Location, Location, Reaction: How Mandatory IP Disclosure Silences Critics and Sparks Backlash[1]

Authors:
Sirui Li, Beijing Normal University
Ping Xu, University of Rhode Island
Yue Guo, Beijing Normal University

Abstract

Previous research has identified two competing outcomes of regulatory disclosure for negative online sentiment: chilling effects, which suppress negative online sentiment, and reactance effects, which intensify it. In this study, we apply these theories to examine the impact of mandatory IP location disclosure on citizens' negative online sentiment regarding posts on Sina Weibo, a leading Chinese social media platform. We argue that the mandatory IP location disclosure can have opposite effects on negative online sentiment depending on the type of social media context. To test this, we use a quasi-natural experiment and analyze more than 160,000 comments posted before and after the implementation of the mandatory IP location disclosure policy. Employing a regression discontinuity design (RDD), we find that mandatory IP location disclosure reduces negative sentiment under government accounts while amplifying it under non-government accounts, particularly in re-post section. These findings reveal the heterogeneous effects of regulatory disclosure and shed light on how citizens adapt their online expression strategies in response. More broadly, the study advances our understanding of the interplay among citizens, platforms, and governments in regulated social media environments.

Keywords: regulatory disclosure | negative online sentiment | chilling effects | reactance effects

---

[1] Please cite: Li, S., Xu, P., & Guo, Y. (2026). Location, Location, Reaction: How Mandatory IP Disclosure Silences Critics and Sparks Backlash. *Policy & Internet*, *18*(1), e70034.

1 | Introduction

The rapid advancement of digital technologies, ranging from artificial intelligence and machine learning to big data and quantum computing, has provided governments with powerful new tools for enhanced administrative efficiency, predictive modeling, and targeted interventions (Chen et al. 2025; Cukier and Mayer-Schoenberger 2013; Kitchin 2014). However, these advancements have also introduced complex challenges related to regulation, privacy, and freedom of expression (Yeung 2018; Zuboff 2019; van Dijck et al. 2018).

A critical yet under-theorized manifestation of this trend is the global expansion of regulatory disclosure mechanisms targeting individual users. Traditionally, regulatory disclosure in public policy has focused on corporate accountability, compelling entities to release data on finance, emissions, or safety to reduce information asymmetries (Fung et al., 2007; Leuz & Wysocki, 2016). However, contemporary digital and algorithmic governance has inverted this logic, extending the mandate of transparency to citizens. By requiring the visibility of user data, such as real-name identities or geographic locations, states seek to dismantle online anonymity and manufacture a legible, traceable public (Creemers, 2018; Li et al., 2025; Zhu, 2024).

China's 2022 directive requiring social media platforms such as Sina Weibo to publicly display users' IP location is a clear example of this strategy. Unlike the real-name registration systems of the early 2010s that operated in the backend (Jiang, 2016; Zhu, 2024), this policy enforces front-stage visibility, attaching a permanent geographic marker to every public expression. Crucially, this logic of mandatory visibility is not unique to China but reflects a deepening global trend toward digital identification. Similar outcomes are emerging in Western

democracies through platform architectures aimed at verifying authenticity. For instance, X (formerly Twitter) has rolled out account labels that display users' country information on profiles, including indicators related to VPN-linked location signals, to combat inauthentic behavior (TechCrunch, 2025). Despite different political and commercial motivations, these converging practices point to a broader shift toward making the digital citizen increasingly traceable, often justified as necessary for maintaining digital legitimacy and public order (Moss, 2025; Schroeder, 2025).

This study examines the implications of such user-targeted regulatory disclosure for citizens' negative online sentiment. Unlike direct censorship or keyword filtering, the mandatory IP location disclosure policy represents a subtle behavioral intervention. It does not remove content outright but alters the structural conditions of speech by displaying users' geographic locations, significantly constraining online anonymity and potentially encouraging self-censorship in response to heightened accountability and traceability (Li et al., 2025; Liu et al., 2024). Existing theoretical frameworks offer divergent predictions. On the one hand, the chilling effect school suggests that regulatory measures suppress negative sentiment, as individuals retreat into silence to avoid legal or social sanctions (Penney, 2019; Schauer, 1978). On the other hand, the reactance effect school proposes that perceived intrusions on autonomy trigger resistance, motivating individuals to amplify negative sentiment as a form of defiance (Brehm, 1966; Zhu & Fu, 2021; Zhu, 2024). However, the boundary conditions under which regulatory disclosure suppresses or amplifies citizens' negative sentiment remain understudied.

We address this gap by analyzing the impact of China's mandatory IP location disclosure policy on negative online sentiment. Here, negative online sentiment is defined as expressions

of dissatisfaction, criticism, or discontent (Liao et al., 2023; Zhang & Guo, 2021). Utilizing a quasi-natural experiment and a regression discontinuity design (RDD) on a dataset of 168,728 comments, we find a nuanced landscape that regulatory disclosure significantly reduces negative sentiment in government social media accounts but increases it in non-government accounts, particularly within re-post section.

These findings contribute to the existing literature in two key ways. First, we extend the theoretical scope of regulatory disclosure from its traditional domain of corporate accountability to user-targeted social media governance by focusing on disclosure requirements imposed on individual social media users. Second, we show that the effects of disclosure-based governance in digital environments are structurally conditional on where and with whom political communication occurs. By modeling account type and interaction area as moderating factors, we reveal a structural division in online political communication between official communication spaces and peer communication spaces. Our results show that the effects of regulatory disclosure are fundamentally shaped by the dynamics embedded in specific digital contexts. We demonstrate that mandatory disclosure does not simply regulate speech but stratifies the digital public sphere into a controlled zone of official communication and a contested zone of peer interactions. This bifurcation clarifies how user-targeted disclosure reshapes the ecology of online public sentiment, with significant implications for digital governance globally.

2 | From Content Control to Mandatory Visibility: The Evolution of China's Digital Governance

Social media governance presents a universal dilemma as digital platforms have become central to public discourse. Governments worldwide struggle to balance protecting free expression with curbing harmful content such as incivility, misinformation, criticism, hate speech, and incitement to violence (Gorwa, 2019). The corporate-led moderation model, historically favored in the United States and parts of Europe, has faced mounting criticism for its failure to manage these risks effectively (Schroeder, 2025). Such failures are particularly evident in the Global South, where unchecked hate speech on platforms has contributed to real-world harms, such as ethnic violence in Myanmar (Fink, 2018) and post-election riots in Indonesia (Lim, 2017). These cases underscore the immense difficulty of regulating powerful transnational platforms, especially in contexts where societal trust is fragile and the potential for offline violence is high.

Different political systems have responded to these challenges in different ways. Western democracies have pursued regulatory reforms grounded in their foundational commitment to freedom of expression (Mansell, 2012; Zittrain, 2008). Rather than imposing direct state control, they increasingly pressure private platforms to enhance transparency and accountability. China, by contrast, represents a distinct model of state-centric regulation that prioritizes direct intervention over corporate self-regulation. Instead of relying on platform-driven moderation, the Chinese state mandates de-anonymization and enhanced visibility as core governance techniques, transforming social media platforms from mere service providers into compliant agents of state authority (Lee & Liu, 2016; Roberts, 2018).

To understand the logic of mandatory IP location disclosure within this state-centric paradigm, we must trace China's digital governance evolution. Over the past two decades, state regulation has shifted from simple content control to complex behavioral governance (McKnight et al., 2023; Gallagher & Miller, 2021). This evolution can be understood through three distinct phases, each representing a deepening of regulatory reach from controlling information to disciplining the digital subject through visibility (Yeung, 2018; Zhu, 2024).

Information Filtering and Content Control

The first phase focused on restricting access to information and managing content. Epitomized by the Great Firewall in the early 2000s, this model systematically blocked foreign websites and enabled the deletion of politically sensitive domestic content (Liu, 2022). As Roberts (2018) characterizes, this strategy relied on selective content moderation. The primary objective was not to eliminate all criticism but to prevent collective action. King, Pan, and Roberts (2013) demonstrate that critical commentary was often tolerated unless it posed a risk of triggering coordinated protest. In this stage, governance was primarily defined by what could not be seen.

Real-Name Registration and Backend Traceability

The second phase marked a shift toward institutionalizing identity traceability. In 2012, China introduced the Real-Name Registration system, requiring users to verify their identities with national ID cards (Fu et al., 2013; Zhu, 2024). Under this dual framework, users remained anonymous to peers while their identities were fully visible to the government (Jiang, 2016). This phase moved beyond controlling content to regulating the subject, ensuring that every digital action could be traced back to a physical body. While enhancing state oversight, public-

facing anonymity was maintained (Lee & Liu, 2016).

Mandatory IP Disclosure and Disciplinary Visibility

The third phase, mandatory IP location disclosure, represents the externalization of regulation to the public sphere. Beginning on April 28, 2022, platforms like Weibo began displaying IP location tags on user profiles and comments at the provincial level for domestic users and country level for international users (Guo et al., 2023; Zhu, 2024). Unlike full IP addresses that can reveal precise geolocation, the displayed information is approximate but mandatory and automatic (Liu et al., 2024). Every post and comment now carries a geographic identifier visible to all users.

This public visibility crucially distinguishes the current phase from the backend monitoring of the Real-Name Registration era. By enforcing front-facing traceability, the policy makes user identity visible not just to authorities but to the entire digital public. While platforms frame this measure as ensuring authenticity (Weibo, 2022), it functions as a disciplinary mechanism of mandatory visibility. By stripping away the public anonymity preserved in the previous phase, the policy leverages social pressure from both regulatory oversight and peer observation to induce self-regulation (Li et al., 2025; Zhu, 2024).

3 | Regulatory Disclosure, Self-Censorship, and Online Negative Sentiment

The mandatory IP location disclosure represents a distinctive form of regulatory disclosure that targets individual users rather than corporations. Regulatory disclosure refers to the compulsory release of structured information to achieve specific policy objectives (Weil et al., 2006). While traditional frameworks focus on corporate accountability, requiring firms to

disclose information to reduce information asymmetry (Fung et al., 2007; Leuz & Wysocki, 2016), the Chinese model extends this mechanism to individual citizens. By shifting the disclosure burden from platforms to users, the IP location policy aims to reduce online anonymity and enhance traceability (Li et al., 2025; Zhu, 2024). In this framework, platforms function not as the objects of regulation but as delegated agents enforcing transparency requirements on behalf of the state (Gallagher & Miller, 2021). Recent research confirms that Chinese citizens are aware of and responsive to such user-targeted surveillance measures, though public support varies depending on the specific behaviors being monitored (Xu et al., 2025).

The critical question is how this macro-level regulation influences micro-level individual expression. We argue that regulatory disclosure operates by triggering or altering the psychological mechanism of self-censorship. Self-censorship is defined as the act of intentionally and voluntarily withholding information or opinions in the absence of formal obstacles (Penney, 2019). It is a strategic calculation where individuals edit their public expression to align with perceived norms or to avoid social and political costs (Büchi et al., 2022; Li et al., 2025).

When confronted with the mandatory disclosure of their location, users experience a psychological dilemma where the need to share information conflicts with the fear of consequences facilitated by increased transparency. We propose that regulatory disclosure does not have a uniform effect. Rather, it activates two competing psychological pathways. First, regulation may strengthen self-censorship. By increasing the perceived cost of expression, the policy may lead to a chilling effect that reduces negative online sentiment (Büchi et al., 2022;

Li et al., 2025). Second, regulation may lower the threshold for self-censorship by triggering resistance. If users perceive the policy as an illegitimate threat to their freedom that exceeds their perceived cost, it may trigger a reactance effect, leading them to deliberately express stronger negative sentiment in order to reclaim their autonomy (Brehm, 1966; Hsieh, 2025; Zhu, 2024).

Our study examines how this specific application of regulatory disclosure affects negative online sentiment and user expression patterns in China's social media environment. Here, negative online sentiment is defined as a proxy for criticism, discontent, misinformation, and incivility (Chen et al., 2025; Guo et al., 2023). It encompasses a range of user expressions, including overt complaints, skeptical or distrustful remarks, sarcastic or hostile comments, and the spread of misleading information (Lyu et al., 2020). As such, negative sentiment reflects not only dissatisfaction with specific policies or events but also broader social frustrations and grievances (Li et al., 2025). This makes it a critical indicator for understanding the dynamics of digital public opinion and the effectiveness of regulatory interventions. Moreover, negative sentiment serves as a primary channel through which citizens express their emotions on public issues, allowing for both individual venting (Wong & Liang, 2023) and collective mobilization (King et al., 2013). By capturing the intensity and prevalence of such expressions, researchers can assess how regulatory disclosure policies shape the climate of online discourse and the willingness of users to voice dissent or criticism, despite the intentional regulatory disclosure to contain it. Our study then also assesses the effectiveness of the policy in regulating or managing online content.

3.1 | The Chilling Effect vs. The Reactance Effect

The chilling effect

Through a first causal pathway, we argue that mandatory IP location disclosure could suppress negative online sentiment by triggering a chilling effect, a phenomenon where individuals voluntarily withhold negative expression due to the fear of potential consequences (Schauer, 1978; Penney, 2019). In the context of digital authoritarianism, regulatory measures have evolved beyond mere administrative transparency into an automated and continuous monitoring of user traces (Büchi et al., 2022). Social media platforms frequently act as proxies for state regulation, implementing disclosure mandates that extend the state's disciplinary power into the everyday digital lives of citizens (Cobbe, 2021; Gallagher & Miller, 2021).

By decreasing anonymity, policies like mandatory IP location disclosure fundamentally shape the structural conditions of public discourse, making user identities visible across the entire digital sphere (van der Nagel & Frith, 2015; Zhu, 2024). When a policy like mandatory IP location disclosure is implemented, it signals heightened regulation, prompting users to reassess the risks associated with their online activities (Li et al., 2025). This heightened awareness leads users to increase their perceived probability of negative outcomes, such as legal sanctions, social ostracism, or privacy intrusions (Li et al., 2025; Penney, 2019). For example, Büchi et al. (2022) reported that an increased sense of digital dataveillance, defined as the automated collection and analysis of digital traces by state and corporate actors in Switzerland, led to greater expectations of negative consequences, which reduced willingness to express negative opinions online. Similarly, Fu et al. (2013) found that after the introduction of Real-Name Registration in China, overall posting frequency remained stable, but political commentary and negative sentiment declined.

In the specific context of mandatory IP location disclosure, the policy signals to users that their digital footprints are not only monitored but also publicly traceable, thereby heightening awareness of potential risks. As a result, users may perceive a greater likelihood of punishment or exposure, and in turn, are more likely to engage in self-censorship and refrain from negative expression. Figure 1 below illustrates the causal mechanism underlying the chilling effect, our first proposed causal pathway. Based on these arguments, we propose the following hypothesis:

H1: Mandatory IP location disclosure can significantly reduce negative online sentiment.

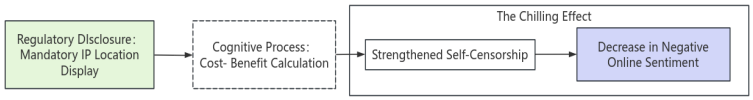

Figure 1 | Hypothetical Mechanisms of the Chilling Effect.

The reactance effect

Conversely, a potential second pathway suggests that mandatory IP location disclosure may inadvertently trigger a reactance effect, intensifying rather than suppressing negative sentiment. This phenomenon is grounded in the theory of psychological reactance (Brehm, 1966), which suggests that when individuals perceive a specific behavioral freedom is threatened or eliminated, they become motivationally aroused to restore it.

While reactance is traditionally associated with the defense of individual autonomy in Western contexts, recent comparative studies demonstrate its robust presence in collectivist societies such as Iran, South Korea, and China (Ng et al., 2021). In these collectivist cultures, regulatory overreach is often perceived not merely as a personal inconvenience but as a violation of the implicit social contract, triggering collective forms of resistance shaped by

prior direct and indirect experience with censorship (Zhu, 2024). In particular, when a policy is viewed as an illegitimate threat to autonomy or freedom rather than a legitimate transparency measure, it can trigger reactance or backfire effects (Zhu & Fu, 2021). In these contexts, enforced mandates have been shown to co-occur with a rise in negativity, anger, and freedom-related language on social media (Hsieh, 2025), as users seek to reclaim their agency through more frequent and emotionally intense dissent (Zhu, 2024).

This second and alternative causal pathway operates through a psychological mechanism of reactance with the goal of autonomy restoration (Brehm, 1966; Hsieh, 2025). Faced with this visible regulatory measure, the psychological cost of compliance outweighs the potential risks of punishment (Lu & Liang, 2024; Roberts et al., 2018). Instead of self-censoring, users may resist constant visibility and monitoring by intensifying negative sentiment, using criticism, sarcasm, or collective venting to reclaim their digital agency (Zhu, 2024; Zhu & Fu, 2021). For example, Roberts et al. (2018) found that online censorship often prompts users to seek more information and express greater dissatisfaction, while Zhu and Fu (2021) observed that post removals can escalate negativity around censored topics. Unlike the chilling effect, these findings indicate that regulatory interventions may inadvertently trigger psychological reactance and increase negative sentiment (Hsieh, 2025; Ng et al., 2021).

In this scenario, visible IP location labels may not deter users. Instead, they may become a focal point for reactance. Figure 2 below depicts the causal mechanism underlying the reactance effect. Based on these arguments, we propose our second hypothesis, which contrasts with H1:

H2: Mandatory IP location disclosure can significantly increase negative online

sentiment.

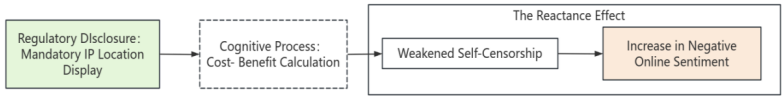

Figure 2 | Hypothetical Mechanisms of the Reactance Effect.

3.2 | Contextualizing the Effects: The Moderating Role of Account Identity and Interaction Area

Theoretical Integration

While the chilling effect and reactance effect offer competing predictions, we argue that they are not mutually exclusive but operate conditionally. As illustrated in Figure 3, we propose a theoretical integration where the social media context, specifically account identity and interaction area, serves as boundary conditions that determine which hypothesized psychological mechanism prevails. High-risk exposure contexts are hypothesized to trigger the chilling effect pathway, while low-risk exposure contexts are hypothesized to trigger the reactance effect pathway.

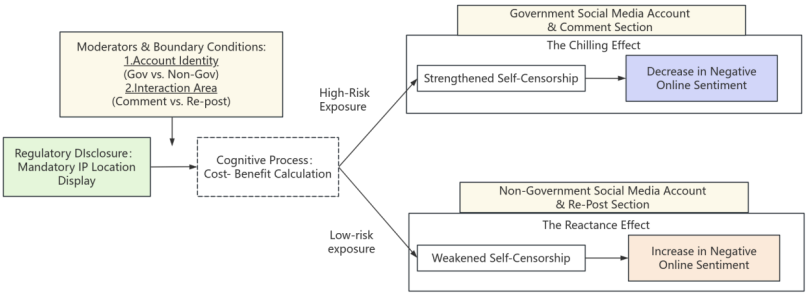

Figure 3 | Conceptual Framework: The Moderating Role of Account Identity and Interaction Area on Regulatory Disclosure Effects.

Government vs. Non-Government Accounts

We first examine account identity as a critical boundary condition. On social media, the distinction between government and non-government accounts delineates zones of differing regulatory intensity, visibility, and risk exposure (Atad et al., 2023; Zhang & Guo, 2021). Account identity serves as a primary signal that fundamentally alters the user's risk assessment and self-censorship calculation.

On Chinese social media platforms, this distinction is structurally reinforced by institutional markers such as the white "V" inside a red circle that certifies official government accounts (see Figure 4). These accounts serve as digital extensions of state authority and expose users to the highest level of scrutiny (Gallagher & Miller, 2021; Jiang, 2016). In this context of heightened risk exposure, the mandatory display of IP locations makes users feel directly monitored by state authorities (Li et al., 2025). As a result, the perceived risk of legal sanctions or social ostracism far outweighs the psychological drive for self-expression (Büchi et al., 2022). Confronted with the real threat of traceability in a government-controlled environment, users increase self-censorship to avoid potential consequences (Fu et al., 2013; Li et al., 2025). Therefore, we expect the chilling effect to be most pronounced in these spaces.

In contrast, non-government social media accounts exist in an environment with lower perceived regulatory oversight. While still subject to platform rules, these spaces present reduced perceived risks of immediate state intervention compared to official accounts (Wong & Liang, 2023; Zhang & Guo, 2021). In these non-official spaces, mandatory IP disclosure is often perceived not as a legitimate transparency measure, but as an intrusive form of digital surveillance that violates the principle of informed consent and spatial anonymity (Liu et al.,

2024; Zhu, 2024).

Because the perceived cost of expression is lower in these non-official spaces, the motive to restore autonomy outweighs the fear of regulatory disclosure (Zhu, 2024). Users are therefore more likely to utilize these spaces as outlets for frustration, engaging in discursive resistance against the policy (Zhu, 2024; Zhu & Fu, 2021). Based on these observations, we propose the following hypothesis:

H3: Mandatory IP location disclosure can decrease negative online sentiment under government social media accounts but increase it under non-government social media accounts.

Figure 4 | Examples of Government and Non-Government Social Media Accounts on Weibo.

Types of Accounts and Interaction Area Jointly Shape the Effect

Beyond account identity, we posit that interaction area works together with account type to shape how mandatory IP location disclosure affects negative online sentiment. By interaction area, we refer to whether a user posts in the comment section or the re-post section. The distinction between comment section and re-post section represents fundamentally different regimes of visibility and regulatory scrutiny (Wong & Liang, 2023).

The comment section functions as the primary, high-visibility display zone (Schroeder,

2025). Within this space, account holders have significant curatorial authority over the content shown. They can determine which comments appear most prominently, control the order of responses, and manage the overall visibility of discussions (Gorwa, 2019; Wong & Liang, 2023). Government accounts, in particular, actively use these moderation tools, selecting favorable comments for top placement, limiting the number of visible responses, and implementing comprehensive content filters (Gallagher & Miller, 2021; Wong & Liang, 2023). As a result, the high visibility and active moderation in this space heighten users' perceptions of regulatory and social risk, leading them to self-censor defensively and sharply curtail critical commentary (Li et al., 2025; Penney, 2019; Zhu, 2024).

In contrast, the re-post section serves as a structurally distinct alternative information space. Unlike comments displayed directly beneath posts, re-post section requires an additional navigational step where users must click on a separate tab to view them (Wong & Liang, 2023; Yao et al., 2019). This architectural difference makes these spaces perceived as lower risk (Zhu & Fu, 2021). Users strategically migrate to re-post section to share critical or non-mainstream viewpoints excluded from curated comment threads (Wong & Liang, 2023; Zhang & Guo, 2021). In these spaces, mandatory location disclosure is less likely to trigger fear and more likely to be interpreted as an unwarranted intrusion. Users are therefore more inclined to overcome self-censorship and amplify negative sentiment as a form of resistance (Guo et al., 2023; Zhu, 2024).

We argue that the interaction between account identity and interaction area creates distinct regulatory effects. In high-risk contexts such as government account comment section, the chilling effect predominates. In low-risk contexts such as non-government account re-post

section, reactance predominates as users resist perceived control attempts. Based on this reasoning, we propose the following hypothesis:

H4: Mandatory IP location disclosure can decrease negative sentiment in the comment section of government accounts while increasing it in the re-post section of non-government accounts.

4 | Research Design and Methodology

4.1 | Research Design

This study examines China's mandatory IP location disclosure policy as a context for investigating the impact of regulatory disclosure on negative online sentiment. Implemented as part of a broader initiative to combat harmful online content, this policy mandates social media platforms publicly display users' IP location information whenever they post or comment (Guo et al., 2023; Zhu, 2024).

We specifically focus on Weibo, China's largest social media platform, which had approximately 582 million monthly active users at the time of policy enactment (SinaTech, 2022). Since April 28, 2022, Weibo has displayed users' IP locations at the provincial level within China and at the country level for overseas users. We treat this policy implementation as an exogenous policy intervention and employ a quasi-natural experimental design to examine its causal effects on user behavior (Guo et al., 2023; Li et al., 2025).

Figure 5 illustrates the data collection process. We gathered data over a 28-day period divided into two phases: 14 days before the intervention (April 14–28) as Period 1 and 14 days after the intervention (April 29–May 12) as Period 2. Comments on posts in Period 1 and Period

2 served as the control and treatment groups, respectively.

To identify relevant discussions, we used Weibo's trending topics, which highlight real-time popular discussions similar to Twitter's trends, and focused on 1,694 pandemic-related topics from April 14 to May 12. We chose the pandemic as the focus of our analysis because COVID-19 remained a consistent and stable trending topic in China throughout the study period with no significant policy changes. This temporal stability ensures that the mandatory IP location disclosure policy represents the primary source of variation during our observation window, satisfying the exogeneity assumption crucial for quasi-natural experimental identification.

We employed a Selenium crawler to extract comments from the top posts for each topic. Post ranking was primarily determined by likes and views, ensuring selection of the most popular and influential posts. Comments on Weibo can be posted in two sections: the comment section, which appears directly beneath posts, and the re-post section, which requires users to click on a separate tab to view. We collected all comments across both sections on pandemic-related posts, totaling more than 200,000 comments. After removing blank comments, emojis, punctuation-only entries, duplicates, and comments outside the time range, 168,728 valid comments remained.

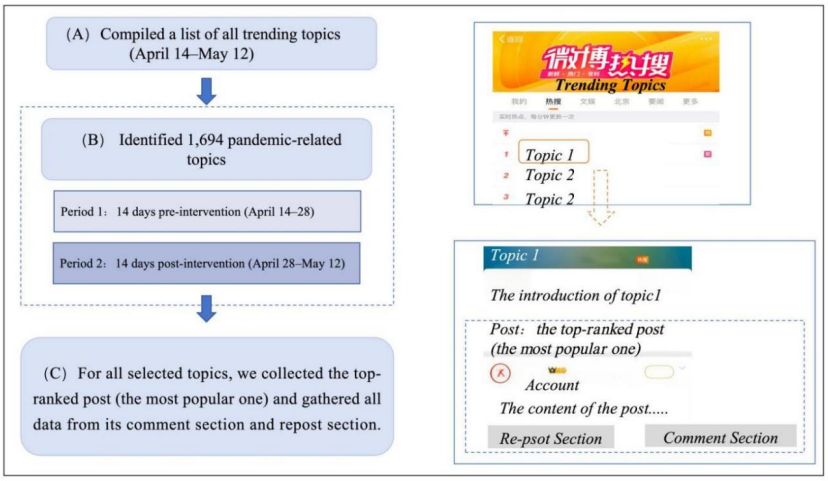

Figure 5 | Process of Data Collection.

We employed a Regression Discontinuity Design (RDD) to assess how mandatory IP location disclosure influences negative online sentiment expressed in users' comments on social media. The RDD is particularly suitable for evaluating policy interventions, as it compares outcomes immediately before and after a clearly defined cutoff date, enabling the estimation of local average treatment effects (Cattaneo et al., 2020).

Our analysis leverages Weibo's mandatory IP location disclosure policy implemented on April 28, 2022, as a sharp discontinuity in the data. Using a local polynomial approach (Cattaneo et al., 2020), we estimate the following regression model:

$$Y_{st} = f(X_{st}) + \beta T_{st} + \gamma Covariates_{st} + \pi_t + \epsilon_{st}.$$

where Y_{st} is the probability that the sentiment in citizens' comments at time t is negative, $f(X_{st})$ is a first-order polynomial of the running variable time X_{st}, and T_{st} indicates the treatment status of users at time t, determined by running variable X, which equals 1 for observations after April 28, 2022, and 0 otherwise. This study investigates the treatment effect

of mandatory IP location disclosure at the policy implementation point. The model adjusts for covariates and day fixed effects,π_t. This specification allows us to estimate the immediate treatment effect of mandatory location disclosure precisely at the policy implementation point, adjusting for potential confounders and temporal effects.

4.2 | Variables and Measurements

Dependent Variable: Negative Online Sentiment

Sentiment analysis is the dominant approach for analyzing sentiment through social media comments, and it is supported by robust technical tools and applications (Yue et al., 2019). Its primary objective is to determine whether a text expresses positive or negative sentiment (Lyu et al., 2020). Two main methods are commonly used: sentiment dictionaries, which rely on reference lexicons for classification, and machine learning models trained on domain-specific corpora, such as those utilizing Long Short-Term Memory (LSTM) networks (Liao et al., 2023; Yue et al., 2019).

Advances in deep learning, particularly with Chinese-language datasets, have further integrated these methods (Liao et al., 2023). Open-source platforms such as Baidu Paddle NLP have demonstrated high accuracy and scalability in large-scale sentiment analysis of Chinese social media content (Liao et al., 2023; Lyu et al., 2020). For example, Liao et al. (2023) used Baidu Paddle NLP to assess the impact of government microblog response strategies on public sentiment, confirming the model's high accuracy. Similarly, Lyu et al. (2020) analyzed shifts in public sentiment during the COVID-19 pandemic, comparing the performance of LSTM, BERT, and Baidu Paddle NLP, and found that Baidu Paddle NLP achieved the highest accuracy for sentiment analysis of Chinese texts on Weibo.

Building on these studies, we employed Baidu Paddle NLP for sentiment analysis. The model transforms each comment into a continuous probability score ranging from 0 to 1, where higher values indicate greater likelihood of negative sentiment. Baidu Paddle NLP is built on the ERNIE architecture, a pre-trained deep learning model designed for Chinese language understanding (Lyu et al., 2020). Unlike lexicon-based approaches that rely on keyword matching, ERNIE captures contextual semantics by learning from large-scale Chinese corpora including social media text. This enables the model to recognize sentiment expressed through indirect or figurative language, including sarcasm, humor, and coded wording common in Chinese online discourse, as it analyzes contextual relationships between words rather than treating them in isolation.

Table 1 provides illustrative examples of how the model scores different types of discourse. While incivility, misinformation, criticism, and discontent are distinct qualitative categories, the model integrates them into a single continuous metric of negative online sentiment. As shown in the table, comments containing uncivil language, aggressive criticism, or expressions of despair all receive high probability scores approaching 1.0. Table 1 thus demonstrates that the increase in negative sentiment observed in our results corresponds to a concrete rise in these specific forms of adversarial expression.

Table 1 | Sentiment Analysis Example.

Category	Comments	Negative Sentiment Score
Civil	Stay strong, Shanghai! Wishing everyone safety. （上海加油！一定要平安）	0.003
Incivility	Bullshit （放狗屁）	0.999
Misinformation	So many shootings and kidnappings	0.868

	（多少枪击案绑架案）	
Criticism	The efficiency is unbearably low. （是真的效率低 太不堪了）	0.999
Discontent	This world is just fucking stupid. (这个世界傻逼透了）	0.999

Running Variable: Time

The running variable, time, represents the specific timing of user comments, and was consistently used to assign treatment and control groups during the study period. As a key variable in the model, time was coded as a continuous measure, reflecting the temporal distance from the intervention date. Each comment's timestamp was recorded to the second and transformed into a range from -14.00 (April 14, 00:00) to 14.00 (May 12, 24:00), with hours and minutes converted into corresponding day fractions.

Covariates

Following best practices for Regression Discontinuity design studies, our model includes covariates determined prior to treatment assignment (Cattaneo et al., 2020). We controlled for two key factors known to influence sentiment expression in user comments (Li et al., 2025; Liao et al., 2023). Discussion heat measures the intensity of online engagement for each trending topic, operationalized as the total number of views reported by Weibo (Li et al., 2025). Sentiment of the original post quantifies the emotional tone of the initial post content associated with each comment thread (Liao et al., 2023). Using the deep learning framework provided by Baidu Paddle NLP, we converted the textual sentiment of original posts into a continuous variable ranging from 0 to 1, where higher values indicate greater negativity (Li et al., 2025; Liao et al., 2023; Lyu et al., 2020). Detailed measurement procedures and variable definitions are available in Appendix Table A1.

4.3 | Validation of the RDD

To ensure the robustness of our Regression Discontinuity Design (RDD) analysis, we conducted multiple validation tests as recommended by Cattaneo et al. (2020).

Bandwidth Selection and Power Analysis

The local polynomial approach in RD analysis fits polynomial regressions within a specified bandwidth around the cutoff date. The choice of bandwidth directly affects the number of observations used and the statistical power of the model. Following Cattaneo et al. (2020), we used a data-driven bandwidth selection method to minimize the Mean Squared Error of the local polynomial estimator. We conducted power analyses for each selected bandwidth to confirm sufficient statistical power. This approach reduces subjective decisions and specification searches, resulting in robust and statistically efficient estimation.

We primarily report results based on the common MSE-optimal bandwidth to maintain parsimony and stability. To address potential concerns about bandwidth sensitivity, we conducted robustness checks presented in Section 5.2 and the Appendix. These tests include comparing different polynomial orders and comparing common bandwidth versus distinct bandwidths on each side of the cutoff.

Covariate Balance Tests

Ensuring covariate balance is critical to validating the RDD assumption that treated and untreated observations near the cutoff differ solely by treatment status. We applied our RDD specification to test whether observable characteristics were balanced across the treatment threshold. Results reported in Appendix Table A2 confirmed covariate balance at the cutoff, providing confidence in the validity of subsequent estimates.

Sorting and Manipulation Check

A fundamental assumption of an RD design is that agents cannot precisely manipulate the running variable to sort themselves into treatment or control groups. In our study, the running variable is time. As highlighted in the literature on Regression Discontinuity in Time, time creates a naturally exogenous running variable because it flows continuously and cannot be paused or reversed by subjects (Hausman & Rapson, 2018).

While users could theoretically engage in strategic timing by rushing to post comments before policy enforcement, this is implausible for two reasons. First, the mandatory IP display was a platform-level update enforced instantaneously at the cutoff, leaving users no opportunity for manipulation. Second, we visually inspected the density of comment volume around the cutoff and observed no significant discontinuity (McCrary, 2008). This continuity supports the validity of our design, confirming that treatment assignment was locally randomized around the implementation threshold.

Placebo Cutoff Tests

Lastly, we implemented placebo cutoff tests to validate the discontinuity exclusively occurring at the true cutoff point. Discontinuities should be absent at placebo points to confirm the true causal effect. Appendix Table A3 demonstrates that no discontinuities appeared at placebo cutoff points, reinforcing the validity of our actual cutoff.

5 | Results

5.1 | RD Results

Table 2 reports the regression discontinuity results at the cutoff point. The coefficients represent the local average treatment effect of the mandatory IP location disclosure policy.

Standard errors were clustered at the individual level, and statistical inference employed robust bias-corrected intervals to account for potential misspecification errors (Cattaneo et al., 2020).

Table 2 | Regression Discontinuity Results.

	Coef.	Std. Err.	P>|z|	[95% Conf. Interval]
Total	-0.008	0.010	0.417	[-0.029, 0.012]
Account Identity				
Government	-0.028	0.009	0.004	[-0.046, -0.009]
Non-Government	0.065	0.022	0.003	[0.022, 0.108]
Account Identity * Interaction Area				
Government	-0.007	0.013	0.604	[-0.031, 0.018]
Comment Section				
Non-Government	0.279	0.047	0.000	[0.188, 0.370]
Re-post Section				

As shown in the first row of Table 2, no significant effect was observed in the full sample (coefficient = -0.008, p = 0.417). Rather than indicating policy ineffectiveness, this null effect suggests that chilling and reactance effects may operate simultaneously in opposite directions, offsetting each other. These results do not support either H1 or H2.

Significant and contrasting effects emerged when examining account identity. For government accounts, mandatory IP location disclosure led to a significant decrease in negative sentiment (coefficient = -0.028, p = 0.004). Figure 6 visualizes this 2.8% reduction in the probability of negative sentiment compared to pre-policy levels. Conversely, for non-government accounts, the policy significantly increased negative sentiment (coefficient = 0.065, p = 0.003). Figure 7 illustrates this 6.5% increase post-implementation. Robustness checks confirmed both findings across alternative bandwidths and functional forms (see Appendix Table A5), providing support for H3.

Further analysis by interaction area revealed additional insights. Results for government account comment section showed no significant separate effect. However, non-government account re-post section showed a pronounced increase in negative sentiment (coefficient = 0.279, $p < 0.001$). As illustrated by Figure 8, this represents a 27.9% increase in the probability of negative sentiment following implementation. This increase reflects an intensification of adversarial expressions including incivility, criticism, and discontent. The pattern suggests that re-post section functions as spaces where users, confronted with mandatory visibility, deploy critical language as psychological reactance. Robustness tests confirmed the stability of this finding (see Appendix Table A5), partially supporting H4.

Figure 9 visualizes the RDD point estimates and 95% confidence intervals across subgroups. Each dot represents the estimated local average treatment effect of mandatory IP location disclosure on negative sentiment at the policy cutoff for a given subgroup. The horizontal bars extending from each dot indicate 95% robust bias-corrected confidence intervals, which account for the approximation bias inherent in local polynomial estimation near the cutoff (Cattaneo et al., 2020). When a confidence interval does not cross the zero line (dashed red line), the estimated effect is statistically significant at the 95% confidence level. Estimates located to the left of the zero line indicate a reduction in negative sentiment (a Chilling Effect), while estimates to the right indicate an increase in negative sentiment (a Reactance Effect). These results highlight the importance of account identity and interaction area as key boundary conditions shaping the effects of regulatory disclosure.

Commented [1]: *Comment 3:*
Figure captions should explicitly indicate what the confidence intervals represent for readers less familiar with RDiT methodology.

We thank the reviewer for this helpful suggestion. Upon review, we confirmed that confidence intervals appear only in Figure 9, which presents the forest plot of heterogeneous treatment effects across subgroups. We have revised both the Figure 9 caption and the accompanying descriptive text in Section 5.1 to serve three functions for readers less familiar with RDiT methodology: (1) explaining that each point represents the estimated local average treatment effect at the policy cutoff, (2) specifying that the horizontal bars represent 95% robust bias-corrected confidence intervals that account for the approximation bias inherent in local polynomial estimation (Cattaneo et al., 2020), and (3) providing a practical reading guide indicating that the estimated effect is statistically significant at the 5% level when bars do not cross the zero line. Please see Section 5.1 (page 26-28) or the revised text below.

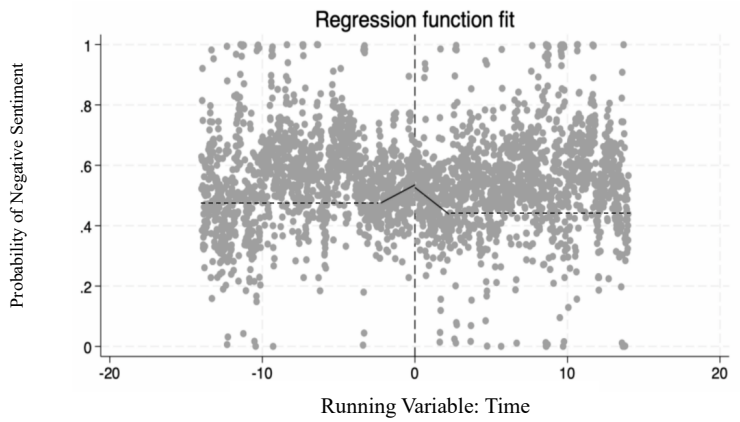

Figure 6 | Treatment Effect on Government Accounts.

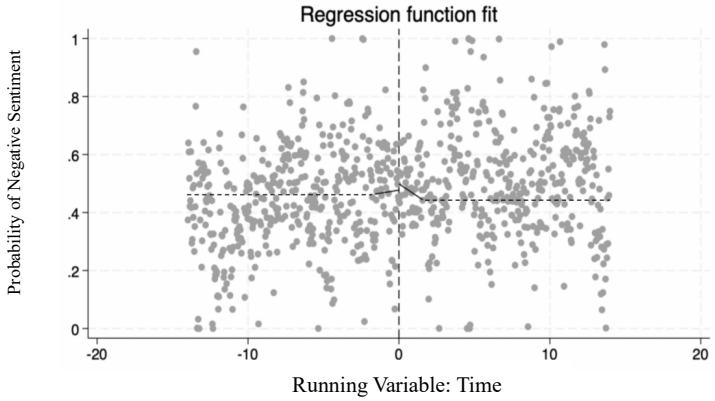

Figure 7 | Treatment Effect on Non-government Accounts.

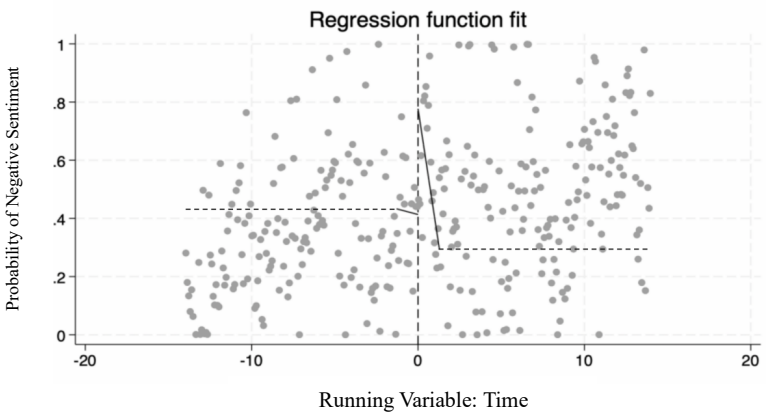

Figure 8 | Treatment Effect on Re-post Section of Non-government Accounts.

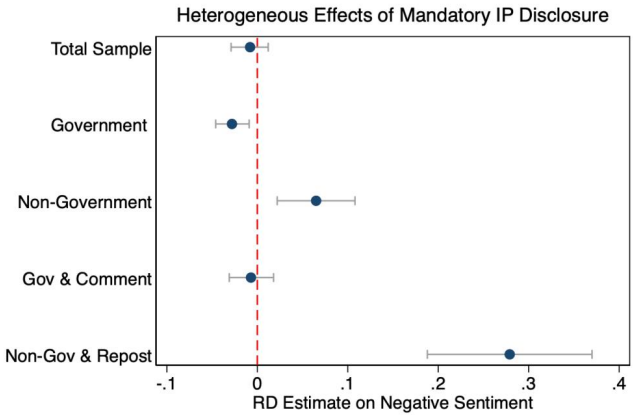

Figure 9 | Heterogeneous Effects of Mandatory IP Disclosure on Negative Sentiment.

Note: Each point represents the estimated local average treatment effect at the policy cutoff. Horizontal bars indicate 95% robust bias-corrected confidence intervals. Effects are statistically significant at the 5% level when bars do not cross the zero line (dashed red). Estimates to the left indicate a reduction in negative sentiment (Chilling Effect); estimates to the right indicate an increase (Reactance Effect). Please revise this note accordingly

5.2 | Robustness Checks

We conducted a series of robustness checks to ensure the reliability of our results. First, we

tested robustness by adjusting bandwidth and polynomial function specifications (see Table A5, Appendix). Second, we accounted for non-linear time trends by including a Time² term in the baseline model (see Model 3, Table A4, Appendix). Third, we applied the donut-hole approach to assess sensitivity to observations near the cutoff (Table A6, Appendix). This method removes observations close to the cutoff that may be susceptible to manipulation or bias, ensuring that estimates are not overly dependent on data at the threshold (Cattaneo et al., 2020). All robustness tests produced results consistent with our main findings.

To validate the sentiment measure, we manually coded a stratified random sample of 1,500 comments following established practices in computational text analysis (Grimmer & Stewart, 2013; Van Atteveldt et al., 2021). The sample was stratified by treatment status, account type, and machine score range to ensure coverage of key analytical conditions. We manually rated each comment on a five-point scale from clearly positive to clearly negative, mapped to corresponding quintiles of the machine probability scores. The validation results indicate strong agreement between human and machine classifications. For binary classification of negative sentiment, the model achieved 81.7% accuracy, 83.3% precision, 81.2% recall, and an F1 score of 82.2%. Agreement within one category of the human rating occurred in 80.1% of cases. These results confirm that Baidu Paddle NLP reliably captures negative sentiment in Chinese social media discourse (see Appendix Tables A7–A9 for detailed validation statistics).

A potential threat to validity in RDiT designs is the occurrence of simultaneous external shocks such as other major policy announcements or viral events coinciding with the treatment threshold. To address this, we systematically compiled and reviewed all trending search topics on Weibo around the policy implementation point. Our screening confirmed that no other

significant internet regulations or polarizing social events occurred at the cutoff. This reinforces our confidence that the discontinuity in sentiment is driven by the mandatory IP location disclosure rather than unrelated trends.

6 | Discussion and Conclusion

Regulatory disclosure has long been a key instrument for policymakers managing complex information environments (Fung et al., 2007; Mahoney, 1995; Weil et al., 2006). While existing frameworks typically focus on corporate accountability (Leuz & Wysocki, 2016), this study examines how digital governance is extending regulatory disclosure requirements to individual citizens (Roberts & Oosterom, 2025). By analyzing China's mandatory IP location disclosure policy (Li et al., 2025; Zhu, 2024), we offer a theoretical reappraisal of how government-implemented user-targeted regulatory disclosure on social media platforms influences online expression.

Our findings advance the literature on regulatory disclosure by moving beyond average treatment effects to theorize and empirically demonstrate context-dependent heterogeneity. Prior research has largely examined whether transparency-based regulation suppresses or constrains online speech in general (Büchi et al., 2022; Li et al., 2025; Penney, 2019). Our analysis shows that the effects of regulatory disclosure are not uniform but are fundamentally shaped by the power dynamics embedded in specific digital contexts. By modeling account type and interaction area as moderating factors, we reveal a structural division in online political communication between official communication spaces and peer communication spaces (Atad et al., 2023; Wong & Liang, 2023; Zhang & Guo, 2021).

In government accounts, institutional trust is pivotal for effective government-citizen interaction (Zhang & Lu, 2025); however, mandatory disclosure disrupts this dynamic. By reclassifying user geolocation data as a public visible resource (Napoli, 2019), mandatory IP location disclosure transforms these spaces from potential sites of dialogue into nodes of regulation (Jiang, 2016; Li et al., 2025). Users read the location tag as a cue of monitoring and asymmetrical power (Büchi et al., 2022; Liu et al., 2024), which in turn encourages strategic avoidance, self-censorship, and performative compliance rather than substantive engagement (Dimitrov, 2015; Li et al., 2025; Liu et al., 2024).

In contrast, peer-oriented spaces including non-government accounts and less visible areas such as re-post section show a reactance effect where higher levels of negative sentiment are observed (Guo et al., 2023; Zhu, 2024). Psychological reactance is expressed collectively as users migrate negative sentiment into these spaces, where it functions as a signal of shared identity and group cohesion under conditions of heightened visibility (Wong & Liang, 2023; Zhang & Guo, 2021; Zhu & Fu, 2021).

Taken together, these findings offer a theoretical reframing of transparency-based governance in digital environments. Mandatory disclosure does not simply regulate speech. Instead, it reconfigures the digital public sphere into a controlled zone of official communication and a contested zone of peer interaction (Schroeder, 2025; Wong & Liang, 2023). This helps explain why transparency policies may simultaneously suppress expression in some spaces while intensifying it in others (Brehm, 1966; Zhu, 2024). More broadly, our study highlights that the consequences of digital regulation cannot be understood without attention to contextual power relations (Cheung & Chen, 2022; Cobbe, 2021), interactional

architecture (Yeung, 2018), and user perceptions of authority (Jiang, 2016; Liu et al., 2024).

Our findings carry significant practical and global implications. The logic of mandatory visibility reflects a broader trend in digital governance, in which states and platforms increasingly leverage user data disclosure to enforce order online (Lee & Liu, 2016). Similar dynamics are emerging in Western democracies (Gorwa, 2019; Schroeder, 2025), exemplified by X's rollout of country-based account labels. While such measures are intended to suppress disruptive content, our findings demonstrate that they may instead displace negative sentiment into less monitored spaces (Wong & Liang, 2023; Zhu & Fu, 2021), creating a filtered feedback loop where regulators receive biased information that misrepresents the popular mood (Dimitrov, 2015; Dragu & Lupu, 2021).

For policymakers and platform designers worldwide, our findings highlight that removing anonymity without addressing underlying grievances may displace dissent rather than resolve it (Büchi et al., 2022; Guo et al., 2023; Zhu, 2024). Effective digital governance requires balancing traceability with the preservation of spaces for autonomous expression (Cheung & Chen, 2022; Zittrain, 2008).

Our study outlines four directions for future research. First, as our analysis relied on pandemic-themed posts, the generalizability of our findings to other discourse topics warrants further verification. However, since mandatory IP location disclosure fundamentally alters the architecture of online visibility regardless of topic, we expect the mechanisms underlying the chilling effect in official spaces and the reactance effect in peer spaces to persist across other domains of public discourse. Although the magnitude of these effects may vary with the political sensitivity of specific topics, the underlying user response to reduced anonymity is

Commented [2]: *Comment1:*
Some paragraphs in the discussion could be streamlined for clarity, as multiple ideas are densely presented.

We thank the reviewer for this helpful suggestion. We revised Section 6 (Discussion and Conclusion) to improve clarity by reducing within-paragraph density and sharpening paragraph-level focus. Specifically, we disentangled multi-step causal claims in the government accounts paragraph by breaking long, concept-heavy sentences into shorter units and separating the trust–visibility mechanism from its behavioral implications.
We also streamlined the peer spaces paragraph to foreground the core reactance mechanism while removing ancillary framing that was previously bundled into the same paragraph. In addition, we tightened the theoretical reframing paragraph by eliminating redundant restatements and improving transitions across paragraphs. Finally, we refined the practical implications paragraph by linking the displacement finding more directly to its governance consequence, and we separated the policy recommendations into a distinct paragraph for easier navigation. Taken together, these revisions present one primary idea per paragraph and reduce unnecessary repetition, while preserving all substantive arguments and theoretical contributions. Please see pages 30–32 (or the revised text below) for these changes.

likely to be similar. Second, we focused on the policy's formal implementation, excluding the agenda-setting phase due to its low public visibility. Third, while our RDD analysis captures immediate policy shocks, it does not account for long-term user adaptation. Future longitudinal studies are needed to determine if these effects persist as the policy normalizes. Finally, by prioritizing regulator-user dynamics, we may have overlooked horizontal pressures. Future research should explore how IP-based regional stereotypes and social stigma drive self-censorship independent of state regulation.

Commented [3]: *Comment 2:*
The conclusion could briefly address the potential generalizability of the findings to non-pandemic topics.

We appreciate this insightful suggestion regarding generalizability. In the last paragraph of Section 6 (Conclusion), we added a brief discussion of how our mechanisms may extend beyond pandemic-related discourse. We clarify that mandatory IP location disclosure alters the architecture of online visibility regardless of topic, suggesting that the chilling effect in official spaces and the reactance effect in peer spaces may also arise in other topics. We further note that the magnitude of these effects is likely to vary with topic sensitivity and the intensity of regulatory attention. Please see Section 6 (page 32) or the revised text below.

Appendix

Table A1 | Variables and measurements

Type	Variable	Measurement
Dependent Variables	Negative Online Sentiment	The probability of a comment being negative, measured as a continuous variable ranging from 0 to 1 where higher values indicate greater negativity. Sentiment was analyzed using the Baidu Paddle NLP model, a widely recognized tool for Chinese social media text.
Running Variable	Time	Temporal distance from the intervention start, recorded to the second. Values range from -14.00 (April 14, 00:00) to 14.00 (May 12, 24:00), with hours and minutes converted to day fractions.
Covariates	Discussion Heat	Total view count for each trending topic as reported by Weibo, collected via Selenium crawler from April 14 to May 12, 2022.
	Original post sentiment	Sentiment of each post analyzed using Baidu Paddle NLP, converted to a continuous variable ranging from 0 to 1 where higher values indicate greater negativity.

Table A2 | Balance tests of the covariates

Variable	MSE-Optimal Bandwidth	RD Estimator	Robust Inference		Num. Observations
			p-value	Conf.Int.	
Discussion heat of each trending topic	1.267	-4338.3	0.670	[-24287.8, 15611.2]	168728
Original post sentiment	0.616	0.002	0.822	[-0.015, 0.019]	168728

Table A3 | Placebo Test

Placebo Test in Government Social Media Account

Alternative Cutoff	RD Estimator	Robust Inference p-value	Robust Inference Conf.Int.
-1	-0.023	0.428	[-0.081, 0.035]
0	- 0.028	0.004	[-0.046, -0.009]
1	0.014	0.662	[−0.049,0.078]

Placebo Test in Non-government Social Media Account

Alternative Cutoff	RD Estimator	Robust Inference p-value	Robust Inference Conf.Int.
-1	0.076	0.146	[-0.026, 0.179]
0	0.065	0.003	[0.022, 0.108]
1	0.030	0.223	[−0.018,0.078]

Placebo Test in the Re-post Section & Non-Government Social Media Account

Alternative Cutoff	RD Estimator	Robust Inference p-value	Robust Inference Conf.Int.
-1	0.095	0.196	[-0.049, 0.239]
0	0.279	0.000	[0.188, 0.370]
1	-0.050	0.204	[−0.128,0.027]

Table A4 | Multiple regression analysis

	(1)	(2)	(3)
	Model 1	Model 2	Model 3
Treatment		-0.035***	-0.035***
		(0.000)	(0.000)
Account Identity	-0.073***	-0.073***	-0.073***
(Non-government=Yes)	(0.000)	(0.000)	(0.000)
Interaction Area	-0.019***	-0.018***	-0.018***
(Re-post=Yes)	(0.000)	(0.000)	(0.000)
Original Post Sentiment	-0.006*	-0.006	-0.006*
	(0.096)	(0.112)	(0.078)
Discussion heat of each trending topic	$9.48e^{-10}$	$1.30e^{-9}$	$1.07e^{-9}$
	(0.451)	(0.302)	(0.394)
T(Time)	0.001***	0.003***	0.003***
	(0.000)	(0.000)	(0.000)
T^2($Time^2$)			0.001 ***
			(0.001)
N	168728	168728	168728
adj. R2	0.007	0.007	0.007

Table A5 | RD regression results with alternative bandwidths

		Mse-optimal bandwidth (mserd)	Mse-optimal bandwidth (msetwo)	Quadratic polynomial
Total	RD estimator	- 0.008	- 0.015	-0.002
		[- 0.029, 0.012]	[- 0.039, 0.008]	[- 0.026, 0.022]
	Effective N	28918	20374	38940
Account Identity Government	RD estimator	-0.028**	-0.046***	-.0369**
		[-0.046, -0.009]	[-0.072, -0.021]	[-0.064, -0.011]
	Effective N	31454	15203	29371
Non-Government	RD estimator	0.065**	0.049**	0.057**
		[0.022, 0.108]	[0.008, 0.089]	[0.014, 0.099]
	Effective N	7772	9010	14162
Interaction Area Government* Comment Section	RD estimator	-0.007	0.001	-0.005
		[-0.031, 0.018]	[-0.030,0.031]	[-0.035,0.025]
	Effective N	18793	12237	24866
Non-Government* Re-post Section	RD estimator	0.279***	0.250***	0.291***
		[0.188, 0.370]	[0.166, 0.335]	[0.199, 0.383]
	Effective N	1744	2463	3518

Table A6 | RD regression results for the donut-hole approach

	Donut-Hole Radius	Msetwo-optimal bandwidth	RD Estimator	Robust Inference	
				p-value	Conf. Int.
Total	0.00	1.422	-0.008	0.417	[- 0.029, 0.012]
	0.10	1.224	-0.022	0.092	[-0.047, 0.004]
Government	0.00	2.323	-0.028	0.000	[-0.046, -0.009]
	0.10	1.272	-0.063	0.000	[-0.091, -0.034]
Non-Government	0.00	1.586	0.065	0.003	[0.022, 0.108]
	0.10	1.414	0.081	0.003	[0.027, 0.135]
Government* Comment Section	0.00	1.763	-0.007	0.604	[-0.031, 0.018]
	0.10	1.710	-0.015	0.324	[-0.045, 0.015]
Non-Government* Re-post Section	0.00	1.287	0.279	0.000	[0.188, 0.370]
	0.10	0.981	0.229	0.000	[0.112, 0.346]

Table A7 | Manual Validation of Sentiment Measure

Item	Description
Sample size	1,500 comments
Sampling method	Stratified random sampling
Stratification variables	Treatment status, account type, sentiment range
Human coding scale	5-point ordinal (1 = clearly positive, 5 = clearly negative)
Machine score mapping	Quintiles (0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, 0.8-1.0)

Table A8 | Validation Performance Metrics

Metric	Value
Accuracy	81.7%
Precision	83.3%
Recall	81.2%
F1 Score	82.2%
Agreement within one category	80.1%

Note. Binary classification uses threshold = 0.5. Agreement within one category refers to cases where machine and human ratings differ by no more than one level on the five-point scale.

Table A9 | Confusion Matrix for Binary Classification

	Machine Non-negative	Machine Negative
Human Non-negative	591	128
Human Negative	147	634

Note. N = 1,500. Non-negative includes human ratings 1-3 (clearly positive to neutral). Negative includes human ratings 4-5 (slightly negative to clearly negative). Machine threshold = 0.5.

References

Atad, E., Lev-On, A., & Yavetz, G. (2023). Diplomacy under fire: Engagement with governmental versus non-governmental messages on social media during armed conflicts. Government Information Quarterly, 40(3), 101835.

Brehm, J. W. (1966). A theory of psychological reactance. Academic Press.

Büchi, M., Festic, N., & Latzer, M. (2022). The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. Big Data & Society, 9(1), 1–14.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020). A practical introduction to regression discontinuity designs. Cambridge University Press.

Chen, X., Guan, T., & Yang, Y. (2025). Allocating content governance responsibility in China: Heterogeneous public attitudes toward multistakeholder involvement strategies. Policy & Internet, 17, e432.

Chen, X., Zheng, P., & Mou, J. (2025). Understanding Chinese Internet users' information sensitivity in big data and artificial intelligence era. Policy & Internet, 17, e419.

Cheung, M., & Chen, Z. T. (2022). Power, freedom, and privacy on a discipline-and-control Facebook, and the implications for internet governance. IEEE Transactions on Professional Communication, 65(4), 467–484.

Cobbe, J. (2021). Algorithmic censorship by social platforms: Power and resistance. Philosophy & Technology, 34, 739–766.

Creemers, R. (2018). China's Social Credit System: An evolving practice of control. SSRN.

Cukier, K., & Mayer-Schoenberger, V. (2013). The Rise of Big Data: How It's Changing the Way We Think About the World. Foreign Affairs, 92(3), 28–40.

Dimitrov, M. K. (2015). Internal government assessments of the quality of governance in China. Studies in Comparative International Development, 50(1), 50–72.

Dragu, T., & Lupu, Y. (2021). Digital authoritarianism and the future of human rights. International Organization, 75(4), 991-1017.

Fink, C. (2018). Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. Journal of International Affairs, 71(1.5), 43-52.

Fu, K.-w., Chan, C.-h., & Chau, M. (2013). Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy. IEEE Internet Computing, 17(3), 42–50.

Fung, A., Graham, M., & Weil, D. (2007). Full disclosure: The perils and promise of transparency.

Cambridge University Press.

Gallagher, M., & Miller, B. (2021). Who not what: The logic of China's information control strategy. The China Quarterly, 248(1), 1011–1036.

Gorwa, R. (2019). What is platform governance? Information, Communication & Society, 22(6), 854-871.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3), 267–297.

Guo, Y., Li, Y., & Yang, T. (2023). Civilizing social media: The effect of geolocation on the incivility of news comments. New Media & Society.

Hausman, C., & Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. Annual Review of Resource Economics, 10, 533–552.

Hsieh, P. H. (2025). Psychological reactance to vaccine mandates on Twitter: A study of sentiments in the United States. Journal of Public Health Policy, 46, 269–283.

Jiang, M. (2016). Managing the micro-self: The governmentality of real name registration policy in the Chinese micro blogosphere. Information, Communication & Society, 19(2), 203–220.

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.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481.

Lee, J.-A., & Liu, C.-Y. (2016). Real-name registration rules and the fading digital anonymity in China. Washington International Law Journal, 25(1), 1–34.

Leuz, C., & Wysocki, P. D. (2016). The economics of disclosure and financial reporting regulation: Evidence and suggestions for future research. Journal of Accounting Research, 54(3), 525–622.

Li, J., Luo, Y., & Yuan, Q. (2025). Uncovering the impact of IP location display on user behavior in China's social platforms: A policy-driven analysis. Telecommunications Policy, 102978.

Liao, W., Li, L., & Lu, X. (2023). Research on the influencing factors of public comments' emotion on government microblogs. Journal of Education, Humanities, and Social Sciences, 18, 165–178.

Liu, Y. (2022). Internet governance in China: Toward a new cyber civilization. China Quarterly of International Strategic Studies, 8(3–4), 359–377.

Lim, M. (2017). Freedom to hate: Social media, algorithmic enclaves, and the rise of tribal nationalism in Indonesia. Critical Asian Studies, 49(3), 411-427.

Liu, Y. L., Wu, Y., Li, C., Song, C., & Hsu, W. Y. (2024). Does displaying one's IP location influence users' privacy behavior on social media? Evidence from China's Weibo. Telecommunications Policy, 48(5), 102759.

Lu, S., & Liang, H. (2024). Reactance to uncivil disagreement?: The integral effects of disagreement, incivility, and social endorsement. Journal of Media Psychology, 36(1), 15–26.

Lyu, X., Chen, Z., Wu, D., & Wang, W. (2020). Sentiment analysis on Chinese Weibo regarding COVID-19. In X. Zhu, M. Zhang, Y. Hong, & R. He (Eds.), Natural language processing and Chinese computing: NLPCC 2020 (pp. 710-721). Springer.

Mahoney, P. G. (1995). Mandatory disclosure as a solution to agency problems. The University of Chicago Law Review, 62(3), 1047–1112.

Mansell, R. (2012). Imagining the Internet: Communication, innovation, and governance. Oxford University Press.

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

McKnight, S., Kenney, M., & Breznitz, D. (2023). Regulating the platform giants: Building and governing China's online economy. Policy & Internet, 15(2), 243–265.

Moss, G. (2025). Digital regulation and questions of legitimacy. Policy & Internet, 17, e433.

Napoli, P. M. (2019). User data as public resource: Implications for social media regulation. Policy & Internet, 11(4), 439–459.

Ng, A. H., Kermani, M. S., & Lalonde, R. N. (2021). Cultural differences in psychological reactance: Responding to social media censorship. Current Psychology, 40, 2804–2813.

Penney, J. (2019). Chilling effects and transatlantic privacy. European Law Journal, 25, 122–139.

Roberts, M. E. (2018). Censored: Distraction and diversion inside China's Great Firewall. Princeton University Press.

Roberts, T., & Oosterom, M. (2025). Digital authoritarianism: A systematic literature review. Information Technology for Development, 31(4), 860-884.

Schauer, F. (1978). Fear, risk and the First Amendment: Unraveling the chilling effect. Boston University Law Review, 58, 685–732.

Schroeder, R. (2025). Content moderation and the digital transformations of gatekeeping. Policy & Internet, 17, e425.

SinaTech. (2022). The adjusted operating profit margin of Weibo Q2 with a revenue of 3 billion yuan

was 32%, a quarter-on-quarter increase. Sina Finance. https://finance.sina.com.cn/tech/internet/2022-09-01/doc-imizmscv8702747.shtml?cref=cj

TechCrunch. (2025). X rolls out 'About this account' feature that displays a profile's country of origin. Engadget. https://www.engadget.com/social-media/x-rolls-out-about-this-account-feature-that-displays-a-profiles-country-of-origin-and-more-160617187.html

Van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. Communication Methods and Measures, 15(2), 121–140.

van der Nagel, E., & Frith, J. (2015). Anonymity, pseudonymity, and the agency of online identity: Examining the social practices of everyday anonymity. First Monday, 20(3).

van Dijck, J., Poell, T., & De Waal, M. (2018). The platform society: Public values in a connective world. Oxford University Press.

Weibo. (2022). IP location feature upgrade announcement. https://weibo.com/1934183965/LqvYeCdBu

Weil, D., Fung, A., Graham, M., & Fagotto, E. (2006). The effectiveness of regulatory disclosure policies. Journal of Policy Analysis and Management, 25, 155–181.

Wong, S. H. W., & Liang, J. (2023). Attraction or distraction? Impacts of pro-regime social media comments on Chinese netizens. Political Behavior, 45(4), 1071–1095.

Xu, P., Krueger, B., Liang, F., Zhang, M., Hutchison, M., & Chang, M. (2025). Media framing and public support for China's social credit system: An experimental study. New Media & Society, 27(2), 995–1013.

Yao, W., Jiao, P., Wang, W., & Sun, Y. (2019). Understanding human reposting patterns on Sina Weibo from a global perspective. Physica A: Statistical Mechanics and its Applications, 518, 374-383.

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. Regulation & Governance, 12(4), 505–523.

Yue, L., Chen, W., Li, X., & Zuo, W. (2019). A survey of sentiment analysis in social media. Knowledge and Information Systems, 60(2), 617–663.

Zhang, X., & Lu, F. (2025). Enhancing public health policy communication through government–citizen social media interactions: The impact of replying agents, inquiry tone, and institutional trust. Policy & Internet, 17, e70000.

Zhang, Y., & Guo, L. (2021). 'A battlefield for public opinion struggle': How does news consumption

from different sources on social media influence government satisfaction in China? Information, Communication & Society, 24(4), 594–610.

Zhu, Y. (2024). Privacy cynicism and diminishing utility of state surveillance: A natural experiment of mandatory location disclosure on China's Weibo. Big Data & Society, 11(2).

Zhu, Y., & Fu, K. (2021). Speaking up or staying silent? Examining the influences of censorship and behavioral contagion on opinion (non-)expression in China. New Media & Society, 23(12), 3634–3655.

Zittrain, J. (2008). The future of the Internet—And how to stop it. Yale University Press.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.