Stanford | Center on China's Economy and Institutions
Freeman Spogli Institute and Stanford Institute for Economic Policy Research

SCCEI CHINA BRIEFS
Updated February 15, 2026

User Location Disclosure Fails to Deter Overseas Criticism but Amplifies Regional Divisions on China's Social Media

Leo Yang Yang and Yiqing Xu (2025). **User Location Disclosure Fails to Deter Overseas Criticism but Amplifies Regional Divisions on Chinese Social Media**. Cornell University *Arxiv* working paper.

China's social media is tightly censored — on Weibo, China's largest microblogging site, 3% of sensitive posts are deleted within 30 minutes and 90% within 24 hours. Yet in 2021, authorities proposed a new control mechanism: requiring social media platforms to display users' IP-based locations — by province for domestic users and by country for overseas users. Officials framed the policy as necessary to combat misinformation and foreign interference. On March 17, 2022, less than a month after the Russian invasion of Ukraine, Weibo began tagging the locations of users posting about the war. On April 28, Weibo extended location tagging to all posts and comments. Other platforms quickly followed. By year's end, location disclosure had become standard across Chinese social media. How has the location disclosure policy affected online discourse?

The data. Researchers continuously monitored 165 prominent government and media Weibo accounts at five-minute intervals from April 18 to May 9, 2022, capturing posts and their top 20 comments before censorship or deletion. This high-frequency approach preserved authentic user responses that would otherwise disappear from standard datasets. A rare implementation glitch briefly displayed location tags on historical comments, enabling recovery of pre-policy geographic data. The researchers classified posts as international (mentioning foreign locations) or domestic, and further categorized domestic posts as local (mentioning particular provinces) or non-local. They also identified comment sentiment and whether replies contained regionally discriminatory language (comments that stereotype, insult, or attack people based on their provincial origin) using supervised machine learning and large language models validated against human coding.

INSIGHTS

■ China's 2022 policy requiring IP-based location display on social media — promoted as a tool to curb misinformation and foreign interference — failed to deter comments from overseas users.

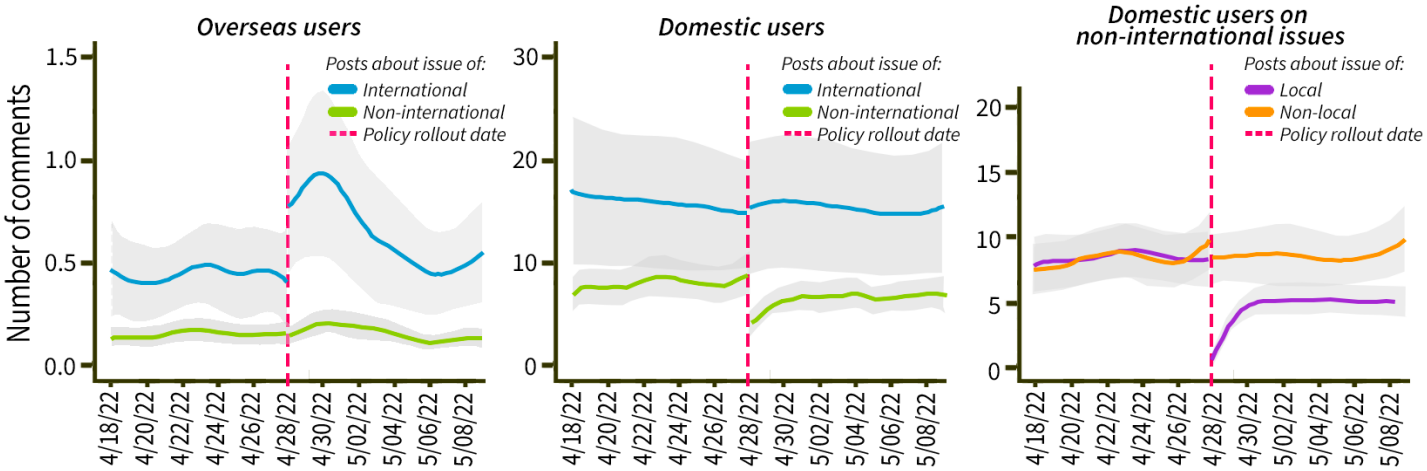
■ However, the policy caused domestic users to sharply curtail engagement with local issues outside their home provinces. Out-of-province comments dropped from 6.10 to 0.42 per post (a 93% decline) and remained depressed.

• • •

Overseas users did not retreat. Contrary to the policy's stated goal, overseas participation did not decline. Comments from overseas users on international topics rose sharply from 0.41 to 0.78 per post immediately after implementation, suggesting a reactive backlash rather than deterrence, before gradually returning to baseline. Engagement on non-international topics remained stable throughout. The result undermines the official rationale that visible location tags would reduce foreign "interference" by exposing overseas commenters to credibility challenges or attacks.

Domestic engagement on local issues collapsed. While the policy failed to deter overseas users, it sharply curtailed domestic discussion of local affairs. Comments from domestic users on non-international posts fell by roughly 50% immediately after the rollout. Disaggregating further, the decline was concentrated entirely in posts about local issues, i.e., those mentioning a Chinese province.

Social media engagement before and after user location disclosure



■ The decline was driven by peer pressure, not direct state repression. Regionally discriminatory replies surged after the policy, concentrated in cross-provincial interactions, raising the social cost of commenting on issues in other regions.

■ The policy reshaped online discourse by activating existing social divisions in ways that reinforce state control without direct censorship.

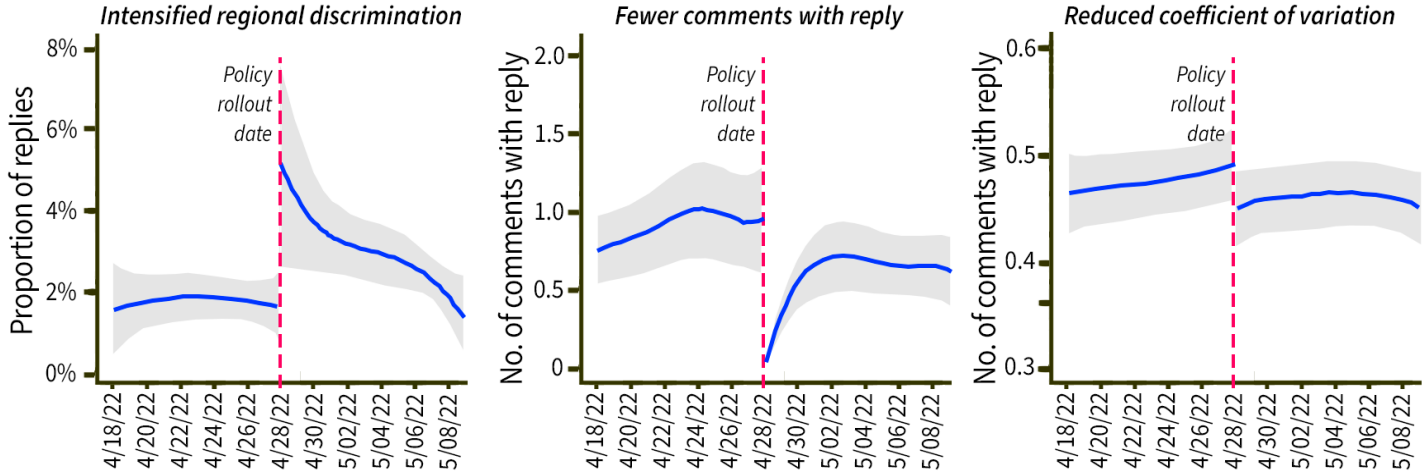
Engagement with local topics plummeted from 8.21 to 0.42 comments per post (a 95% decline), while participation on non-local topics showed no comparable decline. The distribution of posts across positive, neutral, and negative events did not change, ruling out shifts in content as an explanation.

Out-of-province users withdrew. Breaking down comments by user origin reveals that the steepest decline came from out-of-province users — those commenting on posts about regions where they do not reside. Their comments per post fell from 6.10 to 0.42 (a 93% decline) and stayed depressed in the days that followed. In-province users also reduced participation, but the drop was smaller and engagement rebounded quickly. The share of critical or dissenting comments on local posts also declined significantly, driven largely by the withdrawal of out-of-province users who had previously contributed disproportionately to criticism.

Silenced by peers, not police. If fear of state punishment were the primary driver, users might be most cautious when commenting on their own provinces, where risks of identification and retaliation are highest. Instead, the opposite occurred: out-of-province users withdrew while in-province users continued to participate. Domestic users also did not reduce engagement on international or national issues, including politically sensitive topics. The pattern points instead to fear of backlash from other users. Once provincial origins became visible, even neutral observations could be read as attacks from outsiders, and comment sections became sites of interregional confrontation.

Regional discrimination surged. Applying a large language model to detect regionally discriminatory language in replies, researchers found a sharp increase in discriminatory responses after the policy took effect, driven almost entirely by cross-provincial interactions. Within-province exchanges remained stable. Users commenting on other provinces faced greater risks of hostile replies, not just for critical remarks but also for simply engaging. Even neutral comments attracted attacks invoking regional stereotypes. One exchange: a Fujian user commenting on population decline in the northeast invoked stereotypes about "mafia" in Heilongjiang, Liaoning, and Jilin, prompting retaliatory replies about fraud and scams in Fujian. Location tags transformed casual remarks into interprovincial confrontations.

Comment section interactions before and after user location disclosure



Discourse narrowed and fragmented. The rise in regional antagonism reshaped the structure of online discussion. After disclosure, a smaller share of comments attracted replies, signaling more cautious engagement. The coefficient of variation in comment floor numbers — a measure of dynamism in the comment section — also dropped. With users more hesitant to comment across provincial boundaries, top comments were displaced less often, producing a more static and less interactive hierarchy. By making regional identity salient, the policy elevated the risks of cross-provincial participation, curtailed dissent, and produced a narrower, more fragmented conversation.

Divide and conquer. The findings show how authoritarian regimes can shape discourse without direct censorship. By embedding identity disclosure into platform design, China's authorities enabled peer-based sanctions that suppressed participation and narrowed expression. While perhaps not the explicit goal, the policy succeeded in fragmenting domestic discourse and reducing criticism, though it failed to deter overseas users. While such mechanisms may stabilize online discourse in the short run by muting dissent, they risk silencing early signals of discontent with potential costs for governance quality.