptyset-\mu_L}{\sigma_\epsilon})]}, \tag{8}$$

$$C_\emptyset(-1) = \psi(\Delta)\lambda(1-\lambda)\frac{\Phi(\frac{\underline{s}_\emptyset-\mu_F}{\sigma_\epsilon})+\Phi(\frac{\underline{s}_\emptyset-\mu_L}{\sigma_\epsilon})}{(1-\lambda)\Phi(\frac{\underline{s}_\emptyset-\mu_F}{\sigma_\epsilon})+\lambda\Phi(\frac{\underline{s}_\emptyset-\mu_L}{\sigma_\epsilon})}, \tag{9}$$

where $\bar{s}_\emptyset$ is given by (4) and $\underline{s}_\emptyset$ is given by (7).

Proposition 1 *For any $\lambda \in (0,1)$, there exists an equilibrium where agents with signal $s_i < \underline{s}_\emptyset$ choose $d_i = -1$, those with signal $s_i \in (\underline{s}_\emptyset, \bar{s}_\emptyset)$ choose $d_i = 0$, and those with signal $s_i > \bar{s}_\emptyset$ choose $d_i = 1$. Additionally, there exist $(\underline{\lambda}, \bar{\lambda}) \in (0,1)^2$ such that the equilibrium is unique if $\lambda > \bar{\lambda}$ or $\lambda < \underline{\lambda}$.*[7]

3.2.2 Geotagging Environment ($\xi = g$)

Now we turn to the case when the group identity is revealed when commenting. In this case, the social sanction of xenophobia is directly applied to

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

commenters from the opposing group, regardless of the content of comments. Therefore, for any $d = \pm 1$, a local agent $i \in L$ faces the sanction cost of

$$C_L(d) = \psi(\Delta)(1 - \lambda), \tag{10}$$

and a foreign agent $i \in F$ faces the sanction cost of

$$C_F(d) = \psi(\Delta)\lambda. \tag{11}$$

Proposition 2 *For any $\lambda \in (0, 1)$, there exists a unique equilibrium where agents of group $g \in G$ with signal $s_i < \underline{s}_g$ choose $d_i = -1$, those with signal $s_i \in (\underline{s}_g, \bar{s}_g)$ choose $d_i = 0$, and those with signal $s_i > \bar{s}_g$ choose $d_i = 1$.*

3.3 Predictions

Prediction 1 (Labeling Effect) *Suppose that there exists a gap in media slants between the local group and the foreign group (i.e., $\Delta > 0$). When the group identity is revealed, the minority group's likelihood of commenting, both supporting ($d_i = 1$) and criticizing ($d_i = -1$), declines, whereas the majority group's likelihood of commenting ($d_i = \pm 1$) rises.*

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

3.4 Discussion

In the empirical setting, we study the behavioral responses of Sina Weibo users from mainland and non-mainland to the geotagging mandate. The platform operates within China and only attracts a small portion of non-mainland users, making these users the minority F.

Prediction 1 suggests that the revelation policy reduces the proportion of comments from non-mainland users, regardless if the comment shows support or oppose to the original posts. Prediction 2 suggests that the reduction varies by the media slant presented by the posters. A poster presents with more non-mainland favorable slant should find less reduction of non-mainland users than a poster present with less non-mainland favorable slant.

In the following sections, we empirically test these predictions mainly through a RDiT design leveraging the data we scraped from Sina Weibo around the policy implementation date.

4 Data

We focus on the comments responding to posts, viewing them as representations of individual opinions about the news or information mentioned in the post. We developed Python scripts to collect posts and comments from Sina Weibo. Eventually, we compiled data with 17,748 sample posts authored by 509 randomly chosen influential accounts, accumulating 2.8 million comments on Sina Weibo. These posts and their comments were posted between January 1, 2022, and September 1, 2022. We choose this time window to ensure sufficient observations before and after the policy interventions on March 4th, March 18th, and April 28th, 2022.

4.1 Topic-oriented process

To obtain the data, we developed a topic-oriented process. First, we collected breaking news during our sample period and extracted keywords from it. Summarized in Table A.1, we extracted 30 keywords from the news. Many of them represent the most viral topics on Chinese social media during the sample period. For example, Human Trafficking represents a high-profile human trafficking case in Jiangsu province, accidentally revealed by a TicToker; Taiwan represents Nancy Pelosi's visit to Taiwan in August 2022.

Next, we searched these keywords through Weibo's search engine and gathered all posts that received over a hundred likes in the sample period. We then obtained a list of 7,499 posters for these posts. We dropped around 2000 posters that are (1) currently associated with government entities or news agencies, (2) dedicated celebrity fan pages, or (3) being deactivated, to minimize the manipulation while at the same time maximizing the diversity of the content. Due to computational limitations, we eventually drew 10% of the posters from the pool and obtained a set of 509 posters. This selection procedure is dedicated to a set of fairly influential posters on the platform. We argue that stronger policy impacts may be detected from this pool of posters as their contents were widely spread and well-commented by commenters with diverse demographics.

Following the selection of posters, we scraped all the posts written by these posters during the sample period. For the monthly Top 5 most-liked posts, we scrape the last 50 pages of comments due to Weibo's restriction. We obtain both the post and comment characteristics along with their texts, such as the number of likes, reposts, comments, and geographic location, if

available.

We obtained the poster's and commenter's IP location by two methods. We directly scraped the geotag attached beneath the comment for the commenter who commented after April 28, 2022. For the commenter who commented before April 28, 2022, we programmed the Python scraper to click into her user profile and extracted the geotag. The geotag usually stays on user's profile page for one month since the user's last activity. Therefore, the location data was almost perfectly retrieved for comments that appeared after April 28, while it was missed by around 30% for comments that appeared before April 28. This indicates that many users who commented before April 28 became inactive for over a month by the time we conducted scraping.

Eventually, we obtained two levels of text data. The first level is the poster-post level. Among 17,748 posts written by 509 posters, we obtained the content, popularity, and location data of these posts. The second level is the comment-commenter level. Among around 2.3 millions of comments made to these posts, we obtained the content, popularity, and location data of comments made to the posts.

4.2 Text Analysis

Text Analysis with Large Language Models

To assess the topic and the sentiment of the posts and comments, we built a hybrid textual analysis approach that is flexible to the specificity and complexity of the task. For processing data-intense but inference-light tasks, we fine-tuned ChatGLM4-9b, a relatively small-sized LLM that can be implemented locally. This model labels a set of topics and basic sentiments of the posts and comments.

For inference-heavy tasks such as labeling specific sentiments, evaluating sarcastic tones, and navigating dialogues between commenters, we adapted a codebook-based approach illustrated in Xiao et al., (2023) . Compared to the traditional machine learning-based approach in text analysis, the LLMs are pre-trained and can often be accessed publicly through API. It does not require a pre-existing annotated dataset for model training while achieving good performance in deductive coding tasks (Ziems et al., 2023; Xiao et al., 2023). Specifically, we combined the LLM with pre-developed codebooks to assign each post or comment a pre-defined code. Each codebook contains three main elements for each code: name, description, and examples. To

leverage LLMs in context learning capability, we included five examples for each code as in context learning examples. The coding process is as follows. First, we randomly sampled 1.3-20% of the data for each coding task and instructed expert coders to assign labels given an existing codebook. Then, we iterate the prompt with the LLM until it reaches adequate inter-rater agreement with expert coders. We code the rest of the data with the same prompt. To ensure reproducibility, we run the coding process three times, and the final code for each data point is decided based on a majority vote. The model used for coding is gpt-4-32k-0613 with the temperature set to 0 [8].

Posts: Topics and Sentiments

Using the hybrid approach described above, 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 overlap several topics. For example, a post on the COVID lockdown could expand politics, social news, and daily. In preparation for case studies, we also commanded the model to determine whether the post is specific to COVID-19 and the Russia-Ukraine War.

Second, we extracted the key objects from the posts. The key objects are defined by the key person, item, location, and events that happened in the posts. For example, the key objects for a post about the Russia-Ukraine War could be Russia, Ukraine, and War. We then command the LLMs to evaluate the poster's context-specific sentiments towards the key objects. 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 LIKELY NEGATIVE (-1) or NEGATIVE (-2); 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 obtained three sets of information from the posts: (1) Topics of the post; (2) A vector of the post's key objects; (3) A vector of the poster's sentiment towards each key object.

[8]The model temperature controls the randomness of the output. For GPT-4, the temperature could be set between 0 and 1. Lower values make it more deterministic.

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 replies to the comments. These objects are generally considered part of comments but were explicitly made to another comment. We evaluate the tone of these replies and specifically explore whether the replies utilized geographical stigma or stereotyping to express their disagreement. This strategy may become more common on social media when the users' 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.

4.3 Data Summary

In this section, we describe our data in detail. We separately provide details on posts, posters, and comments. Table 1 shows comprehensive summary statistics for posts (Panel A), posters (Panel B), and comments (Panel C).

4.3.1 Posts

In total, we amassed 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 a million likes, over 300 thousand reposts, and over 80 thousand comments. Figure D.3 illustrates the key attributes of

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

	Mean	SD	Min	p25	p75	Max
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 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 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.

these posts.

Table A.2 summarizes the distribution of the post topics. The sum of the proportion across topics exceeds one, as many posts overlap multiple topics. A majority of posts were identified as politics (36.2%) and social news (52.2%). These numbers reflect our selection of keywords to trace the posters. They also reflect that a majority of viral posts written by influential posters on Chinese social media were related to politics and social issues, contrary to what people usually conceive as a highly censored online environment. Following politics and social news is the entertainment post (24.9%). This number is not surprisingly large as the most viral post, by all metrics, in our sample, is a post about a rumor between two Chinese comedians. A

small portion (11%) of posts transmitted knowledge. 747 (38.4%) of them are about COVID-19, such as the transmission knowledge and vaccine. However, evaluating the truthfulness of knowledge post is beyond the scope of this paper but an important topic to be explored in the future. Only 3.7% and 2.9% of posts are advertisements and sports. These small proportions suggest our effective elimination of fan accounts, as these accounts are mostly related to celebrities in sports and are also filled with advertisement posts. What is not shown in this table is the proportion of COVID-19 and Russia-Ukraine War posts, with each of them taking 20.78% and 8.28% of all posts.

The LLM also extracted the key objects from each post and identified if any of them were related to a location. Overall, A majority of (65%) posts are related to China. Posts involving countries in North America follow with around 16%. The majority of these posts are related to the US. Around 8% of all posts involve European countries and Russia-Ukraine. Both Hong Kong/Taiwan-related posts and East Asia-related posts occupy around 5%.

Based on the context of the post, we instructed the LLM to label the poster's sentiments towards the location from NEGATIVE (-2) to POSITIVE (2). Figure D.7 shows a considerable location heterogeneity in the post's sentiments. On average, posts show positive sentiments towards China and European countries. A significant portion of tourist posts contribute to positive attitudes towards Europe. Both the USA and Russia share a similar level of negative sentiment, while a significantly stronger negative sentiment is expressed towards Ukraine. The comparison between Russia and Ukraine is consistent with the anecdotal evidence that Chinese netizens generally support Russia over Ukraine[9]

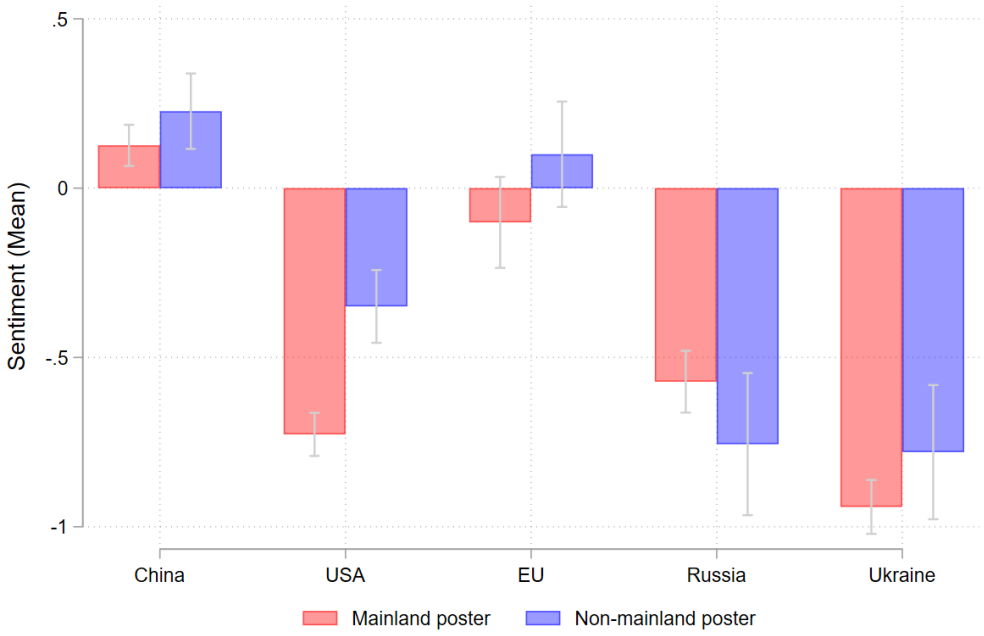
4.3.2 Posters

We sampled 509 posters who are predominantly influential on the platform. As shown in Figure D.5a, on average, these posters own 1.8 million followers, with the most followed poster, Hu Xijin, the retired editor-in-chief of a communist-party flagship newspaper, reaching a staggering 24.8 million followers. According to Figure D.5b, approximately 18% of these posters are non-mainland users.

[9]One caveat is that the figure is plotted without the poster's fixed effects. Anti-Western posters may post disproportionate levels of anti-Western posts that substantially reduce the mean sentiments toward the USA and Ukraine.

Notably, the poster's geographic location predicts their sentiments toward different locations, with the direction consistent with mainland users being more anti-Western. Figure 2 shows that mainland posters expressed significantly stronger negative sentiment towards the USA, European countries, and Ukraine, with the sentiment towards European countries turning marginally negative. Non-mainland posters, while showing negative sentiments toward these regions as well, expressed weaker negative sentiments toward the USA and Ukraine and stronger negative sentiments toward Russia.

Figure 2: Poster Location Heterogeneity in Sentiments



Furthermore, we evaluate the sentimental gap presented by mainland posters versus non-mainland posters. The sentimental gap represents the poster imposing a different sentiment toward an object by manipulating the framing or selectively reporting content that aligns with her partisanship. Figure D.8 shows the potential sentiment bias of mainland and non-mainland posters when discussing COVID-19. While two groups of posters share similar sentiments for China and Shanghai when discussing COVID-19, their

sentiments deviate for the USA. Figure D.9 shows that mainland posters are significantly more positive towards Russia and non-mainland posters are significantly more positive towards Ukraine in posts discussing the Russia-Ukraine War. The sentiment divergence, conditional on the post's topic, suggests the sentimental gap between the two groups of posters.

4.3.3 Comments

We scraped a total of 2.3 million comments made to the almost 18,000 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.

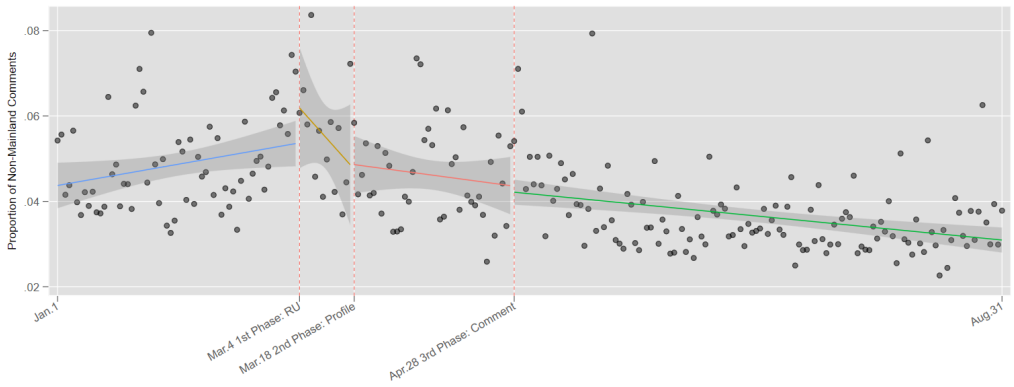
We find non-mainland comments only account for 4.22% of all comments in our data. Figure D.10a shows the geographic distribution of non-mainland comments. The number of comments in each country positively correlates with the number of Chinese immigrants in each country. The US leads the count with over 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 D.10b illustrates the distribution of comments from mainland users. As further depicted in Figure D.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 motivates one of the paper's main results. Each dot represents the daily proportion of non-mainland comments from the sample posts. This figure presents an overall reducing trend of non-mainland comments, starting on March 4, 2022, when Weibo began experimenting with comment geotagging in Russia-Ukraine posts. In the following sections, we use different strategies to identify the extent to which the policy causes the reduction.

Significant differences exist between the comments from mainland and non-mainland users. Table A.3 shows that non-mainland users generally

Figure 3: The Daily Proportion of Non-mainland Comments



display a more subdued sentiment. Their comments are 8% less negative, roughly equally positive, and 10% more neutral compared to those from mainland users. In political posts, non-mainland users also exhibit a lower anti-western sentiment. Only 44% of their responses to international political posts conveyed anti-Western sentiment, in contrast to the 52% observed among mainland comments. This disparity underscores the variations in political attitudes between the two groups.

5 Empirical Strategy

5.1 Regression Discontinuity in Time

The primary goal of this paper is to evaluate the policy's causal impact on the diversity of users and opinions on social media. We estimate the policy impacts using a Regression Discontinuity in Time (RDiT) design.

The identification assumption is that the information environment on the platform is identical in a narrow time window around the policy implementation. Therefore, changes in the user and opinion compositions are caused by the identity revelation instead of other informational channels. We test for discontinuities in post characteristics and sentiments on the policy implementation day. By leveraging the uniquely rich high-frequency text data, we most likely rule out unobserved shocks correlated with time.

Equation 12 is the main RDiT specification:

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

In the primary regression, we test the *Revealing Effect* suggested by Prediction 1. Therefore, y_{ijpt} is a non-mainland dummy that equals one if the comment i's geographic location is somewhere non-mainland. In the second regression, we evaluate the impact of the policy on opinion diversity. Specifically, we examine the change in the xenophobic sentiments in the comments. Therefore, y_{ijpt} is a xenophobic dummy that equals one if the comment i is detected to contain a xenophobic sentiment.

In the main results, we look at the treatment effects of geotagging in the user profile. Therefore, treatment variable $Geotag_t^d$ indicates whether the profile geotagging policy has been implemented at date t. In principle, $d \in \{\text{profile}, \text{comment}\}$ because the policy was implemented in two phases: first in the user's profile and second in the user's comment. However, we pursue a cleaner identification and mainly focus on the analysis of the first phase. This is because the policy's first phase has contaminated the pre-intervention period of the second phase, and social media users might have anticipated the second phase. Therefore, $Geotag_t^{profile}$ equals one if the comment i was made on or after Mar 18, 2022.

The variable $Time_t^d$ is the date centered around the first day of geotagging in the user's profile. $f^a(Time_t^d)$ and $Geotag_t^d \times f^b(Time_t^d)$ 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 include 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 towards the short-run treatment effect.

In the main specifications, we set 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.

In a set of covariates X_{ijpt}, we include the post's number of likes, reposts, and comments to control the post's spread level. We control the post's length

to approximate the post's media composition. We control the number of days between the date of posting and receiving the last comment to evaluate the post's influence on the platform. We include the 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 unobserved heterogeneity within each combination of topics[10]. Last of all, we cluster the standard errors at the post-date level to address the unobserved correlation in commenters' geographic identity and sentiment within a post-day.

5.1.1 Heterogeneity

Prediction 2 suggests the *Revealing Effect* exacerbates with the media gap Δ. We empirically test this prediction through two approaches. In the first approach, we evaluate the poster's location heterogeneity on revealing effect. In the second method, we evaluate the heterogenous effect of media gaps on revealing effect by interacting Equation 12 with a measure of the media gap. Specifically, we approximate the media gap with the sentimental differences observed from the data.

The sentimental differences, illustrated by Figure **??** and **??**, shows that posters from different groups expressed different sentiments towards specific locations after controlling the topic of the posts. Therefore, the sentimental differences are consistent with the definition of Δ. In practice, we set the average sentiments toward the object o by non-mainland posters as a benchmark and denote it as $\bar{\mu}^o_{NM}$. Then, the media gap on object o between a post p and the benchmark non-mainland sentiment is represented as $\Delta^{o,p} = \mu^{o,p} - \bar{\mu}^o_{NM}$. We then aggregate the differences through the Euclidean distance over the set of the object to create an aggregate measure of the media gap:

$$\Delta^p = \sqrt{\sum_o (\mu^{p,o} - \bar{\mu}^o_{NM})^2} \tag{13}$$

For example, for a COVID-19 post that involved the US and China, $\Delta^p = \sqrt{(\mu^{p,usa} - \bar{\mu}^{usa}_{NM})^2 + (\mu^{p,cn} - \bar{\mu}^{cn}_{NM})^2}$.

[10]For example, the sentiment towards a same political figure expressed by the poster in a Political post may be different from that in a Political×Knowledge post, as the latter post may simply introducing the political figure. However, unobserved differences in sentiments towards the political figure may attract different groups of commenters.

5.1.2 Robustness check with Russia-Ukraine posts

To test the robustness of the estimation of RDiT, we exploited the experimental stage of the policy when the platform revealed the geographic location of the comments only for posts related to Russia and Ukraine. These posts were identified to be treated as they explicitly included the keyword "Russia" or "Ukraine" in the content. We can identify the local effects of revealing IP location through a standard difference-in-differences design:

$$y_{ijpt} = \beta_1 Geotag^{RU}_{jpt} + \beta_2 Post_t + \beta_3 Geotag^{RU}_{jpt} \times Post_t + \gamma X_{jpt} + DoW_t + \epsilon_{ijpt}, \tag{14}$$

where $Geotag^{RU}_{jpt}$ indicates whether the post j includes the triggering keywords and $Post_t$ is the conventional status dummy that whether the post j is written after March 4, 2022, the policy implementation day.

While the ideal comparison should be made between the Russia-Ukraine posts that contained the keywords and those that did not, only around 20% of Russia-Ukraine posts exclude both "Russia" and "Ukraine." Rather, given that almost all the Russia-Ukraine posts were political during this time frame, we consider all the other international political posts to be a fair control. We restrict the sample to all the international political posts written between January 1, 2022, and March 18, 2022. We include the same set of controls while dropping the post-type fixed effect, as all the posts included in the regression are political posts.

Selection bias additionally confounds the interpretation of the treatment effect estimated from Equation 14. First, posters can self-select into or opt out of the treatment by intentionally including or excluding keywords to trigger or evade the function. To assess to what extent the poster takes advantage of this loophole, we predict the likelihood of including the keywords in Russia-Ukraine-related posts using a rich set of poster characteristics. If none of the traits strongly predicts the behavior, we can be reassured of the selection concern.

6 Results

We start the discussion of results by first showing that the policy unlikely influenced the group participation by altering the information environment on the policy implementation day. We regress the covariates on the treatment

dummies, linear and quadratic time trends, and fixed effects. The standard errors are clustered at the post-date level.

RD Balance Figure D.13 plots the regression residuals of the post's number of likes, reposts, and comments balanced around the cutoff, indicating the spread of the viral posts on the platform was unaffected. Figure D.14 shows the length and date span of the posts were also unaffected, suggesting the media composition and the influence of the posts remain the same.

6.1 *Revealing Effect*: reduction of Non-mainland users

Table 2 reports the local average treatment effect of geotagging in user's profile on Non-mainlanders' participation. Columns (1) - (3) show the estimated policy's impact on the non-mainlander's direct comment towards the post. Column (1) shows the policy reduced the non-mainlander's direct comments by 20.3% (-0.00966/0.0476). The estimate is marginally significant at the 10% level. Column(2) shows the coefficient from controlling quadratic time trends before and after the policy implementation. The estimated coefficient is comparable to the full regression but no longer significant.

Column (3) shows the coefficient from the regression with linear time trends but dropping one week of observations before and after the policy implementation. The coefficient decreases to -0.02, equivalent to a remarkably 40.3% (-0.02/0.484) reduction in non-mainlanders participation. Comparing column (3) to column (1), we argue that further reduction can be explained by a short-run positive effect of geotagging that disproportionately attracts non-mainland users. We will discuss this phenomenon more extensively in the later paragraphs.

In terms of reply to the comment, columns (4) and (5) suggest the policy did not influence the non-mainlander's participation in the reply to the comment. One possible reason is that Sina Weibo "hides" the reply to the comment in a drop-down list that could not be directly observed from browsing the post, reducing the social stigma imposed on non-mainland users.

6.1.1 Approval or Disapproval

Prediction 1 suggests that geotagging reduces the participation of both compliant and dissent non-mainlanders. We measure the compliance and dissent using the level of approval detected from the comments. Table 3 shows the estimation using Equation 12 but using indicators for approval or disapproval

Table 2: Geotagging in Profile on Non-mainland Participation

	Direct Comments			Reply to a Comment	
	(1)	(2)	(3)	(4)	(5)
Geotag	-0.00966*	-0.0105	-0.0200**	-0.00593	-0.00963
	(0.00494)	(0.00750)	(0.00859)	(0.00691)	(0.00993)
Control Mean	0.0476	0.0476	0.0484	0.0631	0.0629
Observations	385,086	385,086	316,537	256,745	211,811
Linear	✓	✗	✓	✓	✓
Quadratic	✗	✓	✗	✗	✗
Donut	✗	✗	✓	✗	✓

Notes: This table reports the estimation results from an RDiT using the non-mainland comment indicator as the outcome. The bandwidth is set at 40 days, the number of days before the policy phase change. 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 Quadradic stands 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.

of non-mainland comments as the outcome variable. We drop the first week observations around the cutoff date for all of the regressions.

Consistent with the model prediction, we obtain negative estimates from all variations of the estimating equations. However, most of these coefficients are insignificant due to small sample sizes and the magnitude of change. For example, only 2,742 non-mainland comments expressed disapproval within the 40-day bandwidth, with two more weeks being dropped to create the donut, and the proportion of disapproval non-mainland comments is less than 0.1%.

Columns (1) and (2) show the RD estimates with linear time trends. The estimates show a 32.4% and 30.1% reduction in non-mainland users commenting on approval and disapproval of the post. Columns (2) and (5) show the RD estimates with additional control in the average (dis)approval rate of the post. Adding this variable controls the level of controversy of the post. For example, a controversial post disproportionately attracts disapproval comments. In this case, a disapproval comment suggests compliance instead of dissent. In other words, controlling the average (dis)approval rate reduces the measurement error of the comment's compliance or dissent and,

indeed, increases the magnitude and precision of the estimates[11].

Columns (3) and (6) show the estimates using quadratic time trends. Changing to quadratic trends substantially increases the magnitude of estimates. We suspect the higher-order polynomial is sensitive to the variation within a narrow time bandwidth. Nevertheless, these estimates are consistent with linear estimates and theoretical predictions.

Table 3: Geotagging in Profile on Non-mainland Approval and Disapproval Comments

	Approval			Disapproval		
	(1)	(2)	(3)	(4)	(5)	(6)
Geotag	-0.00948	-0.0147**	-0.0377***	-0.00262	-0.00284	-0.00633
	(0.00629)	(0.00634)	(0.0118)	(0.00187)	(0.00183)	(0.00445)
Post approval rate		0.0600***	0.0605***			
		(0.00600)	(0.00599)			
Post disapproval rate					0.0334***	0.0332***
					(0.00147)	(0.00262)
Control Mean	0.0293	0.0293	0.0293	0.0087	0.0087	0.0087
Observations	316,537	316,537	316,537	316,537	316,537	316,537
Linear	✓	✓	✗	✓	✓	✗
Quadratic	✗	✗	✓	✗	✗	✓

Notes: This table reports the estimation results from an RDiT using the non-mainland (dis)approval comment indicator as the outcome. The rest of the specification is similar to that of Table 2 except we drop one week of observations before and after the cutoff date. We only detect the approval and disapproval sentiments among the direct comments. Post (dis)approval rate represents the proportion of (dis)approval in the post.

6.1.2 Attraction effect

Recall that we have shown the comparison columns (1) and (3) in Table 2 suggest a potential attraction effect of the policy towards the non-mainland users. Figure D.12 shows that each policy implementation phase triggered a surge of comments that discussed users' IP locations. Non-mainland users may be more inclined to test the precision of the geotagging function when

[11] We can measure the user i's social compliance in post p by $|Approval_{i,p} - \overline{Approval}_p|$, where $\overline{Approval}_p$ is the proportion of approval comments received by post p. Therefore, not including $\overline{Approval}_p$ is an omitted variable problem that induces a mismeasurement of i's compliance.

it was first implemented, therefore, to some extent, mitigating the negative effect of the policy on their participation in the short run.

The attraction effect of geotagging is amplified in the short run when the policy is implemented in the comment. A surprisingly drastic contrast emerges when estimating the local impact of geotagging in comments. Table D.22 shows this latter phase of the policy increased the ratio of non-mainland comments significantly by 20.8% in a 30-day bandwidth. However, the estimated policy effect is sensitive to the choice of bandwidth, with the magnitude of the effect almost halved when extending the bandwidth to 40 days. The sharp contrast between the effects of policy can be explained by the two

Table 4: Geotagging in Comment on Non-mainland Participation

	Non-mainland Comment				
Bandwidth Days =	**30**	20	25	35	40
Comment geotag	**0.0101****	0.0142**	0.0155***	0.00823*	0.00590
$t \geq$ Apr.28	**(0.00478)**	(0.00623)	(0.00573)	(0.00469)	(0.00446)
Observations	**464,000**	281,924	358,752	561,447	642,200
Control Mean	**0.0486**	0.0462	0.0508	0.0483	0.0482

Notes: This table presents the impact of geotagging in comments on non-mainland user's participation. The major difference between this table and Table 2 is that this table reports the estimate varying the bandwidth.

policies' different abilities in encouraging comments. Revisiting Figure D.12, the proportion of comments related specifically to IP location surged around five times higher during the first few days of comment tagging than the first few days of profile tagging. When the geographic location is attached to the profile, users can observe the policy change simply by viewing their profiles. Meanwhile, when the geographic location is attached to the comment, users must comment on something to witness the policy change. Therefore, non-mainlanders' positive response to the comment tagging n can be explained by their disproportionate interest in testing the policy.

By extending the bandwidth horizon, Figure D.22 illustrates the drastic difference when using a narrow bandwidth (30 days) and a wide bandwidth (120 days). While the left panel shows a positive policy impact at the cut-off, the fitted time trend displays a negative slope that consistently evolves, eventually leading to the negative estimate on the right when more weights

are assigned towards the potentially negative long-run effect. Table A.4 illustrates how the positive impact is gradually displaced by the negative effect, addressing the inconsistent treatment effect issue we encountered when estimating the local policy effect.

6.1.3 Group difference

Prediction 2 suggests that the revealing effect is amplified with the group differences in the media narratives. We test this prediction using two methods. In the first method, we estimate the RD specification with the poster's location heterogeneity. In the second method, we construct a measure approximating the sentiment bias between posts and then estimate the RD specification by interacting with the measure.

The intuition of the first approach is that the poster's location predicts the mean sentiment toward a specific location illustrated in Figure 2. Table 5 implies that the heterogeneous poster's location indeed impacts the policy in the direction that is consistent with the Prediciton 2. Profile Tag coefficient is the policy effect on non-mainland users' participation in the posts written by mainland posters. Profile Tag $\times$ NM poster coefficient is the additive effect when non-mainland posters wrote posts. We find Profile Tags are significantly positive (at 10% level) in columns (3) and (4), where the estimates are based on the donut sample. Their magnitudes are comparable to the ones we obtained from the original RD specification, which does not incorporate the heterogeneity of the poster location.

On the other hand, none of the Profile Tag $\times$ NM poster coefficients is significant. Moreover, none of the Wald tests of the equivalence between the first variable and the summation of the first and second variables has its null hypothesis rejected. Nevertheless, the reduction rate of non-mainland users in the mainland posters' posts consistently dominates the reduction rate in the non-mainland posters' posts. For example, according to column (3), the reduction rate is 41.7% (-0.0168/0.0403) for mainland posters and 33.5% ((-0.0168-0.0101)/0.0802) for non-mainland posters.

In the second approach, we use the sentimental bias measured by Equation 13 to evaluate the group differences. The measure is straightforward in measuring the sentimental distance expressed in one post compared to the benchmarked sentiments of non-mainland posters. We conduct this analysis among a subset of posts discussing COVID-19. Therefore, the gap in location sentiment can be interpreted as media bias when different posters describe

Table 5: Poster Location Heterogeneity on Revealing Effect

	(1)	(2)	(3)	(4)
	Non-mainland comments			
Profile Tag	-0.00829	-0.00635	-0.0168*	-0.0353*
	(0.00568)	(0.00866)	(0.00963)	(0.0207)
Profile Tag × NM poster	-0.00382	-0.00821	-0.0101	-0.0493
	(0.0114)	(0.0160)	(0.0178)	(0.0467)
Control Mean M	0.0399	0.0399	0.0403	0.0403
Control Mean NM	0.0782	0.0782	0.0802	0.0802
Observations	385,086	385,086	316,537	316,537
Linear	✓	✗	✓	✗
Quadratic	✗	✓	✗	✓
Donut	✗	✗	✓	✓

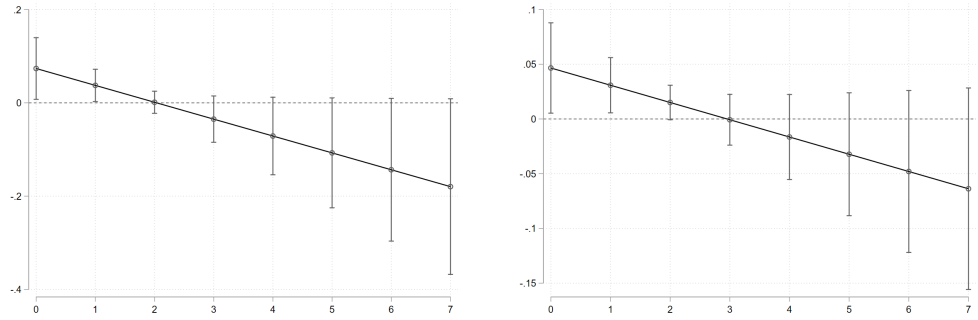
Notes: The coefficients are obtained by regressing Equation 12 with a non-mainland poster indicator interacting the treatment dummy and time trends. The rest of the regression specifications are the same as in the previous tables. Control Mean M (NM) represents the proportion of non-mainland comments in mainland (non-mainland) posters' posts before the policy implementation.

the location.

Figure 4 shows the marginal effects of profile tagging at different levels of sentimental gaps. The coefficients are derived from estimating Equation 12 with quadratic time trends and interacting the treatment dummy and time trends with the sentimental gaps. The left figure uses all the posts, while the right figure uses only the posts written by mainland posters. The right figure addresses the correlation of sentiments between non-mainland posters.

These two figures show several patterns mostly consistent with Prediction 1. First, the revealing effect is enhanced by the sentimental gap. We find the marginal effects of profile tagging decrease monotonically with the standardized sentimental gap. Second, the revealing effect is non-negative when the gap is closed. We find the marginal effects are weakly positive when the gap is zero and insignificant from zero when the gap is at its minimal (around 1.35). Being restricted to a specific topic truncates the sample size and the statistical power. Nevertheless, most of the marginal effects with significant gaps statistically differ from those when the gap is zero in both figures.

Figure 4: Marginal Effects of Profile Tagging At Different Levels of Sentimental Gaps



Notes: The horizontal axis represents the standardized sentimental gap. The sentiment gap is calculated using the post's sentiments toward US, China, and Shanghai. Missing sentiments are replaced with 0. The minimum gap is 1.35, and the maximum is 6.74. The vertical axis shows the reduction in non-mainland comments. The confidence intervals are at the 90% level.

6.1.4 Potential channels

Geographical stigma

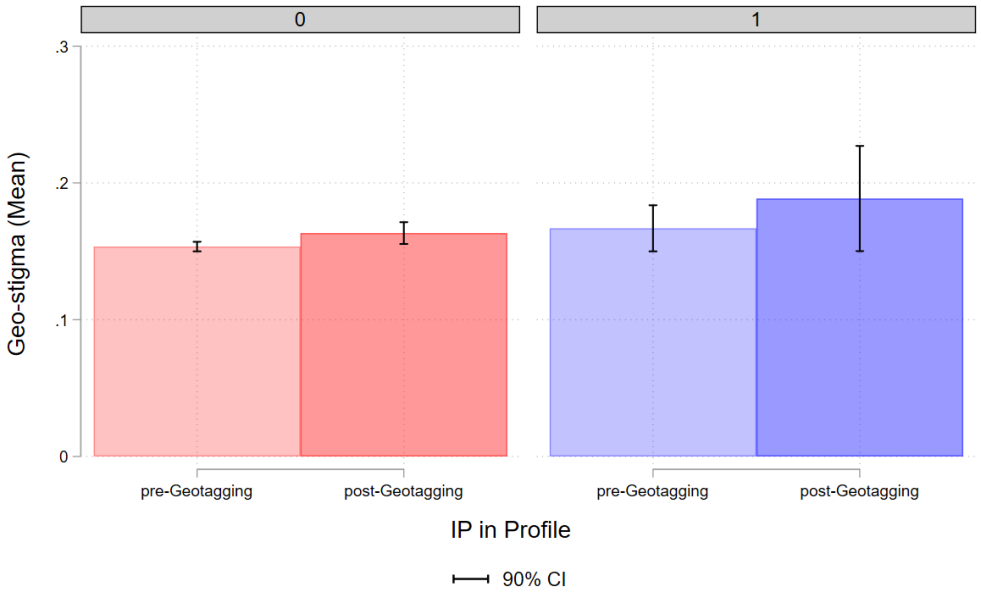
The heterogeneous result points to a potential channel that identity revelation exacerbates the geographical stigma. It is plausible that geographical stigma prevailed in posts written by posters who shared divergent beliefs with their users.

Figure 5 presents suggestive evidence that non-mainlanders received a more substantial geographical stigma after the platform geotag users in the profile. The left panel shows the proportions of geo-stigma received by disapproval comments made by mainland commenters, while the right panel shows the proportions received by non-mainlanders. Comparing the bars within each plot, we see that both groups of commenters received a stronger geo-stigma. Comparing the pre-Geotagging bars across the plot, we see non-mainlanders received a weakly higher geo-stigma than mainlanders before geotagging. Comparing the post-Geotagging bars across the plot, we see that geo-stigma increased more for non-mainland commenters who disagreed.

Contents and sentimental changes

It is also plausible that posters foresaw the reduction of non-mainland users and responded by producing contents or sentiments that are less favorable

Figure 5: Mean Geo-stigma Received by Mainland vs. Non-mainland Disapproval Comments



by non-mainland users. Though such equilibrium shifts are plausible in the long-run, they are unlikely to be true in a narrow 40-day bandwidth. Nevertheless, we create measures for posters' contents and sentimental changes respectively and find neither of the measures show discontinuity at the policy implementation date.

We test this possibility by running the RD balance test. In terms of the content, we average eight types of posts across poster-date to measure the daily trend post types each poster created. Figure D.18a shows the RD estimates of profile revelation on post types each poster create. These coefficients suggest that there is no evidence showing that posters systematically responded to the policy by altering the content they created.

We conduct a similar exercise for sentiments. To measure the daily trend of sentiments produced by each poster, we average the post sentiments barometer toward specific locations across poster-date. Figure D.18b suggests that there no systematically sentimental change found among our

posters.

Censorship

The reduction of non-mainland commenters can also be a result of targeted censorship. The pro-Western sentiments expressed by the non-mainland commenters could lead to stronger censorship after policy revealing their identities.

The strength of censorship is notoriously hard to measure. The traditional procedure requires repeated scraping of the same URL to identify the censorship (King et al., 2013). However, as we use historical data, censorship of the post should have been done already.

Fortunately, we can exploit a platform loophole to indirectly evaluate the strength of within-post censorship. One prevailing method of censorship on Sina Weibo is *Shadow Banning*. If a comment is shadow-banned, it is only visible to the poster and original commenter. Therefore, while the number of comments displayed on the post does not change, shadow-banned comments cannot be seen or scraped[12]. This means that we indirectly measure the strength of within-post censorship by calculating the number of missing comments.

Figure 6 shows the RD regression of the number of missing comments, average within poster-date cells. We study a subsample of posts that shows a comment number less than 500 as the Weibo data repository that we scraped only displayed the newest 500 comments. The outcome variable measures the strength of within-post censorship conducted towards a poster on a specific date. The vertical axis's positive (negative) value indicates if the policy increased (decreased) shadow-censoring. The plots suggest that there is no evidence that the strength of censorship has correlated with the geotagging. Therefore, these insignificant results rule out the possibility that the reduction of non-mainland comments results from differentiated censorship.

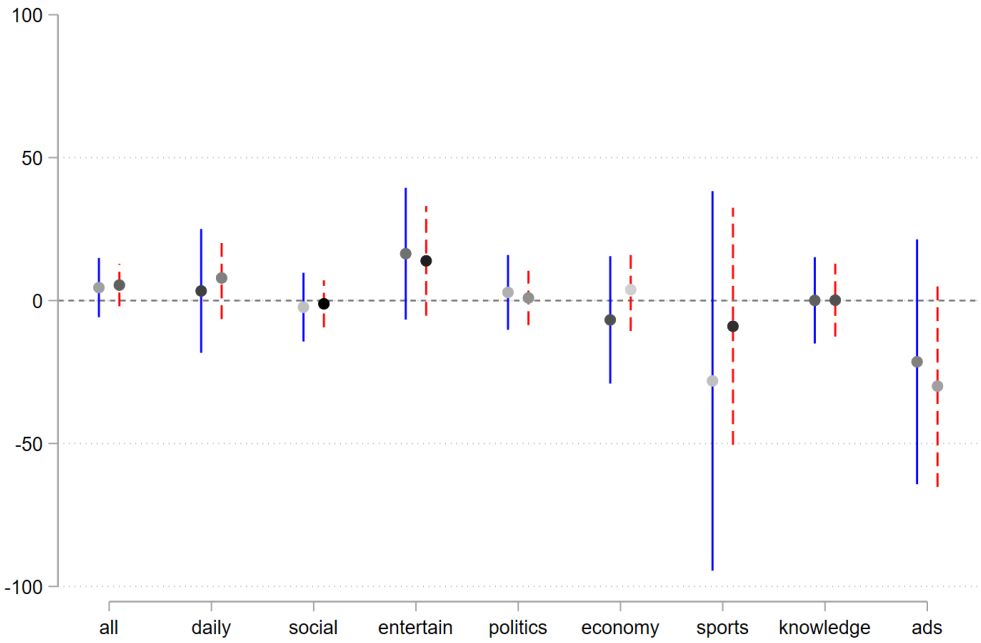
We admit that there are two limitations associated with this test. First, it only measures within-post censorship. Censorship towards the post cannot be measured unless we repeat a real-time scraping of the same post's URL. On the other hand, it is plausible that the strengths of these two types of censorship are highly positively correlated. Therefore, measuring within-post censorship predicts post-wise censorship, and we postulate that the strength

[12]On the other hand, the number of comments decreases when the poster deletes the comments.

of post-wise censorship did not vary due to the policy.

Second, there are other potential reasons that lead to the mismatch of comment numbers. Anecdotally, Weibo over-displays the comment numbers for paid promotion. However, there is no clear incentive for varying promotion strategies due to geotagging. For example, as shown in Figure 6, there is no strong evidence that within-post censorship varies significantly due to the policy in entertainment and advertising posts, where paid promotion usually prevails.

Figure 6: Continuity of censorship



6.2 Anti-Western sentiments

Next, we study if the policy influenced the opinion diversity on the platform. Hypothetically, reducing non-mainland users would amplify the representation of mainland users, who are, on average, more inclined to a xenophobic narrative. We test this hypothesis by estimating Equation 12 with a set of xenophobic dummies as the outcome variable. We restrict the analysis to a

subsample of posts involving at least one foreign entity to ensure some users were triggered to express their sentiments toward foreign countries.

Table 6 shows the results of the donut RD specification with the anti-Western comment dummy as the outcome. The coefficients represent the change of anti-Western sentiments in the comments. In general, we observed a substantial increase in anti-Western sentiments in comments. Column (1) shows the level of sentiment increased by 34% (0.174/0.5114) in the specification with linear time trends. Column (3) shows the effect of policy among mainland comments and finds that mainland comments increased the sentiment by a magnitude equivalent to column (1). Columns (2) and (4) report a similarly large magnitude using quadratic time trends, but the estimates are imprecise.

Table 6: Policy Effects on Anti-Western Sentiments

	Comment contains anti-Western sentiments			
	(1)	(2)	(3)	(4)
Profile Tag	0.174***	0.213	0.171***	0.168
	(0.0565)	(0.141)	(0.0563)	(0.140)
Comment Sample	ALL	ALL	Mainland	Mainland
Control Mean	0.5114	0.5114	0.5200	0.5200
Observations	14,049	14,049	13,161	13,161
Linear	✓	✗	✓	✗
Quadratic	✗	✓	✗	✓

Notes: The coefficients are obtained by regressing Equation 12 with the anti-Western indicator as the outcome. All four columns use the donut sample that we drop one week observations before and after the cutoff date. The rest of the regression specifications are the same as in the previous tables.

We also study where the policy exacerbates the anti-Western sentiments. Figure D.17 plots the coefficient of estimates restricted to six topics: politics, social news, economy, knowledge, entertainment, and sports. Unsurprisingly, anti-Western increased significantly in topics that involved more controversies and was unaffected in topics less likely to include controversies. The sentiment increased significantly in social news (0.312, p-value = 0.66), increased but marginally significant in politics (0.201, p-value = 0.20) and economy

(0.298, p-value = 0.42[13]). The sentiment was unaffected or even marginally decreased in entertainment, sports, and knowledge.

6.2.1 Bridging group participation with opinion change

We have shown that the geotagging policy has reduced the participation of non-Mainland users and increased the anti-Western sentiments. Are both phenomena co-moving with each other, and does there exist a causal link connecting them? To explore the co-movement, we repeat the practices we conducted to verify Prediction 2.

Figure D.19 shows the marginal effects of poster location heterogeneity on anti-Western sentiments, and Figure D.19b shows the sentimental gap heterogeneity. Posts written by mainland posters lost most of the non-mainland commenters and experienced an increase in anti-Western sentiments in their posts. A similar pattern can be found in posts that expressed sentiments that further deviated away from the non-mainland posters' mean sentiments.

Combining these two findings, we hypothesize that geographic identity revelation caused the stronger anti-Western sentiment by crowding out non-mainland users. The immediate and substantial policy effect is not only due to the selection effect in which the policy incentivizes the commenters to be more gathered to their favored posts instead of the real-time attitudinal change. The crowding out of non-mainland users likely causes the increase in anti-Western sentiments among mainland users. The demography shift leads to the norms shift on social media regarding anti-Western sentiments, making these sentiments more acceptable than ever before.

6.3 Robustness

In the last part, we test our findings' robustness by leveraging the geotagging policy's experimental phase. Before the universal implementation of IP-location revelation, the website briefly experimented with comment anonymity reduction with only posts that included keywords "Russia" or "Ukraine." The experiment started on March 4, 2022, two weeks before geotagging in profile. We test whether the findings from the previous sections are replicated in this short-period experiment through a difference-in-differences

[13]The imprecise estimates of policy impact in economy posts may result from low sample size: only 775 comments were made to economy posts that contain at least one foreign entity in the sample bandwidth.

design. Given that, almost all Russia-Ukraine posts included these two keywords, and that all of the Russia-Ukraine posts that appeared during this period are political, we use all the other international political posts as the control group. Figure D.21 shows the pretrends test for the four interested outcomes. We do not find notable pretrends that could confound the analysis.

Table A.5 shows the difference-in-differences estimates of impacts of geotagging on two outcomes among the posts that included keywords "Russia" and "Ukraine." The variable RU_KW×Post RU reports the DID estimate. Last, we do not observe that the policy increased the anti-Western sentiment in these posts.

A major challenge of the identification of using DID in our context is that the poster can intentionally select into or opt out of the treatment by including or excluding the keywords "Russia" and "Ukraine." To address this concern, we predict the likelihood of using the keywords in Russia-Ukraine-related posts using the poster's baseline (before March 4, 2022) characteristics. Figure D.20 presents the coefficients of the estimate of regressing whether the poster included two keywords in the Russia-Ukraine posts after the policy on a set of poster's baseline characteristics. We find no strong evidence that a particular type of poster is more likely to include the RU keywords other than the ones that already using these words in the baseline.

The DID estimates provide additional results that could not be obtained using RD. First, we find a marginally significant increase in non-mainland comments in untreated political posts during that time. This evidence may suggest the potential crowding out when the policy was implemented locally. Non-mainland users were driven to the untreated posts to evade the geographical stigma. Second, we find substantial decreases in anti-Western sentiment in the untreated political posts. This decrease resonates with the increase of non-mainland comments in these posts.

Parallel to the robustness test, the already-treated Russia-Ukraine Posts also present a credible placebo test. As the comments made to these posts are already tagged with commentators' IP locations, the later universal policy implementations should not have any impact other than spillovers. In Table A.6, we show that for all of the outcomes we studied, comment-level anonymity-reduction only significantly increased the non-mainland comments in these posts. The drastic increase, instead of being caused by the policy, may be explained by the spillover of the temporary increase in non-mainland comments throughout the whole platform.

In general, we do not find strong evidence that the effects on non-mainlanders'

presence and anti-Western sentiment are the artifacts resulted from other shocks.

7 Conclusion

This article examines the effects of reducing a social media platform's anonymity. It identifies the causal effects by leveraging a unique policy experiment in China.

Overall, we find reducing anonymity crowds out the participation of non-mainland Chinese users on one of the largest Chinese social media platforms. Besides the impact of policy on the comments' sentiments, the effects vary with the posters and posts' geographic identities and sentimental gaps. We provide suggestive evidence that the policy enhances the geographical stigmatization of non-mainland users, whose exile further exacerbates the anti-Western sentiments on the platform.

The findings have two implications for understanding social media ecology and policy design. First, a moderate level of anonymity regulation can affect the representativeness of the social media platform. Systematic behavioral change is likely triggered by discriminative treatment, such as the geographical stigma against specific geographic groups. While our study is China-specific, the result may be generalizable, given that many global platforms have implemented verification programs to inhibit the transmission of misinformation and harmful action and may extend to platform-wise anonymity reduction within a legal scope.

Second, as one type of public opinion, social media opinions and sentiments have been evaluated by policymakers when designing policies. In this paper, we show that regulation crowds out dissent and segregates the opinions and user groups. This may lead to social media opinions deviating from public opinion and distorting the policy design based on social media data. We believe that more efforts should be made to examine the backfire effect of Internet regulations on future policy design.

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.

King, G., Pan, J., & Roberts, M. E. (2017). How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review*, *111*(3), 484–501.

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.

Singer, P. (2011). Visible man: Ethics in a world without secrets. *Harper's Magazine*, *34*, 47.

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.

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 Tables

Table A.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 A.2: 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 A.3: Mainland vs. Non-mainland Comments

	Mainland			Non-mainland			
	N	mean	sd	N	mean	sd	diff
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***
Sentiments							
Negative	13,209	0.40	0.49	12,505	0.36	0.48	-0.037***
Neutral	13,209	0.39	0.49	12,505	0.44	0.50	0.046***
Positive	13,209	0.21	0.40	12,505	0.20	0.40	-0.009*
Anti-western	40,000	0.52	0.50	4,300	0.44	0.50	-0.075***

[1] 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

[2] The difference is calculated by regressing the characteristic on a non-mainland dummy without adjusting the standard error.

Table A.4: Impact of Reveal IP Location in Comment on Non-mainland Participation, Extended Bandwith

	Non-mainland Comment			
Bandwidth Days =	**120**	60	80	100
IP in Comment	-0.00707***	0.00476***	-0.00168**	-0.00303***
	(0.000631)	(0.000888)	(0.000765)	(0.000679)
Observations	1,891,324	995,425	1,304,044	1,639,899
Control Mean	0.0515	0.0523	0.0518	0.0521

Notes: This table provides the additional result to the main table by extending the bandwidth. The main results are derived by estimating the main regression discontinuity equation, with the outcome being the presence of non-mainland comments. The control mean is the proportion of non-mainland comments within the bandwidth days before the policy implementation. The regression includes the set of controls defined in the previous section, as well as day-of-week and post-type fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table A.5: Robustness Check Using Treated Russia-Ukraine Posts

	Non-mainland	Anti-Western
$Treat^{RU}$	-0.0241***	0.0229
[=1 if keywords]	(0.00829)	(0.0224)
Post	0.0499*	-0.0848**
[=1 if t>Mar04]	(0.0257)	(0.0337)
$Treat^{RU} \times Post$	**-0.0455***	**-0.0437**
	(0.0263)	(0.0448)
Observations	116,830	15,483
Control Mean	0.0710	0.5216

Notes: This table presents the difference-in-differences estimate of comment de-anonymization during Weibo experimented with the function only on Russia-Ukraine Posts. The outcomes are the four comment features used in the RD regressions. The estimation sample comprises all the Russia-Ukraine posts and international politics posts that appeared in our data between January 1 and March 4, 2022. $Treat^{RU}$ indicates if the post includes the keyword "Russia" or "Ukraine." Post indicates if the post was written on or after March 4, 2022. $Treat^{RU} \times Post$ captures the DID estimate. The regression includes the set of controls as well as the day-of-week fixed effect.

Table A.6: Placebo Test Using Treated Russia-Ukraine Posts

	(1) Non-mainland		(3) Anti-Western	
	profile	comment	profile	comment
Reveal	0.0147	0.0213**	-0.0293	0.0387
	(0.00957)	(0.00996)	(0.0723)	(0.0665)
Observations	27,669	20,044	2,346	1,631
Control Mean	0.0466	0.0385	0.0443	0.0925

This table presents the effects of policy on two key outcomes. The estimation sample is the Russia-Ukraine posts with the keyword "Russia" or "Ukraine" that triggered the comment deanonymization on March 4, 2022. The table shows the estimation results from the main regression discontinuity model using the 40-day bandwidth.

B Omitted Proofs

B.1 Proof of *Proposition 1*

By (4) and (7), we have $\underline{s}_\emptyset \leq \bar{s}_\emptyset$ if $\underline{s}_\emptyset$ and $\bar{s}_\emptyset$ exist (and the inequality holds strictly if the media slant gap $\Delta > 0$). We show the existence and uniqueness of $\underline{s}_\emptyset$; the existence and uniqueness of $\bar{s}_\emptyset$ can be established by symmetry, which we ignore in the proof. Define $F(s) = f_1(s) - f_2(s)$ where

$$f_1(s) = -\frac{1}{k}(s+b), \tag{B.1}$$

$$f_2(s) = \psi(\Delta)(1-\lambda)\frac{\hat{\Phi}(s)+1}{(1/\lambda-1)\hat{\Phi}(s)+1}, \tag{B.2}$$

and $\hat{\Phi}(s) = \Phi(\frac{s-\mu_F}{\sigma_\epsilon})/\Phi(\frac{s-\mu_L}{\sigma_\epsilon}) \in (0,1]$. Note that $0 \leq f_2(s) \leq 2\psi(\Delta)(1-\lambda)$ for any $s \in \mathbb{R}$; in particular, if $\lambda = 1/2$, then $f_2(s) = \psi(\Delta)/2$, which is constant over $s \in \mathbb{R}$. Therefore, $\lim_{s\to-\infty} F(s) = \infty, \lim_{s\to\infty} F(s) = -\infty$. Since $F(\cdot)$ is continuous, by *Intermediate Value Theorem*, there exists an $\underline{s}_\emptyset \in \mathbb{R}$ such that $F(\underline{s}_\emptyset) = 0$, and thus, (7) and (9) hold simultaneously, i.e.,

$$C_\emptyset(-1) = f_1(\underline{s}_\emptyset) = f_2(\underline{s}_\emptyset) > 0.$$

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

$$\lim_{\lambda\to1} F'(s) = -\frac{1}{k} - \lim_{\lambda\to1}\psi(\Delta)(1-\lambda)\frac{(2-1/\lambda)\hat{\Phi}'(s)}{[(1/\lambda-1)\hat{\Phi}(s)+1]^2} = -\frac{1}{k} < 0, \tag{B.3}$$

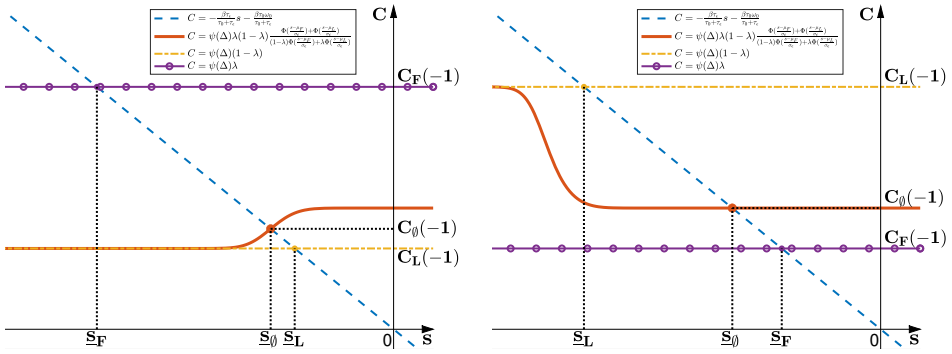
$$\lim_{\lambda\to0} F'(s) = -\frac{1}{k} - \lim_{\lambda\to0}\psi(\Delta)(1-\lambda)\frac{\lambda(2\lambda-1)\hat{\Phi}'(s)}{[(1-\lambda)\hat{\Phi}(s)+\lambda]^2} = -\frac{1}{k} < 0, \tag{B.4}$$

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

B.2 Proof of Proposition 2

Fix any group $g \in G = \{L, F\}$. By (4) and (7), we have $\underline{s}_g \leq \bar{s}_g$ if $\underline{s}_g$ and $\bar{s}_g$ exist (and the inequality holds strictly if the media slant gap $\Delta > 0$). We show the existence and uniqueness of $\underline{s}_g$; the existence and uniqueness

Figure B.1: The Threshold of Observed Signal for Criticism ($d_i = -1$)



(a) $\lambda > 1/2$ (b) $\lambda < 1/2$

of $\bar{s}_g$ can be established by symmetry, which we ignore in the proof. Define $\hat{F}(s) = f_1(s) - \kappa$ where $f_1(\cdot)$ is given by (B.1),

$$\kappa = \psi(\Delta)[(1-\lambda)\mathbb{1}\{g = L\} + \lambda\mathbb{1}\{g = F\}] > 0, \tag{B.5}$$

and $\mathbb{1}\{\cdot\}$ is the indicator function which takes the value of 1 if the statement is true and takes the value of 0 otherwise. For any $\lambda \in (0,1)$, $f_1(\cdot)$ strictly decreases in s, ranging from ∞ to $-\infty$. Therefore, the unique solution to $\hat{F}(s) = 0$ is

$$\underline{s}_g = f_1^{-1}(\kappa) = -k\kappa - b,$$

where k is given by (5), b is given by (6), and κ is given by (B.5). Thus,

$$C_g(-1) = f_1(\underline{s}_g) = \kappa > 0.$$

B.3 Proof of Prediction 1

Consider the case when the local group accounts for the majority of users on social media, i.e., $\lambda > 1/2$ — the case when the foreign group accounts for the majority $\lambda < 1/2$ is analogous and thus ignored in the proof. Drawing on Proposition 1 and 2, we need to show that (i) $\underline{s}_F < \underline{s}_\emptyset < \underline{s}_L$, and (ii) $\bar{s}_F > \bar{s}_\emptyset > \bar{s}_L$. By (4) and (7), it suffices to show that $C_F(d) = \psi(\Delta)\lambda > C_\emptyset(d) > C_L(d) = \psi(\Delta)(1-\lambda)$ for $d = \pm 1$. Indeed, by (8) and (9), it follows

from $\lambda > 1/2$ that

$$C_\emptyset(1) = \psi(\Delta)(1-\lambda)\frac{\tilde{\Phi}(\bar{s}_\emptyset)+1}{(1/\lambda-1)\tilde{\Phi}(\bar{s}_\emptyset)+1} > \psi(\Delta)(1-\lambda), \tag{B.6}$$

$$C_\emptyset(1) = \psi(\Delta)\lambda\frac{\tilde{\Phi}(\bar{s}_\emptyset)+1}{\tilde{\Phi}(\bar{s}_\emptyset)+\lambda/(1-\lambda)} < \psi(\Delta)\lambda, \tag{B.7}$$

$$C_\emptyset(-1) = \psi(\Delta)(1-\lambda)\frac{\hat{\Phi}(\underline{s}_\emptyset)+1}{(1/\lambda-1)\hat{\Phi}(\underline{s}_\emptyset)+1} > \psi(\Delta)(1-\lambda), \tag{B.8}$$

$$C_\emptyset(-1) = \psi(\Delta)\lambda\frac{\hat{\Phi}(\underline{s}_\emptyset)+1}{\hat{\Phi}(\underline{s}_\emptyset)+\lambda/(1-\lambda)} < \psi(\Delta)\lambda, \tag{B.9}$$

where $\tilde{\Phi}(s) = [1-\Phi(\frac{s-\mu_F}{\sigma_\epsilon})]/[1-\Phi(\frac{s-\mu_L}{\sigma_\epsilon})] \geq 1$ and $\hat{\Phi}(s) = \Phi(\frac{s-\mu_F}{\sigma_\epsilon})/\Phi(\frac{s-\mu_L}{\sigma_\epsilon}) \in (0,1]$. See Figure B.1 for an illustration.

B.4 Proof of Prediction 2

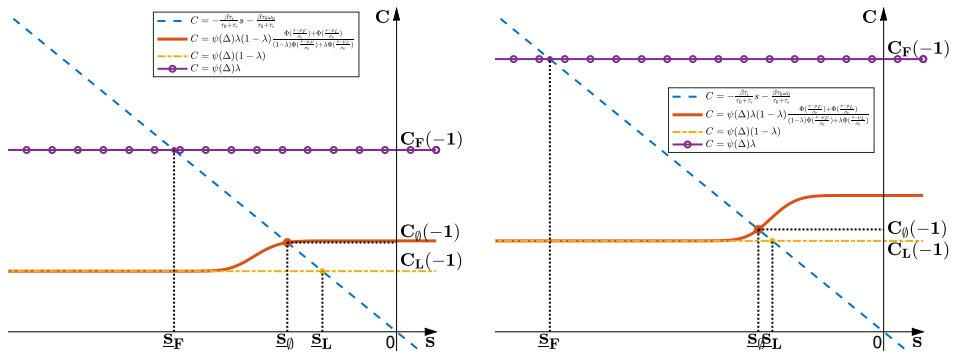
From the proof of Prediction 1, the labeling effect is driven by the gap in sanction costs across different information environments for both groups, which is in turn positively proportional to $\psi(\Delta)$. Since $\psi(0) = 0$ and $\psi'(\cdot) > 0$, we have the labeling effect is zero if there is no gap in the media slants across different groups (i.e., $\Delta = 0$) and it is amplified if the media bias gap Δ increases.

B.5 Additional Theoretical Prediction

Prediction B.5.1 (Echo Chamber) *The labeling effect is amplified by the gap between the fractions of different groups $|\lambda - 1/2|$, and it disappears when two groups are of equal size (i.e., $\lambda = 1/2$).*

Prediction B.5.1 asserts that, with group identity revealed on social media, the participation from the majority group further expands whereas that from the minority group further shrinks. This implies if one group, influenced by a particular media slant, dominates social media, then revealed group identities decrease the chances that users encounter different voices to the subject under discussion, and therefore, commenters are more likely to meet minds alike. This tendency is aggravated by algorithms that present posts according to user's preferences. Consequently, users tend to read posts

Figure B.2: Media Slants Gap Between Two Groups Amplifies Labeling Effect

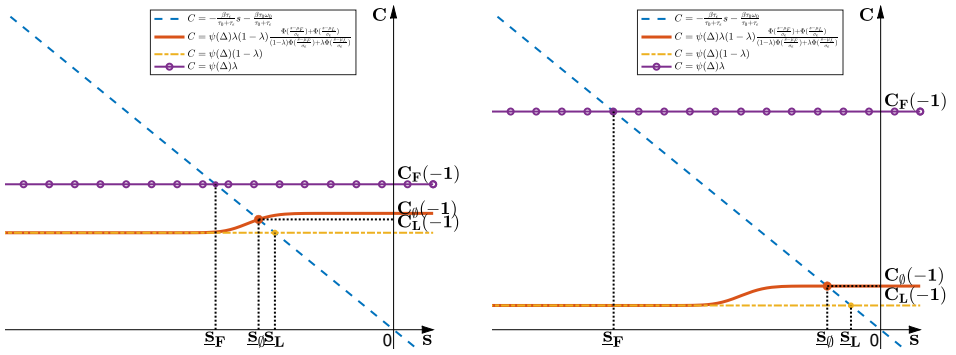


(a) Δ Small (b) Δ Large

with their group being the major audience, leading to a natural segregation of communication. This gives rise to the phenomenon of *Echo Chamber*, which contributes to polarization in opinions regarding important issues in life that deserve discussion.

Proof of Prediction B.5.1 From the proof of Prediction 1, the labeling effect increases as the sizes of different groups get more unbalanced (i.e., $|\lambda - 1/2|$ increases). In particular, if $\lambda = 1/2$, then different information environments present the same sanction costs for both groups, neutralizing the labeling effect. Intuitively, as $|\lambda - 1/2|$ grows, the majority group accounts for a larger share of the users, thus exhibiting fiercer social sanctions against the minority group when group identity is revealed.

Figure B.3: Ratio Gap Between Two Groups Amplifies Labeling Effect



(a) λ Close to 0.5 (b) λ Far Above 0.5

C VPN Usage in China

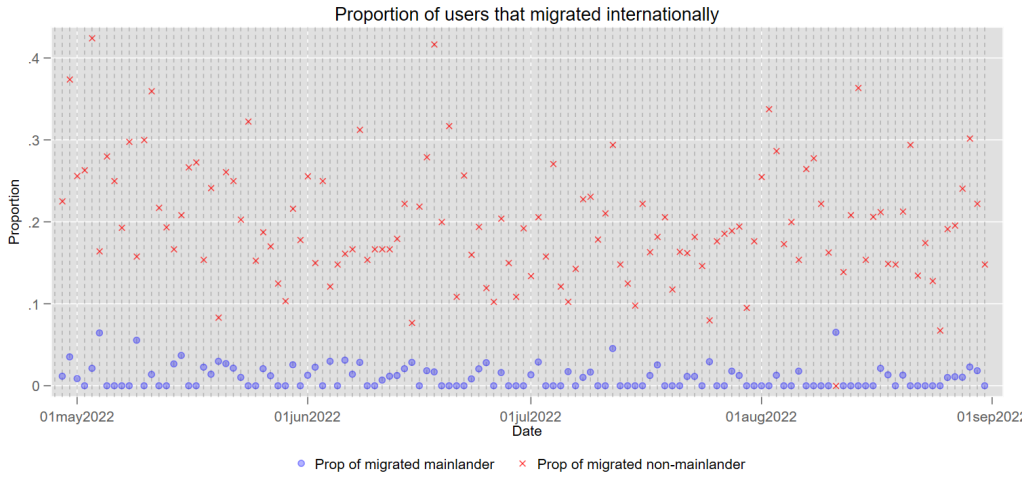
Due to the expansion of the Great Firewall project, through which the Chinese government barred mainland Chinese access to a large set of foreign platforms, including YouTube, Facebook, X (formerly Twitter), and many other highly visited websites, Chinese netizens anecdotally increased their use of VPN 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 C.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 C.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 C.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.

D Figures

Figure D.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].

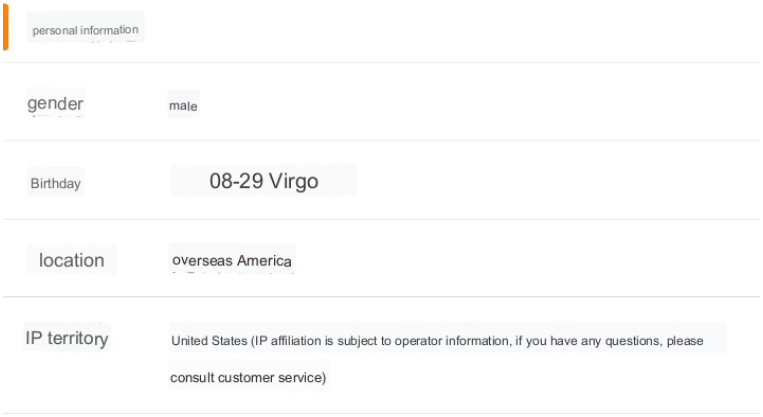
Context: [CONTEXT]

Text: [TEXT]

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

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

(a) In Profile

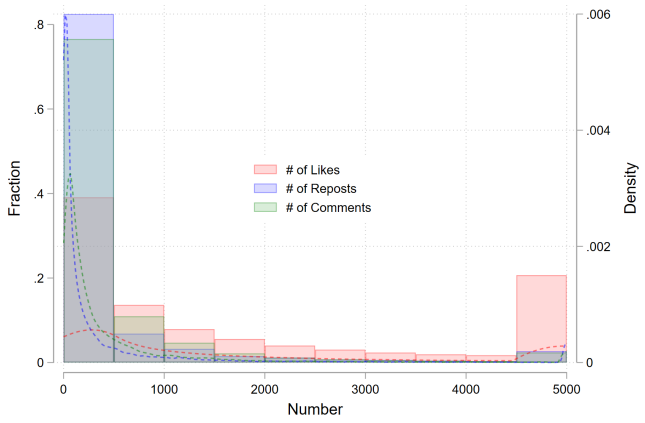


(b) In Comment



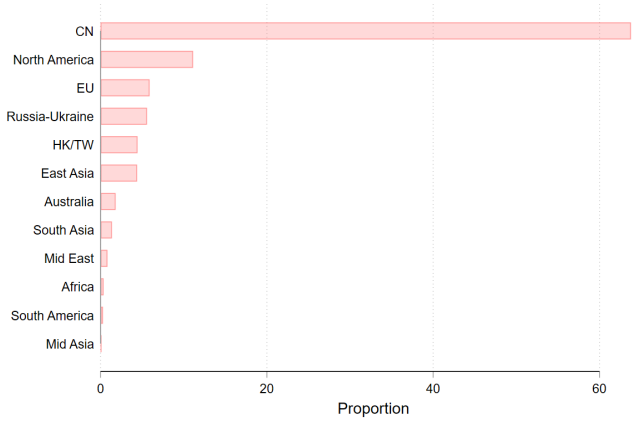
Note: The *IP territory* is the IP location

Figure D.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.

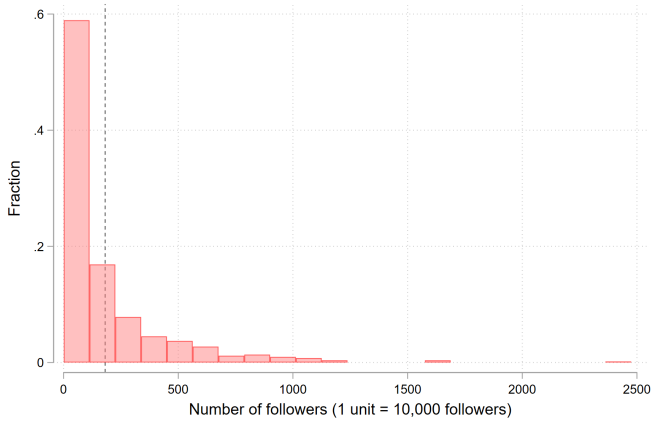
Figure D.4: The Distribution of Countries/Regions Mentioned in the Post



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 D.5: Poster's Basic Statistics

(a) The Distribution of # of Followers



(b) The Poster's IP-Location

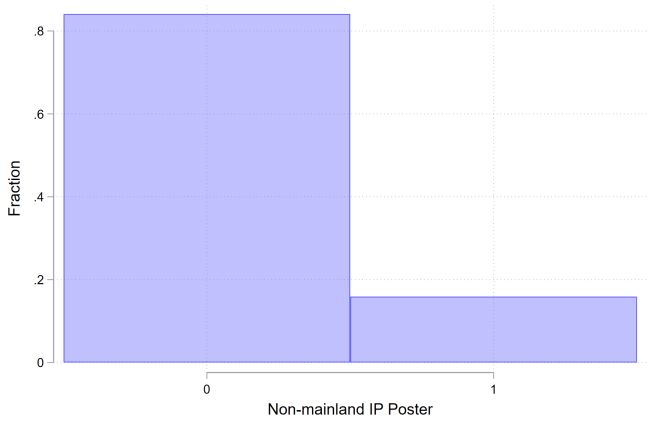
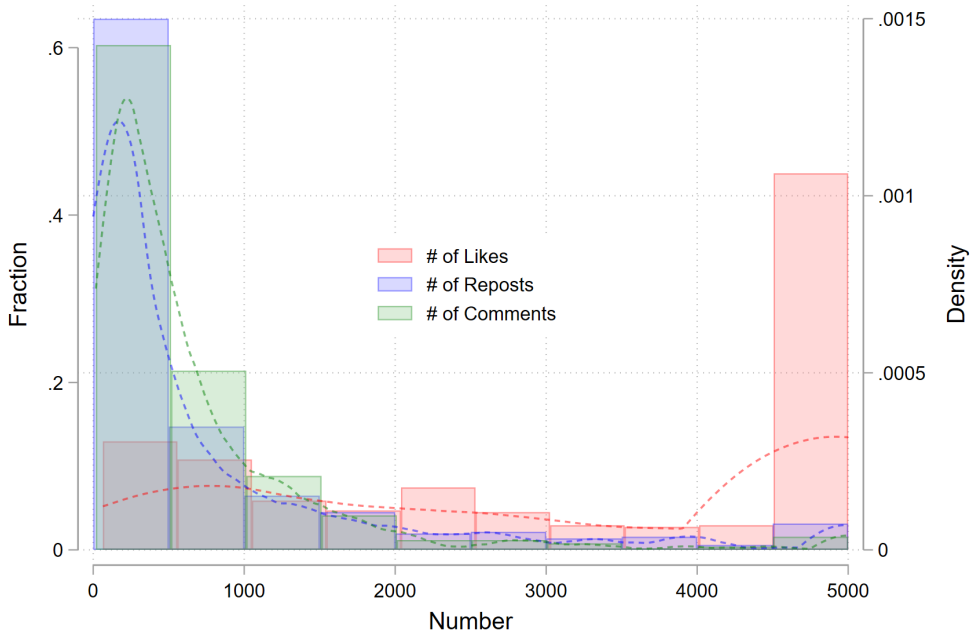


Figure D.6: The Distribution of # of Likes, Reposts, and Comments



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 D.7: Mean Sentiments Toward Locations Mentioned in the Posts

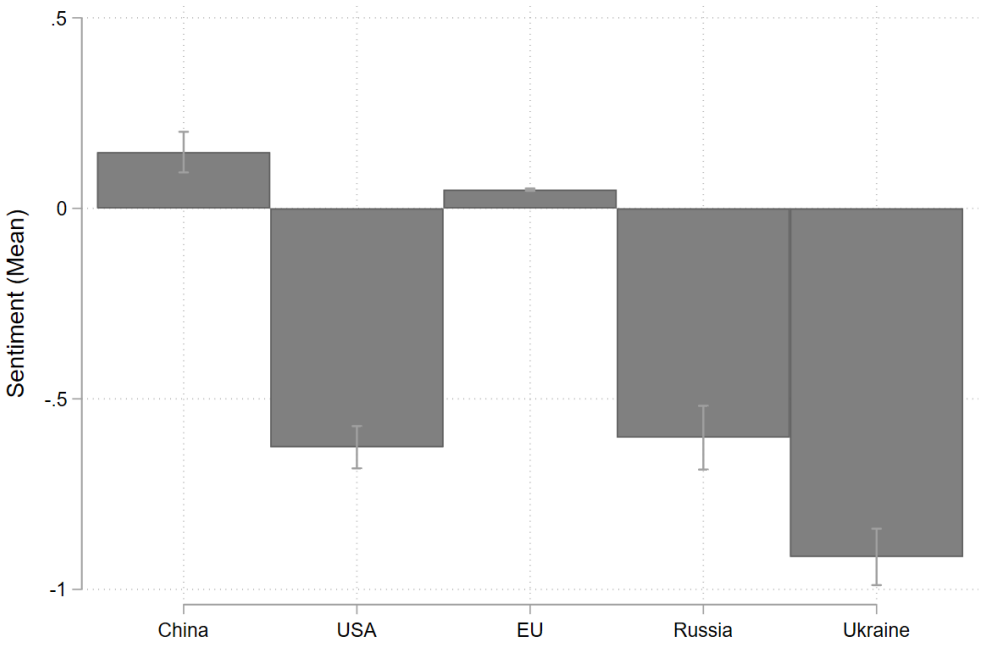
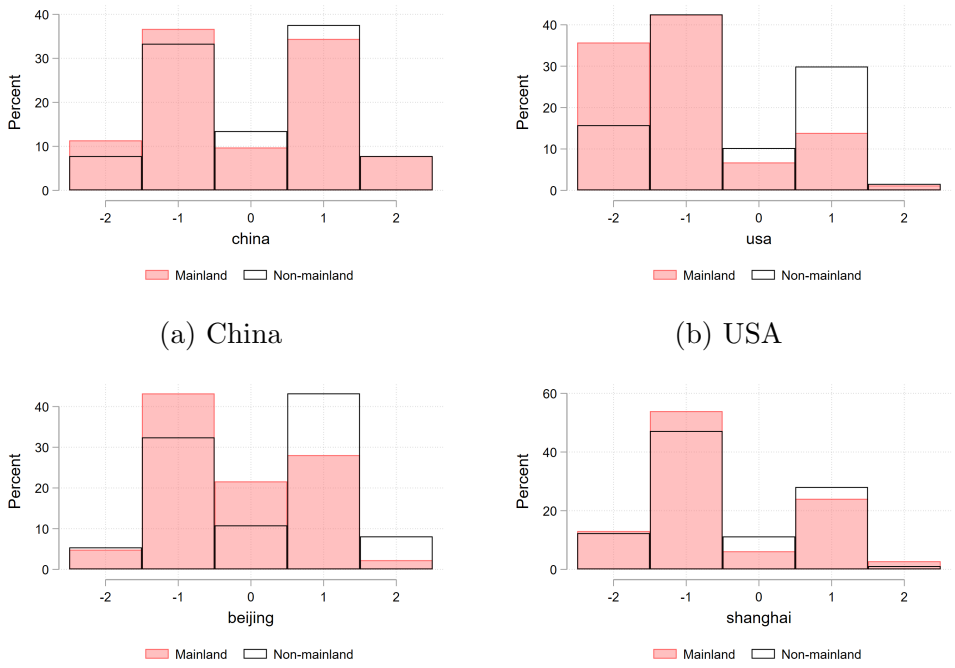


Figure D.8: Sentimental Differences in COVID-19 Posts

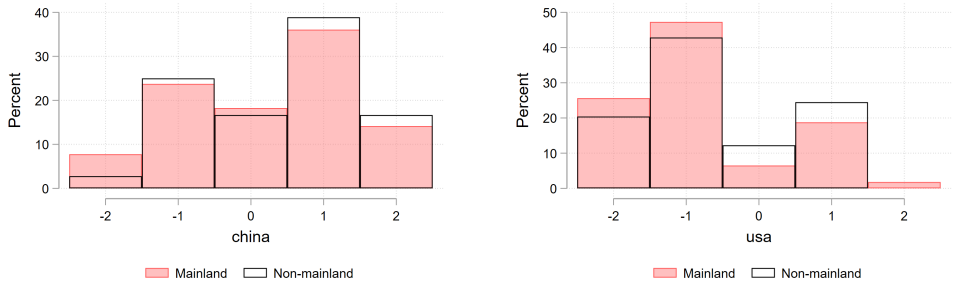


(a) China (b) USA

(c) Beijing (d) Shanghai

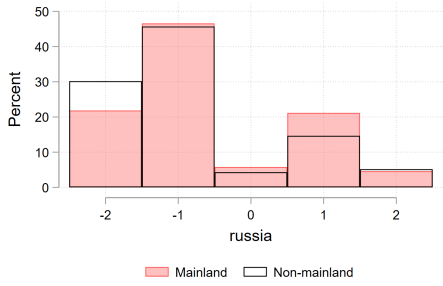
Notes: All four figures plot the proportion of sentiments, ranging from -2 (very negative) to 2 (very positive), towards four different locations. The LLM first identifies the location from the post and then evaluates the poster's sentiment towards the location based on the post's content.

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

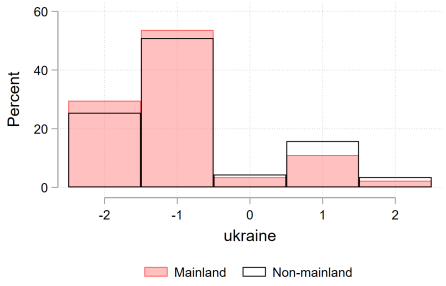


(a) China

(b) USA



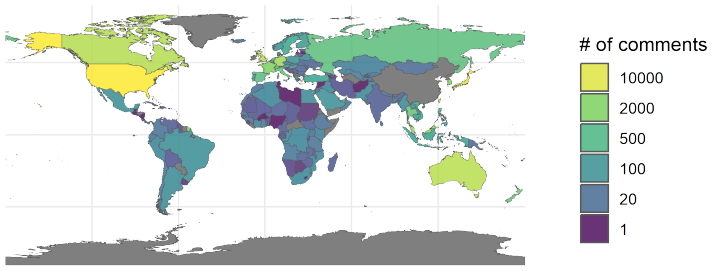
(c) Russia



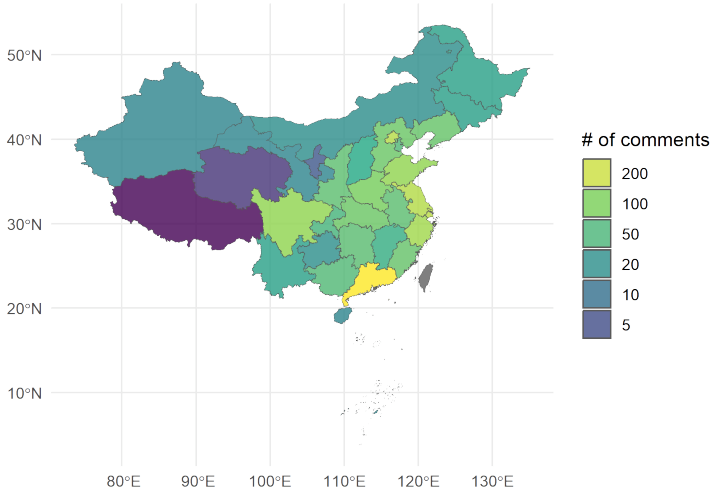
(d) Ukraine

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



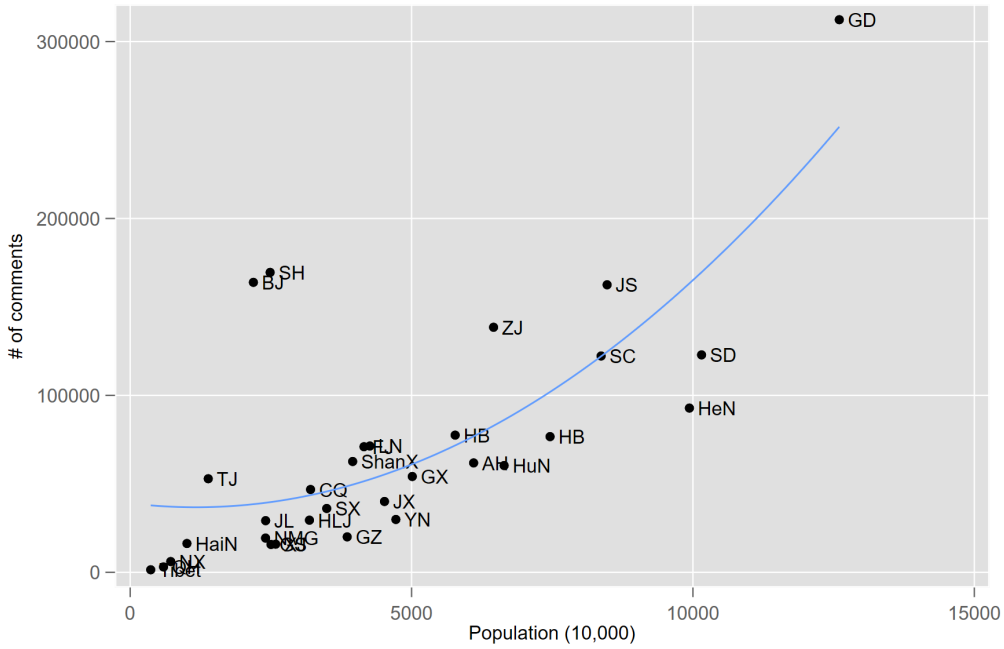
(a) Non-mainland



(b) Mainland

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

Figure D.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 D.12: The Daily Proportion of IP-related Comments

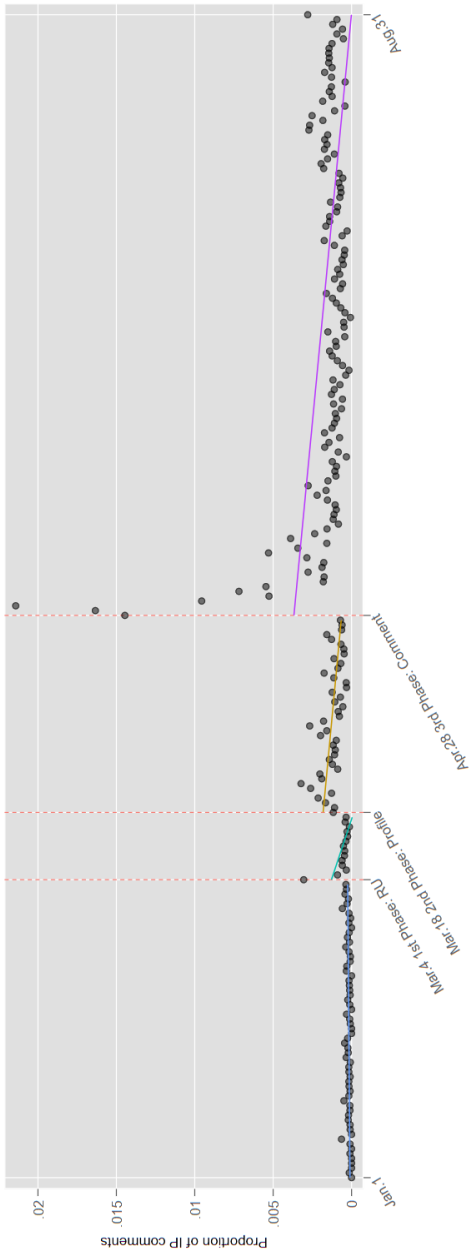


Figure D.13: RD Balance: Post's # of Likes, Reposts, and Comments

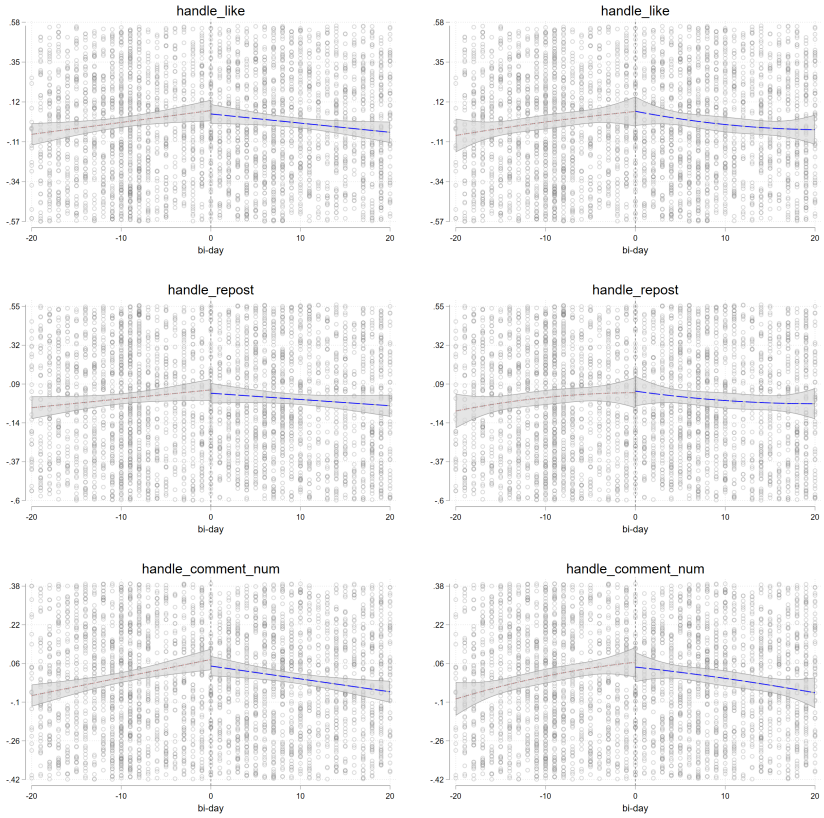


Figure D.14: RD Balance: Post's Length and Date Span

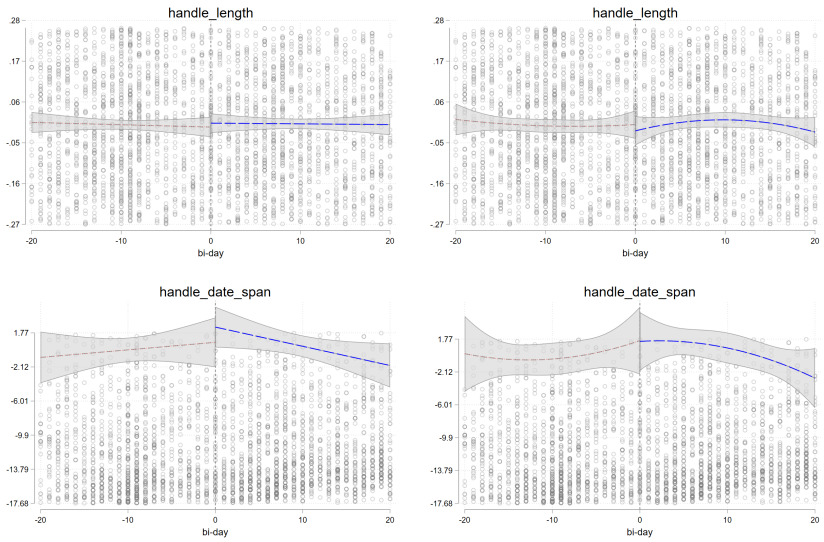


Figure D.15: RD Balance: Post's Location Sentiments

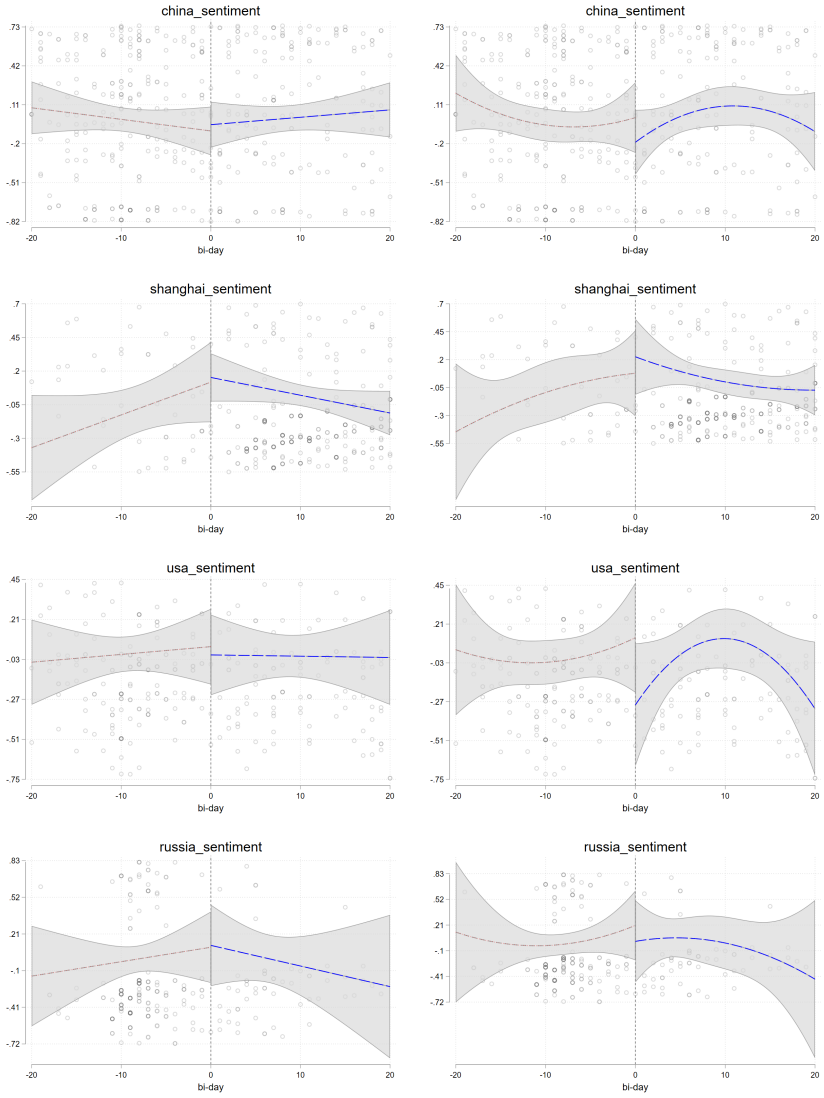


Figure D.16: Mean Geo-stigma Received by Mainland vs. Non-mainland Disapproval Comments in Mainlander's Posts

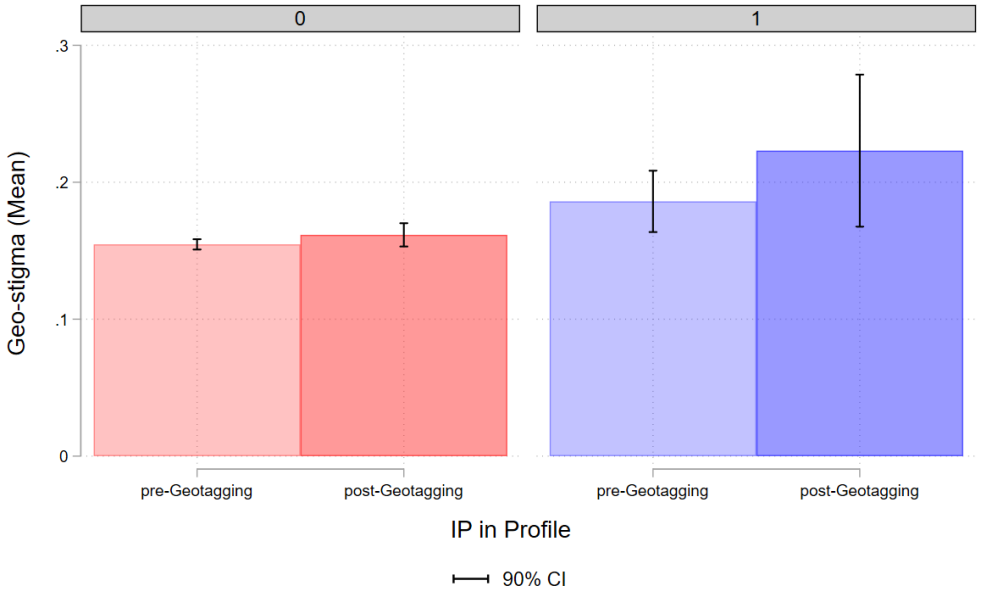
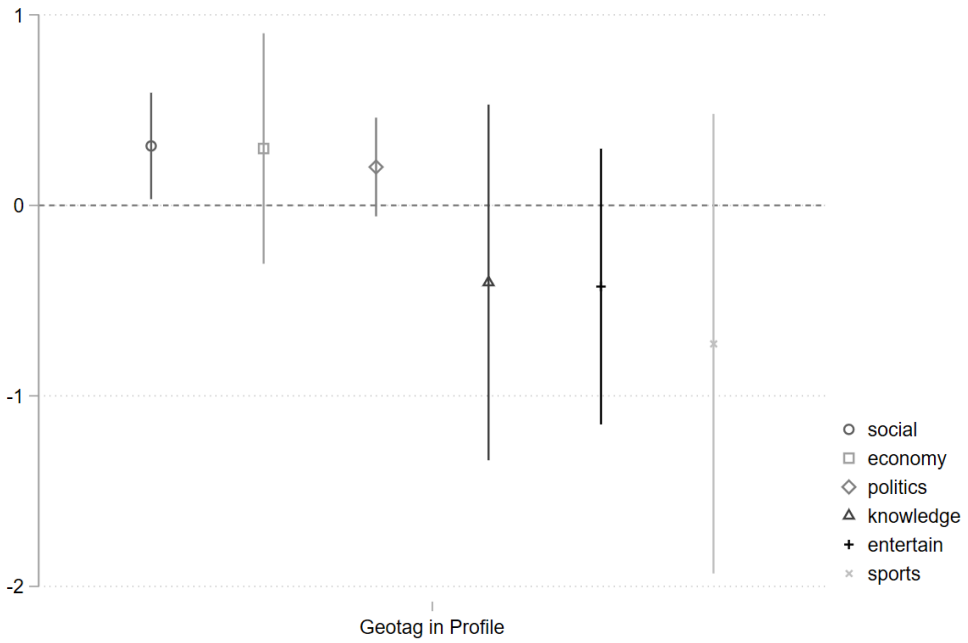
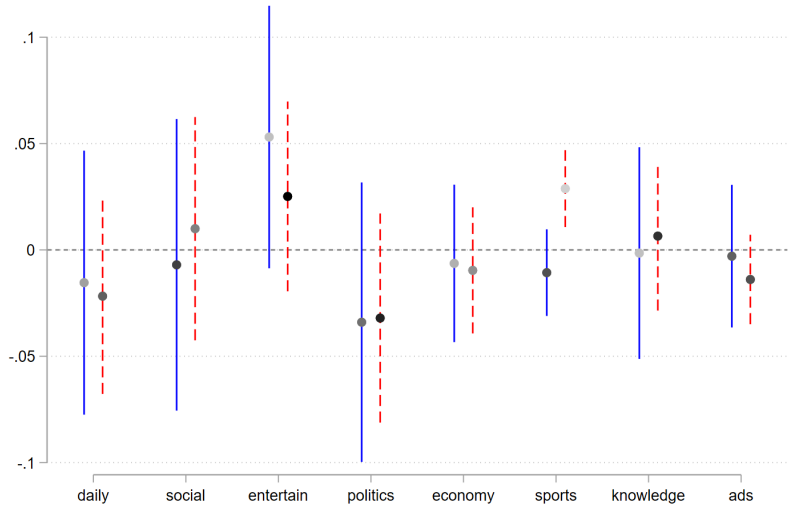


Figure D.17: Policy Effects on Anti-Western Sentiments in Different Topics

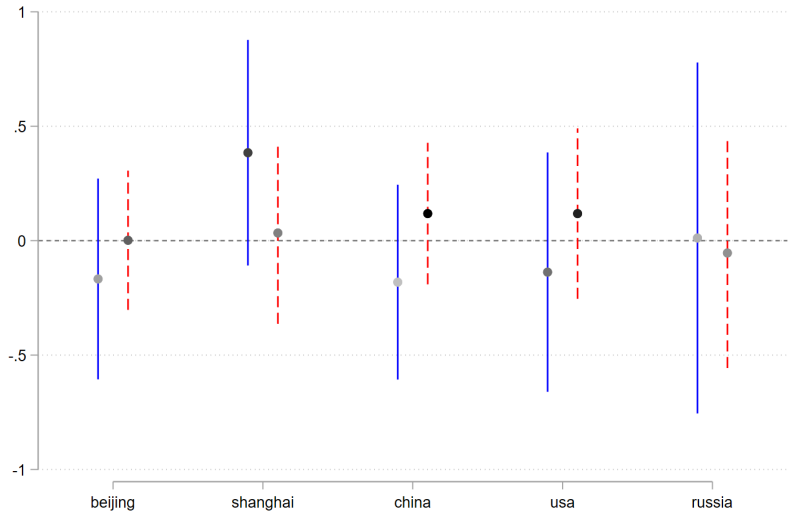


Notes: The coefficients are obtained via anti-Western donut RD regression with quadratic time trends and sample restricted to certain topics.

Figure D.18: The RD estimates of profile revelation on contents and sentiments



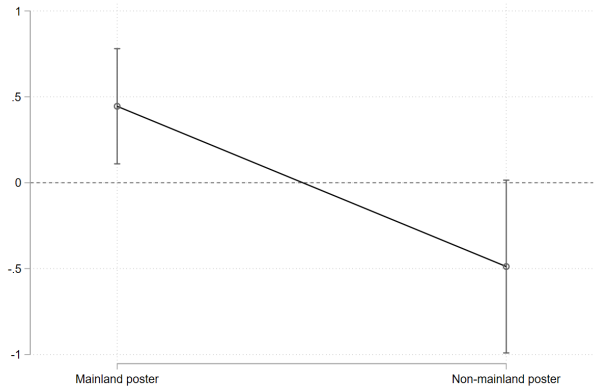
(a) Continuity of contents



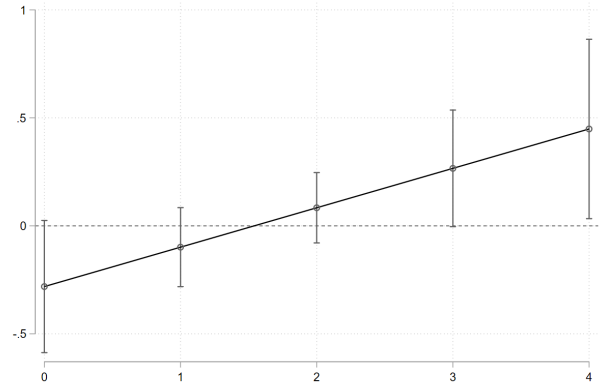
(b) Continuity of sentiments

Notes: The coefficients are obtained via RD regressions on the content indicators or sentiments barometers average across posters by date.

Figure D.19: Revisiting Poster Location Heterogeneity and Sentimental Gaps in Anti-Western Sentiments



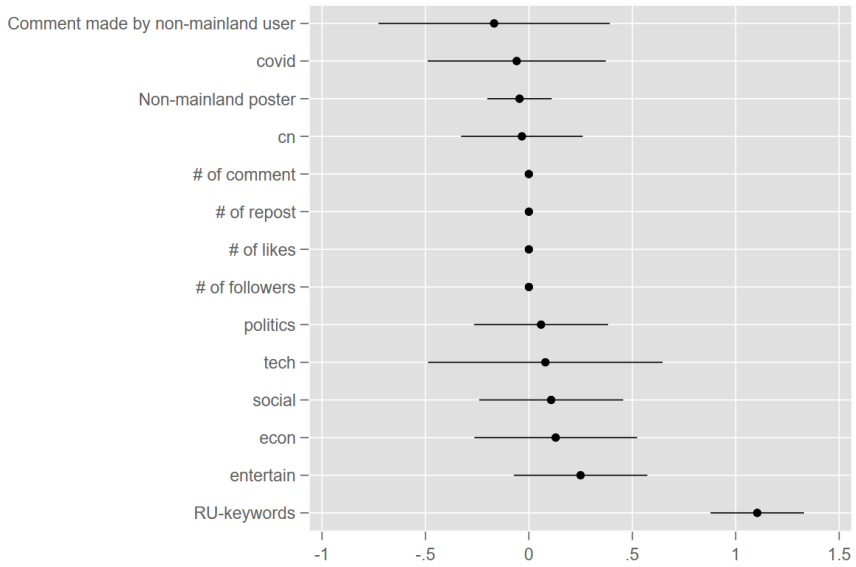
(a) Poster Location Heterogeneity



(b) Sentimental Gap Heterogeneity

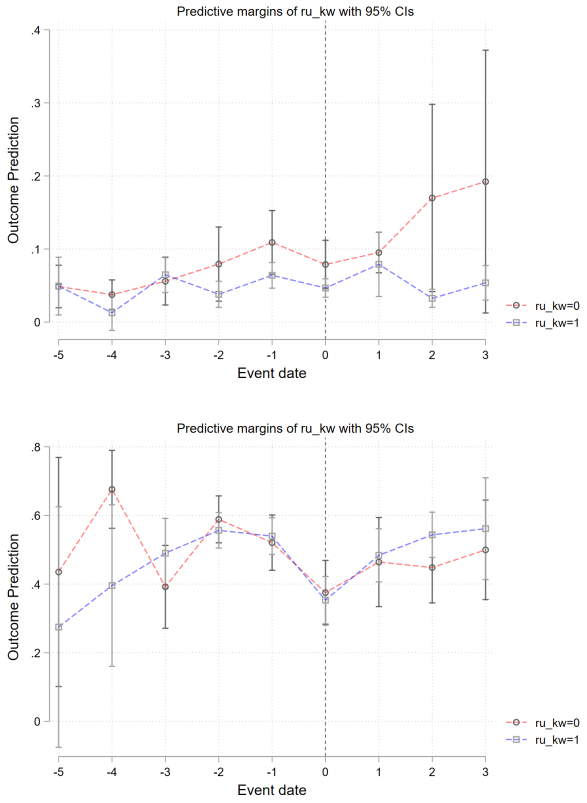
Notes: The left figure is the marginal effects of poster's location on anti-Western sentiments. The coefficients are obtained via anti-Western RD regression with the treatment dummy and quadratic time trends interacted with the poster's location dummy. The right figure is the geotagging's impact on anti-Western 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 D.20: Predicting the Poster's Likelihood to Include Keywords



Notes: The coefficients are obtained from an OLS regression that predicts the probability of a post is related to the Russia-Ukraine War. The only predictor is that it includes Chinese words of "Russia" or "Ukraine"

Figure D.21: Parallel Pretrend Test



Notes: The coefficients of these figures are obtained using the traditional event-study regression that interacts with the treatment availability with the time dummies.

Figure D.22: Policy Effect Estimated Using Narrow and Wide Bandwidth

