Journal of Quantitative Description: Digital Media 5(2025), 1–47 10.51685/jqd.2025.013

Unveiling Strategic Governance and User Dynamics in Weibo's Community-driven Content Moderation System

Andy Zhao

Cornell University, Ithaca, NY, USA

Lily Hu

The University of Hong Kong, Hong Kong, China

Social media companies continuously experiment with various platform governance models to tackle content moderation challenges, which calls for a comprehensive and empirical understanding of how a content moderation system evolves over the long term. Our study aims to fill this gap through a quantitative and qualitative study of Weibo's community-driven content moderation system, leveraging eleven million public moderation cases and decision data from 2012 to 2021. Based on reporting activities, platform decisions, and jury actions, we investigated the motivations and behavior patterns of three important actors in this governance model: reporting users, the platform, and user jurors. We suggest that users who frequently reported content and initiated the community-driven content moderation process tend to exhibit patterns of voluntarily policing the community or abusing others, sometimes coordinately, and were also treated differently by the platform. We indicate that Weibo's strategic moderation decisions have significantly distinctive preferences over cases from various topic categories and different levels of harmfulness, and the cases involving socially sensitive issues were given more consideration and penalized more severely than common misbehaviors. We also

Andy Zhao: dz352@cornell.edu
Date submitted: 2025-04-15

explore how the platform leveraged the usually one-sided votes of digital jurors to endorse its final decisions and find that the reason notes given by crowdsourced jurors also revealed a serious issue of decaying motivation. Our findings offer important insights into the coordination between a social media platform and its volunteer moderators to moderate an online community and address the question of how an autonomic content moderation model can prevail or perish.

Keywords: Content Moderation, Platform Governance, Digital Jury, Weibo

Introduction

Governing an online community with millions of users is a demanding problem for social media platforms worldwide. Many modes of platform governance were tested to explore a feasible and efficient way to moderate user-generated content. Weibo, a leading Chinese social media that rose rapidly in the early 2010s, participated in this expedition with its international peers. Flooded with massive inappropriate content, this microblogging platform chose to learn from the common law judicial system and establish a novel content moderation system with crowdsourcing features in 2012 (Lei et al., 2013), which evolved and lasted till today.

While asking ordinary users to report their subjectively undesired content is a typical approach on social media, directly involving users in the moderation decision-making process is also not uncommon. For instance, Facebook first experimented with crowdsourced content moderation in 2010 (Eldon, 2010) and later launched the Oversight Board, a quasi-judicial body of third-party experts, to make consequential content moderation decisions (Klonick, 2019). X also had a crowdsourced note-taking system called Community Notes, originally known as Twitter Birdwatch in 2021, to leverage users to moderate inaccurate information (Pröllochs, 2022).

Comparatively, Weibo's moderation experiment was surprisingly radical and transparent in 2012, even though it was operated in an authoritarian society. The early 2010s was the time when liberal ideas were prevalent and widely accepted by Chinese netizens, and the Chinese government struggled to regulate all online discourses well (DeLisle et al., 2016). So many Chinese digital platforms gradually took some responsibilities of institutional development and

came up with experimental designs on the Internet, a process called "institutional outsourcing" by Liu (2024). Even though the government still gave directives to social media platforms to regulate content, the discretionary power of content moderation remained centralized within platforms like Weibo, similar to that of Western social media platforms like Facebook and Twitter (Miller, 2018).

A divergence occurred in 2012 that made Weibo's moderation more unique. In May, Weibo drafted a community convention and multiple regulations with experts from different domains and established a community-driven content moderation system in imitation of the common law system, in which ordinary users can possess the power to report inappropriate content and judge some cases by votes. We describe Weibo's system as community-driven in this paper because all moderation cases were reported by ordinary users and some cases were also judged by other users, rather than purely initiated and judged by the platform. Because of the autonomous nature of this governance model, Weibo has publicized millions of user-reported moderation cases, corresponding decision processes, and verdict results on its Community Management Center since the beginning.

In academia, there is always a solid call to understand how democratic components can be embedded in the content moderation process (Gillespie, 2018; Seering, 2020) and how digital juries may work in the context of online communities (Fan and Zhang, 2020). Weibo's large-scale and high-granular moderation data and policy documents offer a valuable window to answer these questions and understand how this community-driven content moderation system works. Furthermore, China's authoritarian governance provides a distinct perspective through which to examine how democratic content moderation systems on social media might evolve and adapt under the increasing regulation pressure. As more social media platforms (like X and Facebook) altered their moderation practices as the result of political and ideological influence (Miller, 2022; McMahon et al., 2025), the lessons of Weibo struggling to maintain a healthy moderation system become increasingly relevant in Western contexts (Flew, 2021).

Our core guiding question is *how Weibo facilitated a community-driven content moderation system, though not very successful, based on collaborative efforts from different users and the platform.* Focusing on the content moderation as a critical component of Weibo platform governance, this paper aims to answer this question from the perspective of three prominent

actors: the users who chose to report, the Weibo platform, and the users who played the roles of community jurors. Therefore, we have three specific research questions in this paper:

- (RQ1) How do users differ in their reporting practices, and what patterns characterize their reporting behaviors?
- (RQ2) How did the platform handle and respond to different types of reported cases?
- (RQ3) How did users participate in content moderation as jurors, and how did they engage with the system?

By exploring these research questions, we aim to make a theoretical contribution to the study of content moderation by highlighting both the potential and the limitations of incorporating democratic elements into moderation systems within authoritarian contexts. While prior studies also examined the crowdsourced components in content moderation (Pröllochs, 2022), they usually lack a more internal perspective to understand the interaction between different stakeholders in the content moderation. Our study extends the prior studies with an empirical and quantitative analysis of millions of moderation records from Weibo. We examine how users were motivated to file reports to Weibo and how the platform chose to process various reported cases in a different way strategically. We also explore how users, as digital jurors, engaged in the moderation decision-making process over time. In summary, this paper investigates how bottom-up participation intersects with top-down governance in content moderation, offering new insights into hybrid models of platform governance emerging from an understudied social media context.

Literature Review

In this section, we discuss the studies relevant to our core guiding research question: *how Weibo facilitated a community-driven content moderation system, though not very successful, based on collaborative efforts from different users and the platform.* We first introduce literature about content moderation in online communities to give a general theoretical background for our study. Then, we situate our study in the context of Weibo and Chinese internet.

Content Moderation in Online Communities

Content moderation in online communities has long involved both top-down governance by platforms and bottom-up efforts by users to restrict the undesirable contents (Gillespie, 2018; Gosztonyi, 2023). Aiming to align the user-generated contents to the norm of the platforms or the community, content moderation has been a challenge for social media for years.

Top-down governance, or the centralized platform-controlled content moderation remains the dominant model on social media. Gillespie (2018) argued that pursuing a truly open social platform is infeasible and that moderation is necessary for all online communities. Given that traditional top-down moderation approaches such as adopting teams of human moderators or algorithms automatic moderation, have been criticized for inconsistency and bias, Gillespie (2018) raised questions for platforms about when, how, and why to intervene. Gorwa (2019) argued that finding an appropriate balance between external governance, self-governance, and co-governance will be key to answering the question of how to govern the platform.

In contrast, a growing body of research has examined bottom-up participation in content moderation, highlighting the potential of community-driven approaches. Surowiecki (2005) contended the potential for collective intelligence to outperform individual judgment, a theory that forms the foundation for community self-moderation. Kraut and Resnick (2012) argued that community members who have been deeply embedded in the social context are capable of enforcing community norms and standards. Further exploring this concept, Matias (2016) explored the role of volunteer moderators in Reddit communities and proposed that community self-moderation can be both reactive—where users report offensive content—and proactive—where users jointly create and uphold regulations. He noted the difficulties and potential for misuse in this grassroot approach, such as bias, mob mentality, and a lack of transparency. Bernstein et al. (2011) found that crowdsourced moderation can lead to 'popularity bias', where the most popular content receives the most attention, regardless of its quality or relevance.

Meanwhile, there is a strong need to understand how the platform, users, and policies shape platform governance mechanisms and other external factors (Bucher and Helmond, 2017; Gorwa, 2019). Bajpai et al. (2022) proposed a framework of 'members, infrastructure, and content' to understand the moderation practices on different platforms macroscopically. Seer-

ing (2020) discussed the two perspectives in moderation research: the platforms and policies perspective, and the community perspective. Attempt to understand how the self-moderation social structure functions and how it might be improved, Seering (2020) also raised six important questions for future works about self-moderation and system design. These works guided our analysis about different perspectives in Weibo's moderation.

Apart from Seering (2020), other scholars also examined the practical implementation or implications of crowdsourced self-moderation. Kiene et al. (2016) analyzed Wikipedia's moderation system, demonstrating how community-led efforts can create a robust and decentralized moderation system. They suggested that the size of the community and platform structure are the factors to a more amenable self-moderation structure. De Gregorio (2020) argued for a more democratic, transparent, and accountable process for content moderation, as business interests could harm the fundamental value of free speech. Fan and Zhang (2020) discussed the idea of digital juries as a model of civics-oriented content moderation, which involves online community members adjudicating moderation cases. They suggested that, while jury decisions are perceived to have better procedural justice than existing moderation methods, users still have no consensus on whether the platform action should be bonded with jurors' votes. However, an experiment within subjects suggests that users perceive expert panels with higher legitimacy than digital juries (Pan et al., 2022).

Twitter's Birdwatch, now known as X Community Notes, was also a valuable experiment of crowdsourced content moderation. Pröllochs (2022) found that users did make helpful fact-check notes to counter misinformation, but the effectiveness of these fact-checks were associated with other factors like source, sentiments, or user consensus. Chuai et al. (2024) also took advantage of this natural experiment on Twitter/X and investigated the true effects of community notes, and found that interventions from ordinary users actually take too long to be effective in addressing misinformation.

Furthermore, some studies investigated the dynamics in the hybrid moderation models on social media platforms that involved more top-down and bottom-up elements, which brought us more insights about how corporate interests may drive the content moderation or platform governance. Wang et al. (2024) examined the user-initiated discussions about platform governance on Twitter/X. They found that user efforts were usually fragmented and constrained

by the platform's own operation goals, which undermined the crowdsourced power to drive meaningful governance change on the platform. Scharlach et al. (2024) studied the policy documents on multiple popular social media to discover which values these platforms pay attention to. Scharlach et al. (2024) found that platforms are more likely to mentioned "expression, community, safety, choice, and improvement", but also selectively shoulder off their duties to users or strategically reinterpret these values to accommodate their own interests.

In summary, these above studies provide valuable guidance for our analysis and a theoretical background of hybrid or user-driven content moderation. Our work extends these studies and makes valuable empirical contributions by examining another unique case of community-driven moderation, in terms of context, case discovery and assignment method, user participation criteria, involvement mechanism, and the general governance institution.

Content Moderation on Weibo and in China

In China, content moderation is intertwined with the censorship, since the natures of both are to restrict the information and many people may also interpret them sometimes interchangeably (Roberts, 2018; Myers West, 2018; Gosztonyi, 2023). It is true that the content moderation on Chinese social media is more state-driven compared to American and European platforms (Van Dijck et al., 2018), and Chinese content moderation in practice is close to the technical level of censorship by restricting netizens from posting and browsing content (Zittrain and Edelman, 2003; MacKinnon, 2008). However, Li (2023) pointed out that most studies on Chinese social media, particularly Weibo, have ignored the issue of platform governance and tend to adopt a dichotomy view between censorship versus online activism. Li (2023) found out in the early days of Weibo (which is also the focus of our study), the platform actively promoted the generation and spread of online debates to establish itself as a key channel for users to discuss and follow the development of social events. In recent years, the traditional semantic moderation practices on Chinese platforms evolve to a more proactive and nuanced approach to control information flow (Li and Zhou, 2024), which we believe aligns closer to censorship practice.

Our work is not the first study to examine Weibo's content moderation system. Kou et al. (2017) used an interview to suggest that while users are passionate about joining Weibo's

self-governance, they usually struggled with the moderation system. Chin et al. (2022) extended Gorwa (2019)'s framework to the governance of false information on Weibo and highlighted Weibo's governance model as a state-led model that combines external, co-, and self-regulation. Chin et al. (2022) posited that the absence of a substantial system for accountability, transparency, recourse, or due process in both procedural and final decision-making stages creates a spiral of privatized regulation on Weibo.

Our paper extends from the above works about Chinese content moderation to a more microscopical but large-scale data analysis. We confirm the existing gap between the platform's expectations and how users perceive their work by investigating the strategic decisions from the platform and the diverse motivations from the users.

Data and Methods

In this section, we first give a brief background introduction about the history and the operation of Weibo's content moderation, as a necessary base for our introduction about data collection. Next, we give detailed descriptions about Weibo moderation case and what information we have. Last, we introduce the methods we used in this study.

Background

Millions of moderation cases, dated back from 2012, are publicly accessible on the website of the Weibo Community Management Center, which is still in use to store new moderation cases till now. However, it is noteworthy that this content moderation system which significantly involved the user participation in the moderation decision was de facto suspended in 2014 for unclear reasons.[1] The role of users in Weibo content moderation was also degraded in the following years, as shown in the Weibo's updated community conventions. Even when this moderation system of digital jurors was revived on Weibo in 2021, the significances of users were largely weakened compared to those of users in 2012. For more detailed background about this community center, please refer to Appendix A.

[1]As a matter of fact, the suspension of Weibo's community-driven content moderation occured concurrently with the rise of Cyberspace Administration of China, one of the most powerful governmental agencies that could supervise or direct social media companies.

While the internet is rigidly controlled and the content moderation is always intertwined with censorship in China, social media still had limited freedom to facilitate regular public discussion and foster a relatively healthy environment (Qin et al., 2017). Although Chinese content moderation has become more aggressive in the recent years (Li and Zhou, 2024), we still believe the Weibo community management center is strictly and naturally separated from censorship or other moderation actions. This is because Chinese censorship is usually centralized from and initiated by the platform, and the information control practice is difficult to detect (Roberts, 2018). In contrast, community-driven moderation initiated by Weibo users is under public scrutiny and does not cover the contents that are forbidden and should be deleted right away by the platform.

Data

We collected eleven million moderation cases, ranging from 2012 to 2021, from this website by programmatically scrapping. We obtained all the cases accessible on the Weibo Community Management Center, however, we still do not guarantee that our collected data is a complete dataset of all Weibo user-reported moderation actions. This is because Weibo never made a public announcement to promise all user-reported moderation cases in the past ten years were stored in the same way in this center. Because the full picture of the Weibo content moderation system is still a black box, we cannot verify the completeness of our data coverage.

The pipeline of the Weibo community-driven content moderation system is shown in Figure 4 in the Appendix. Users could file a report on the subjectively undesired content to Weibo, and the platform could make a decision about how to process it: (1) *accept* this case and directly judge it by Weibo; (2) accept this case and *forward* it to the community committee, and let Weibo jurors make a decision; (3) explicitly *dismiss* this case with a reason by Weibo; (4) *ignore* this case and just leave it unanswered as a mistrial. After meeting some minimum requirements like being at least 18 years old, all active users can apply to join the community committee. For each case that Weibo chooses to accept and forward to the community committee, several committee members would be randomly selected to form a jury. Most cases would be forwarded to regular committee juries, which are made up of ordinary users; misinformation cases and appeal cases would be forwarded to expert committee juries, which are made up of verified experts in some domains. These jurors need to vote independently

in 24 hours to judge whether the reported side violates the platform policy and post a note of their vote reasons.

In each moderation case (see Figure 5 in the Appendix), there is a reporting side (plaintiff) and a reported side (defendant), whose profiles are all public. One or several reporting users would make a statement about why this content should be subject to moderation. The reported user has the right to defend the content with one or more statements. The reported content is also public unless it was deleted by the reported user or hidden by Weibo to cover private or harmful information. For all cases, moderation results are public; if committee members judged a case, then there is jurors' vote information along with their reasons (see Figure 6 in the Appendix).

All cases have an associated topic category selected by the reporting users in Weibo's system. These topic categories, pre-set by Weibo, include verbal abuse, plagiarism, harassment, misinformation, impersonation, spam ads, doxing (leaking other people's privacy), and others. In addition, given the strict censorship in Chinese cyberspace, we do not expect (or see) many reported contents that conflicted with the state ideology and could reach to the reporting stage. There were indeed some cases that users argued over some facts about social or public affairs. However, the 'misinformation' cases received an increasing dismissal rate in the following years, which we would show later in our paper.

Methods

The primary method in this study for all research questions is exploratory data analysis (EDA). Then, as the foundation of our data analysis on the reporting behaviors (RQ1), we classify users who made at least ten reports into two types: low diversity frequent reporters (LDFR) and high diversity frequent reporters (HDFR). We define LDFR as users whose report number is *at least three times* the number of unique accounts reported by them, which means that they made at least three reports to each reported user on average. On the contrary, we define HDFR as users whose report number is *at most two times* the number of unique accounts reported by them, which means that they usually would not file multiple reports on a single user. (There is a long tail distribution of report numbers per unique reported users after around 2 times.) Under this definition, we pull out the reporting relations between users, calculate the total report numbers

and the unique reported users for each reporting user. Dividing the total report numbers by the unique reported users for each reporting user, we identify 22,014 LDFR and 112,212 HDFR in total.

Next, we also leverage network analysis for RQ1, as the reporting relationship in content moderation can be naturally modeled as a directed edge from the reporting user to the reported user. But in our study, we build up another edge from the reporting relationship: we identify the reporting users who both reported the same target users and regard this co-reporting relation as the subject we study here. This is because we are interested in the coordinated behaviors in the reporting activities, and lack the user profile information of both the reporting and the reported users.

Then, we also use regression analysis as well to examine which features of the reported content are statistically associated with the platform's moderation decisions (RQ2). This approach allows us to identify potential patterns or predictors that may influence how content is handled, offering insight into the factors that shape enforcement outcomes.

In our regression analysis for RQ2, we use the content moderation API provided by Baidu AI services [2] to detect harmful content in Chinese text. Politically sensitive content was clearly indicated by Baidu's model response and was also rare in the reported content (we believe they were censored before being reported). Since the criteria for political sensitive content is very subjective, arbitrary and unclear, we exclude the political text and built a binary harm indicator for each reported content based on Baidu's responses. We note that the presence and exclusion of politically sensitive content do not imply direct government influence over user reporting behavior; rather, they primarily reflect Baidu's classification criteria, not users' perceptions.

Last, information entropy is employed to measure user engagement (RQ3), because it captures the diversity and uniqueness of responses across text sets of varying lengths. The less likely a vote reason is to be found in all cases, the higher the information entropy, and the higher the juror's engagement. Higher entropy reflects more efforts being put in the jury duty, making it a suitable metric for comparing engagement levels across different user groups. For

[2]https://ai.baidu.com/solution/censoring

example, the most common reason 'Decide.', which appears more than 1.2 million times in our dataset, has only 2.12 bits of information entropy. 'The reported side has evidence, fact, and basis' is a reason that has only been written once, so it has the highest entropy: 22.14 bits. This method provides a quantitative way to reveal the level of cognitive and physical investment by the Weibo jurors in their moderation tasks.

Results

In this section, we present our results following our three research questions. First (RQ1), we show the results about the reporting activities from users that formed the basis of community-driven content moderation. We examine two special reporting behaviors: coordinated reports and frequent reports, and then infer the motivations behind these actions. Next (RQ2), we present the findings about Weibo platform actions in terms of its case process and penalty decisions. We demonstrate that Weibo did not treat all types of cases in an equal way and strategically leveraged the user jury system. Last (RQ3), we uncover the user voting patterns in Weibo jury system and then investigate how user engagement, which is essential to the output side of this community-driven content moderation system, changed over time.

Driven by Reporting Activities

All cases in the Weibo content moderation center were made possible by users who filed reports, so the moderation practices in this system are initiated and driven by the whole Weibo community. More than 4.18 million unique users in our dataset have filed a report to Weibo, and 824,433 users made at least three reports. These ordinary reporting users were essential actors in Weibo's system. This community-driven content moderation could happen without these users' inputs, and the reporting pattern could also shape the work contents of this moderation system.

To answer RQ1, we choose to investigate the key patterns in the reporting activities made by reporting users. We examine these actions with questions whether the reports were coordinated and what characteristics are in the users who initiated reports. We do not dive into the report reasons because most reporting statements were short and straight with simple terms like 'abuse' and 'swear words'.

Coordinated Reports

A large network of reporting activities brings up the question of coordinated reporting behaviors. We define co-reporting relations between two users as making reports to the same target at least three times and then identified 928 users who have frequent co-reporting relations. We use the co-reporting relations as the edges and the users as the nodes to identify 227 clusters, which are defined as the group of nodes connected by edges, or the users who shared the same reporting target at least three times. The biggest cluster we find has 68 nodes, followed by a cluster with 26 nodes. The mean size of our identified reporting clusters is 4.09, and the median size is 3. The users in these co-reporting clusters contributed fourteen thousand reports in total.

After manually examining the nodes in a few biggest clusters, we find that many users are fans of singers who reported subjectively undesired content in a coordinated way as Zhao and Chen (2023) described (but on a much smaller scale). Besides, many users in big co-reporting clusters frequently reported inappropriate posts together, who may simply agree on the issues with each other rather than strategical coordination. Our results suggest that mass reporting behaviors did exist on Weibo, but not on a large scale as expected in the content management center.

Frequent Reporters

Then, we turn to the frequent reporters who did not necessarily co-report with other users, as defined in the section 3.3. We study these frequent reports, low diversity frequent reporters (LDFR) and high diversity frequent reporters (HDFR), based on three variables: content category, the status of their reported cases, and winning chance.

Table 1 shows that LDFR made more reports about verbal abuse and harassment than average and cared less about spam ads cases, and HDFR had disproportionately more reports on spam ads but noticeably less on verbal abuse and harassment cases.

As for the case status, Weibo ignored about seventy percent of all cases reported by

Table 1: Percentage of case category of high and low diversity frequent reporters

	All Users	*LDFR*	*HDFR*
Spam Ads	56.70%	32.69%	70.29%
Verbal Abuse	31.41%	50.05%	20.42%
Harassment	5.15%	8.59%	2.15%
Impersonation	2.47%	1.75%	4.43%
Misinformation	2.11%	1.75%	1.47%
Others	0.73%	2.53%	0.51%
Doxing	0.65%	1.33%	0.22%
Plagiarism	0.60%	1.24%	0.41%
Total	100%	100%	100 %

HDFR, as Table 2 shows, most of which were about spam ads. The remaining cases from HDFR, primarily related to verbal abuse, were more likely to be accepted by the platform. The good news for HDFR is that their cases were less likely to be dismissed. Comparatively, LDFR had less chance of being ignored and higher rates of reviewed by the Weibo jury.

Furthermore, we find that all frequent users have a higher chance of winning than average cases (58.5%) if judged, which is not surprising if they mastered the reporting strategies. Fortunately, HDFR were treated better this time and had a higher winning chance (73.7%) than LDFR (61.1%).

Concisely, users who initiated moderation reports were usually driven by different motivations, which led to their coordinated or frequent reporting activities. Weibo did not have an equal way of processing the reported cases from these users, partly because of the nature of different cases reported by different types of users.

Strategical Governance of the Platform

How Weibo chose to process moderation cases and what types of penalties to use reflect the underlying intention of its governance strategy that is not publicly written in policy documents. Additionally, the Weibo platform retains significant power throughout many stages of this

Table 2: Case status of high and low diversity frequent reporters

	All Users	*LDFR*	*HDFR*
Accept	23.60%	36.00%	20.65%
Dismiss	17.78%	25.21%	7.74%
Ignore	56.33%	31.29%	69.93%
Transfer	2.29%	7.51%	1.68%
Total	100%	100%	100 %

community-driven content moderation system, which allows it to quietly bend final outcomes in its favor. Therefore, investigating the case decisions of Weibo platform offers an opportunity to reveal the governance mind of a business-driven platform in content moderation. To answer RQ2, we explore the governance strategies from the Weibo platform from multiple dimensions. We investigate how Weibo platform processed and allocated different types of cases. We examine the factors associated with the platform decisions on case outcomes. We explore the reasons that Weibo cited for case dismissals. We also study the exact penalties decided by Weibo over different content categories. In summary, our analysis in this section aims to expose what Weibo is more concerned about from the platform actions in different ways.

Reported Case Processing and Allocation

After a case was reported by users, the first fork of its lifeline is the process decision made by Weibo: how to handle this case? Accept and judge by Weibo? Accept and transfer to community juries? Dismiss? Or just ignore it?

The category information of reported cases, which were selected by reporting users, revealed the general picture of Weibo moderation. For cases reviewed by the Weibo authority, which count around 41.38% of total cases in our data, 72.25% of them are 'verbal abuse' cases and 10.60% cases are 'harassment', followed by 5.97% 'impersonation' and 5.10% 'misinformation'. Figure 1 shows the status and platform decision of cases judged by Weibo in these categories. Plaintiffs won in most situations if their cases were accepted. The biggest winning gaps happened in verbal abuse and misinformation cases, where the won and the lost numbers

differ by three orders of magnitude. The misinformation category has the lowest number of defendant victories from Weibo's direct decision: 39.

However, Figure 1 suggests a different tendency to dismiss a case under a different topic category. In some categories, Weibo chose to dismiss more cases than accept. Only 14.8% 'harassment' cases (harassing or disturbing other users with repeat posts) and 16.9% 'misinformation' cases were not dismissed by the Weibo authority. For categories like 'verbal abuse' (attacking other users with abusive language), roughly two-thirds of cases could be accepted and reviewed by Weibo. This difference reveals where Weibo wanted to shy away and where it had more concerns. Another noteworthy category for dismiss action is 'spam ads', which we discuss in Appendix C because most of 'spam ads' cases were mistrials.

Our next focus is the cases transferred to the community committees. The remaining 2.30% of total cases in our dataset were accepted by Weibo and then transferred to Weibo community committee jurors. Narrowing down the time window to 2012 and 2014, when the Weibo jury system was still active, we find roughly 11.9% of total cases were transferred to jurors during that time. As Table 3 shows, in these jury-verdict cases, 64.67% cases are 'verbal abuse', 33.26% cases are 'harassment', 1.94% cases are 'plagiarism', and 0.12% cases are 'misinformation'. Comparatively, harassment cases were more likely to be transferred to committee members, and misinformation and impersonation cases were much less likely to be handled by committee members. Meanwhile, the winning chance of defendants (the reported users) increased from 0.23% (platform-verdict) to 2.70% (jury-verdict) in verbal abuse cases, from 2.84% to 7.40% in harassment cases, 4.01% to 7.26% in plagiarism cases, and 0.10% to 9.84% in misinformation cases. Also, the expert committee handled appeal cases. But only 1.4% of 25,489 appealing cases, mostly spam ads, received a reversed verdict.

The trend of committee cases in each category reflects how Weibo collaborated with community members over time. In the first year, Weibo mainly transferred plagiarism cases to the regular committee and was willing to ask for help with misinformation cases. However, this pattern quickly changed in late 2013, and the verbal abuse and harassment cases increased by about a thousand times and dominated the task of committee members.

Figure 1. Waffle Plot of all accept and dismiss decisions from Weibo by content categories. Each square, no matter the size, represents one thousand cases (less than one thousand cases would be ignored)

Table 3: Number of community committee decision by categories

Category	**Winning Side**	
	Defendant	**Plaintiff**
Harassment	6364	79651
Verbal Abuse	4518	162724
Plagiarism	365	4663
Misinformation	31	284
Impersonation	2	19
Doxing	2	1

Factors Associated with Platform Decisions

As Figure 1 suggests, Weibo did not process different types of reported cases in an equal way. Therefore, we would like to investigate what factors were associated with the Weibo platform's decision to disclose the moderation strategy behind.

Since some old user profiles cannot be obtained now, we turned to the unchanged information: the reported content. Filtering around four million cases with publicly available and informative reported content, we obtained roughly 2.2 million cases with positive harm indicators and 1.8 million cases with negative harm indicators. Next, we build our models with the variables including the case category, case process decisions, case outcomes, and the harmfulness indicator identified by Baidu AI, as we explain in the section 3.3.

Our first multinomial logistic regression model (Table 4) examines how the case topic category and whether being harmful or not would affect Weibo's process decisions (reviewed by Weibo, ignored by Weibo as a mistrial, and forwarded to committee members). Generally, different case topic categories indeed could determine the case process choice significantly, and such effects vary category by category. Weibo also tends to make a judgment themselves when the reported content contains some harmful information. Additionally, we do not assume the influence of harmfulness is equal across all case categories and the linguistically benign content may still pose serious danger, so we chose to include multiple interaction terms. We find that the interactions between case categories and harm indicators have significant implications on

case process choices.

Our second two-way logistic regression model (Table 5) examines how the case topic category, whether harmful or not, and the process method could collectively affect the moderation decision (plaintiff win or loss). While the case topic category and being harmful or not have significant impacts on the final decisions, the process method itself has no significant effects after controlling other variables (for example, mistrial cases do not appear to meaningfully affect outcomes). However, several interaction terms between the committee process and specific case categories (on the right column) indicate that committee members indeed were much more likely to support the plaintiff involving misinformation, verbal abuse, harassment, plagiarism, or containing harmful content. Also, the effect sizes of these interactions are usually much larger than other significant interactions with 'Harm Indicator'. This suggests that the committee does not exert a uniform effect across all content types but instead plays a more targeted role in certain types of disputes.

Reasons about Case Dismissals

To further understand the motivations behind Weibo moderation decisions, we turn to the reasons cited by the platform when cases were dismissed. We extract all verdict texts from dismissed cases and identify five frequent reasons by counting and manual reviewing for relatively unstructured verdicts.

Figure 2 shows how Weibo referred to different reasons to dismiss different types of cases. Ineligibility is one of the most frequent reasons cited in verdicts, especially for misinformation and spam ads cases. Also, 'not a party to the case' is another frequent reason, and common in law practice, for more personal reports like verbal abuse and harassment. For controversial categories like misinformation or plagiarism, an unsubstantial ground from the plaintiff is another frequent reason Weibo chose to dismiss a case. Meanwhile, Weibo sometimes dismissed harassment and misinformation cases because the reported content was already invisible to the public.

It is noteworthy that the misinformation category has the highest proportion of cases

Table 4: Regression Table of the Case Process Model. The default level of the case category is 'Others', the default level of the Baidu API harm indicator is 0 (not harmful), and the default level of case process choice is Weibo's review.

	Case Process Choice (default: Weibo review)	
	Mistrial	Committee
Misinformation	-3.922^{***} (0.000)	3.812^{***} (0.702)
Troll	7.752^{***} (0.167)	6.830^{***} (0.700)
Impersonation	5.739^{***} (0.435)	0.618 (0.743)
Spam Ads	20.092^{***} (0.150)	-11.289^{***} (0.000)
Plagiarism	8.804^{***} (0.240)	7.800^{***} (0.700)
Doxing	-1.799^{***} (0.00001)	2.181^{**} (0.908)
Harassment	4.843^{***} (0.596)	8.594^{***} (0.700)
Harm Indicator (Baidu Moderation API)	-1.899^{***} (0.308)	-5.131^{***} (0.175)
Misinformation:Harm Indicator	0.300^{***} (0.000)	4.879^{***} (0.229)
Troll:Harm Indicator	1.298^{***} (0.322)	6.069^{***} (0.175)
Impersonation:Harm Indicator	-8.918^{***} (0.000)	4.962^{***} (0.859)
Spam Ads:Harm Indicator	1.628^{***} (0.309)	0.561^{***} (0.000)
Plagiarism:Harm Indicator	-50.816	5.280^{***} (0.180)
Doxing:Harm Indicator	-1.324^{***} (0.000)	-10.373^{***} (0.000)
Harassment:Harm Indicator	2.306^{**} (0.918)	5.499^{***} (0.175)
Constant	-16.715^{***} (0.150)	-10.206^{***} (0.700)
McFadden's R^2	0.573662	
Observations	4,056,289	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 5: Regression Table of the Case Decision Model. We used a two-way logistic regression model because it has a good balance between the model complexity and the Akaike information criterion. The default levels of involved variables are 'Others', not harmful, Weibo's review, and plaintiff loss respectively.

	Decision Result		*Decision Result*
Misinformation	−1.380*** (0.010)	Misinformation:Harm Indicator	−0.328*** (0.025)
Verbal Abuse	−0.308*** (0.009)	Verbal Abuse:Harm Indicator	0.649*** (0.019)
Impersonation	0.816*** (0.010)	Impersonation:Harm Indicator	−0.906*** (0.024)
Spam Ads	2.175*** (0.032)	Spam Ads:Harm Indicator	−1.216*** (0.094)
Plagiarism	−0.518*** (0.012)	Plagiarism:Harm Indicator	−1.626*** (0.048)
Doxing	−0.504*** (0.023)	Doxing:Harm Indicator	0.645*** (0.060)
Harassment	−1.494*** (0.010)	Harassment:Harm Indicator	−0.665*** (0.021)
Harm Indicator	0.364*** (0.019)	Verbal Abuse:Mistrial	−5.881 (35.010)
Mistrial	−7.650 (33.784)	Impersonation:Mistrial	−2.517 (47.067)
Committee	0.014 (1.414)	Spam Ads:Mistrial	−14.008 (36.785)
		Plagiarism:Mistrial	−2.522 (44.759)
		Misinformation:Committee	3.520** (1.427)
		Verbal Abuse:Committee	2.659* (1.414)
		Impersonation:Committee	1.930 (1.750)
		Plagiarism:Committee	3.092** (1.415)
		Doxing:Committee	−0.190 (1.871)
		Harassment:Committee	3.810*** (1.414)
		Harm Indicator:Mistrial	1.249 (34.561)
		Harm Indicator:Committee	1.117*** (0.024)
		Constant	−0.014 (0.009)
McFadden's R^2			0 .2224144
Observations			4,056,289
Note:			*p<0.1; **p<0.05; ***p<0.01

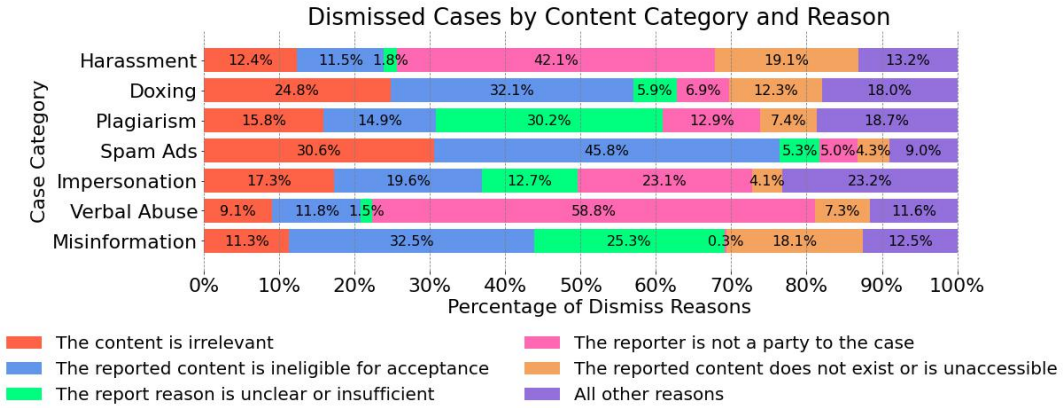

Figure 2. Dismiss reasons in verdict text for all types of cases

(57.8%) dismissed in the name of ineligibility or insufficiency, the two most ambiguous reasons that can be arbitrarily cited and require more efforts from plaintiffs to improve. Figure 8 in the Appendix also shows the trend of acceptance and dismissal of misinformation cases between 2012 and 2014, which suggests a clear tendency to stay away from this special category.

Furthermore, we wonder if the number of plaintiffs is associated with the decision to transfer a case or not. Filtered out all ignored cases, the moderation cases with only one plaintiff have a 2.3% chance of being judged by the Weibo jurors. In comparison, 10.1% of cases with two plaintiffs and 18.7% of cases with three plaintiffs were transferred by the Weibo platform to the jurors. For cases with more than three plaintiffs, Weibo chose to transfer 37.8% of them to the jurors. While it is difficult to argue this is a causal relation, our results imply that a case that more users reported has a higher chance of being transferred to crowdsourced moderators rather than directly judged by the platform.

Penalty and Sentence in Moderation Decisions

In addition to general process decisions, the choice of penalties and their intensities for different cases reveals the governing mind of Weibo's authority. Therefore, we want to use the penalty information to understand the moderation decision pattern of Weibo platform.

Since Weibo moderation intentionally mirrored the judicial system, we borrowed the understanding of the primary purposes of criminal punishments: retribution, incapacitation, deterrence, rehabilitation, and reparation. Given the nature of social media, retribution, reparation, and rehabilitation are largely impractical, leaving deterrence and incapacitation as the most plausible goals. We extract the most frequent sentence statements from Weibo verdicts and classify all penalties into nine types, shown as Table 7 in the Appendix.

Table 6 shows the percentage of penalties used in different content categories. Since one case may incur multiple penalties, the sum of all penalties is above 100%. We find that most Weibo penalties are indeed close to deterrence, for example, credit deduction. As one of the most frequent penalties, credit deduction reflects how Weibo wants to govern with fewer efforts and more effects. Other penalties may share similar characteristics with incapacitation; for instance, account suspension nullifies the action capacity of a user.

Table 6: Percentage of penalty decision under each content category

	Misinfo	**Abuse**	**Impersonate**	**Plagiarism**	**Doxing**	**Harassment**
Suspension	12.17%	0.02%	0.00%	6.01%	0.86%	0.02%
Follow Ban	11.62%	0.00%	0.00%	1.85%	0.35%	0.01%
Credit Deduction	60.99%	99.12%	16.24%	91.76%	52.21%	88.95%
Dismissal	0.51%	0.25%	2.12%	2.77%	20.01%	0.46%
No Penalty	0.13%	0.20%	0.01%	4.12%	0.13%	4.70%
Account Deletion	0.37%	0.26%	79.56%	0.05%	21.55%	3.46%
Content Deletion	0.03%	62.83%	2.04%	0.12%	39.39%	50.12%
User Info Deletion	0.01%	0.18%	10.74%	0.00%	0.12%	0.00%
Flag	36.57%	0.00%	0.00%	0.40%	0.00%	0.00%

Weibo differentiated its governing strategies by different violations and their policy. Surprisingly, misinformation is (again) a special major category that has incapacitation-type penalties like account suspension and follow ban. Meanwhile, being the most socially sensitive and impactful category, misinformation is the only category that received a significant amount of 'flag' penalties (explicitly labeled as wrongful content), resembling a damage control strategy. (We do not identify a clear pattern about what types of misinformation cases got penalized most) For other categories close to interpersonal conflicts like verbal abuse, harassment, or plagiarism cases, Weibo was more likely to erase harmful information and use credit deductions to deter such misconduct. Doxing cases have a similar pattern where both deterrence and damage control strategies are important. Impersonation is a different category because most

violators were directly deleted.

Furthermore, some penalty types have continuous levels by days or by points, which allows us to examine the intensity of sentencing decisions and what cases would have the worst penalty. (Because account suspension and follow ban are almost identical in our data, we only consider account suspension in this analysis.) Table 8 in the Appendix shows different levels of points deduction in different categories, which tells a similar story that misinformation usually received higher credit deduction (5 or 10) while verbal abuse, plagiarism, and harassment usually got 1 or 2 points credit deduction.

In general, Weibo platform strategically handled moderation cases with different channels and implemented different levels of various penalties, based on the topic categories, the harmfulness levels, and its capacity. Socially sensitive topics like misinformation usually received higher dismissal rate and a harsh punishment, especially after 2013.

Users as Jurors

Weibo platform promised to make final sentencing decisions based on the jury results, so investigating how users participated in this content moderation system as crowdsourced jurors is essential to understanding the role of the community. The analysis of voting records and stated reasons could provide important insights about the contributions and the dynamics of jurors on Weibo. In this section, we first examine the voting behaviors of committee members as jurors and explored their decision patterns and their relation to Weibo's sentences. Then, we leverage the reason text in the jury votes to measure the changing trends in the juror engagements.

Voting Behaviors

Weibo community committee was popular among users. There were 76,131 unique users who have served as Weibo jurors till the end of 2021. The median number of votes per juror is 9, and the most enthusiastic user voted around 1,300 times.

Jurors predominantly favor the plaintiff in all categories. 84.5% of all votes went for the

plaintiff. This strong tide to stand with the plaintiff is not surprising, given that these cases were reported by users who thought the reported sides violated some norms. Most users cast most of these votes for the plaintiff. 47,282 unique jurors cast at least five votes during their careers, and 2,862 of them, roughly 6.1%, always say 'yes' to the plaintiff. Only 769 (1.6%) users have voted for the plaintiff less than once in three times.

While most cases received one-sided votes, Weibo's jury system still offered an opportunity to examine the (mild) public opinion conflicts among Chinese users. We have put our analysis of contentious cases into Appendix D.

Juror Engagement

Next, the engagement of crowdsourced workers like moderation jurors is always a necessity to sustain high-quality work. So, we leverage the notes of vote reason to measure the dedication of jurors to this moderation work. Because uncommon and customized reasons require more effort and may signal the authors' commitments, we chose information entropy, as explained in the section 3.3 to measure the engagement levels in the jury duties.

In addition, to understand the changes in juror dedication, we group users who cast their first votes in the same month as cohorts and calculate their average information entropy per vote reason in all following months. Figure 3 (a) shows how each cohort's information entropy of vote reasons changed over months. There are several noteworthy insights. First, information entropy, or dedication to jury work, generally decreases over time, especially in the last few months, and this trend is consistent with all cohorts. For example, in 2013, more and more users chose to just use a blank space or a random punctuation as a placeholder for the vote reasons, instead of typing a short sentence. Second, with decreased positivity over time in existing cohorts, newcomers also started with a decreasing enthusiasm level. Meanwhile, Figure 3 (b) shows that the winning rate of the plaintiffs in jury-verdict cases had an opposite trend to the engagements of jurors and gradually stabilized to a very high level.

In summary, crowdsourced users had a strong tendency to support the reporting users in moderation decisions, but their dedication to this work faded away after one-year of the

operation of community committee.

Discussion

The experiment of Weibo's community-driven content moderation system is a unique case worthy of deep examination. Not only because this experiment was a rare democratic practice in an authoritarian society, but also because it pioneered an autonomous model of content moderation with over 200 millions of users in 2012 that involved both the platform and the ordinary users to moderate a large online community. The success, or failure, of this experiment should be examined from both top-down and bottom-up perspectives.

Top-down Governance

Our work bridges the knowledge gap about the platform's preference and tendency to deal, or not deal, with different types of misconduct on social media. Content moderation usually works in the shadow, so it is hard to understand the propensity of a moderation system without knowing what cases are ignored and dismissed in this blind spot. Therefore, our study examines millions of moderated cases, above or under the water, and exposes a systematic pattern of processing and penalizing cases by categories and reports on Weibo. Our results suggest that the case topic category and the harmfulness collectively affect how a platform decides to process a case, though the exact effects vary by category. The final case decisions were also significantly affected by the case category and harmfulness.

Misinformation, rather than prevalent interpersonal issues like verbal abuse and harassment, received the most severe punishments when the reported contents were proven by fact-checks or clear evidence to be false. Weibo also tended to dismiss misinformation cases with vague reasons like ineligibility rather than accept, and these dismissed misinformation reports were usually the cases without official fact-checks or grounded evidence. This contradiction reflects that the platform intends to address this challenge but is reluctant to step into muddy water. Misinformation category is socially sensitive because it is a battleground for different arguments on the veracity of many claims related to public or social affairs in China. And fact-checking some claims may raise attention to some public issues or even challenge governmental legitimacy (Fang, 2022), like Figure 5 in the Appendix. Considering the difficulty of fact-checking and the political risk, it is understandable that Weibo must avoid tricky cases

and only step in when there is an official endorsement. Meanwhile, it is unknown whether misinformation content would also be moderated through other channels, for instance, directly restricted.

Moreover, how and when to leverage crowdsourced powers is important in Weibo's process decisions. Our regression results suggest that the user committee does not significantly differ from the centralized platform in terms of decision outcomes. This indicates that the value of the jury lies not in systematically different ruling criteria, but in introducing more deliberative scrutiny in cases where centralized content moderation may fall short. However, only a small part of cases was transferred to juries, which seemingly did not significantly reduce the workload of Weibo staff. The cases with more plaintiffs were more likely to be transferred to the jury, and plaintiffs could win dominant votes in most jury-verdict cases, which indicates that these cases are probably not ambiguous and challenging. This phenomenon results in a big puzzle about the transfer decision: why bother (transferring these easy cases)?

We think the following evidence could partially explain this puzzle. First, Table 3 and Figure 1 suggest the categories needing more subjective judgments are more likely to be transferred (and dismissed). More plaintiffs may not (only) suggest that a case is clearer and easier but rather indicate more public attention. So, Weibo platform could leverage committee members to handle cases, mostly about interpersonal conflicts, that may incur discontent if misjudged. The votes from other users could be seen as an endorsement of mass opinion to prevent potential resistance if anyone is not satisfied with the final verdict. This explanation is also consistent with Weibo's new community committee system, which has hundreds of jurors for each case and plaintiffs still win most cases by a large landslide[3]. Leveraging crowdsourced jurors for demanding and subjective cases signals the bureaucratic motives of this content moderation system to dodge tricky tasks (Habib and Nithyanand, 2022). In other words, echoing Wang et al. (2024)'s finding that users are often constrained by the platform in their governance efforts, Weibo subtly shaped user participation in content moderation, even when the community took a greater role than usual, to align with its own interests.

[3]https://service.account.weibo.com/h5/council/votedpush

Bottom-up Participation

Crowdsourced moderators are expected to help a centralized content moderation system greatly (Fan and Zhang, 2020), but it is also difficult to maintain the motivations of crowdsourced moderators. While Weibo's initial documents emphasized how users are valued in this community-driven content moderation model, it still failed to mobilize their users to consistently and positively contribute to this moderation system.

First, decreased information entropy in vote reasons indicates that there are still no substantial incentives to motivate crowdsourced members for a long time and such low morale is contagious among participants. Reading undesired content and making a decision are demanding tasks, even for professional moderators (Roberts, 2019), let alone crowdsourced workers. Meanwhile, a 'dismiss' decision on reported cases may also disappoint the reporters and discourage them from filing more reports. Without an effective mechanism to mobilize users, it is not easy to facilitate positive and consistent contributions in a system that relies on crowdsourced powers.

Sometimes the concern about the motivations of users is not on the lack, but on the nature. The divergence in the case categories from reported cases, as well as the coordinated reporters, suggest that different motivations from users. Targeting on the verbal abuse cases from a few other users, LDFR were expected to be someone who engaged in personal conflicts or leveraged the reporting function to attack others repeatedly. In contrast, HDFR showed more altruistic behavior, consistently reporting both spam and abusive content to improve the platform, even though spam reports, which made up most of their activity, were largely ignored by the platform. This is ironic because users who voluntarily moderate this community usually get ignored, but aggressive users, who may behave like a troll, instead get rewarded by the platform with more attention and higher acceptance rates. Even though we hypothesize that Weibo may come up with other technical solutions to address overwhelming spam ads, this widespread and unbalanced silence may still hurt the enthusiasm of neighbor watchers.

In summary, our results point to the risks of redundant or merely symbolic participation, as well as the limited motivation among digital jurors. So we suggest that platforms aiming to develop autonomous content moderation systems should facilitate meaningful and healthy

user participation through mechanisms like incentives, or opportunities for deliberation. At the same time, it is essential to implement safeguards, such as oversight mechanisms, to deter coordinated or strategic manipulation from users. Eventually, the success of a community-driven content moderation system lies not only in including users but in designing institutional frameworks that balance autonomy, accountability, and procedural transparency.

Implication and Conclusion

Attaching different goals to Weibo's experiment may lead to different answers to the outcome of this moderation system. However, imitating a democratic governance system online in an authoritarian country is an uncommon practice and a remarkable step in public reasoning and digital governance. This experiment offers a valuable and rare opportunity to view authentic behaviors when people defend themselves with reasoning and persuasion rather than power and settle conflicts with votes rather than orders. Nevertheless, this work still fills the gap of dismissed and ignored parts of content moderation, which largely broadens our understanding of platform decisions and motives with details.

While this paper examines various achievements and failures of Weibo moderation system, we do not answer whether this experiment successfully restrained the undesired behaviors on Weibo. However, the natural experiment in Weibo's moderation setting provides a good opportunity to causally investigate the effects of such moderation model. Future works could investigate the effects of community-driven moderation and practice, diverting attention from English-centric platforms to understudied social media which probably have more abundant lessons to take. The online governance experiments out of the scope of the English world are wild and active and worthy of more studies. We believe that institutional competition and imitation between social media platforms will bring us more potential to build and govern a trustworthy information ecosystem.

Acknowledgments

We wish to express our sincere gratitude to our mentors Will Hobbs, Mor Naaman, and King-Wa Fu; to our cohorts and colleagues who attended our talks; and to the patient reviewers for their insightful comments and suggestions. Additionally, we extend our thanks to the participants at ICA 2023 for their valuable feedback and engaging discussions.

References

Bajpai, T., Asher, D., Goswami, A., and Chandrasekharan, E. (2022). Harmonizing the cacophony with mic: An affordance-aware framework for platform moderation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22.

Bernstein, M., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., and Vargas, G. (2011). 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 50–57.

Bucher, T. and Helmond, A. (2017). The affordances of social media platforms.

Chin, Y. C., Park, A., and Li, K. (2022). A comparative study on false information governance in chinese and american social media platforms. *Policy & Internet*, 14(2):263–283.

Chuai, Y., Tian, H., Pröllochs, N., and Lenzini, G. (2024). Did the roll-out of community notes reduce engagement with misinformation on x/twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–52.

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review*, 36:105374.

DeLisle, J., Goldstein, A., and Yang, G. (2016). *The internet, social media, and a changing China*. University of Pennsylvania Press.

Eldon, E. (2010). Facebook begins testing advanced crowd-sourced content moderation.

Fan, J. and Zhang, A. X. (2020). Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Fang, K. (2022). "rumor-debunking" as a propaganda and censorship strategy in china: The case of the covid-19 outbreak. *Disinformation in the global south*, pages 108–122.

Flew, T. (2021). *Regulating platforms*. John Wiley & Sons.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

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

Gosztonyi, G. (2023). *Censorship from Plato to Social Media: The Complexity of Social Media's Content Regulation and Moderation Practices*, volume 61. Springer Nature.

Habib, H. and Nithyanand, R. (2022). Exploring the magnitude and effects of media influence on reddit moderation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 275–286.

Kiene, C., Monroy-Hernández, A., and Hill, B. M. (2016). Surviving an" eternal september" how an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1152–1156.

Klonick, K. (2019). The facebook oversight board: Creating an independent institution to adjudicate online free expression. *Yale LJ*, 129:2418.

Kou, Y., Gui, X., Zhang, S., and Nardi, B. (2017). Managing disruptive behavior through non-hierarchical governance: Crowdsourcing in league of legends and weibo. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–17.

Kraut, R. E. and Resnick, P. (2012). *Building successful online communities: Evidence-based social design*. Mit Press.

Lei, L., Zou, S., and Wei, Z. (2013). Weibo court: Internet case, netizens judge (in chinese).

Li, L. and Zhou, K. (2024). When content moderation is not about content: How chinese social media platforms moderate content and why it matters. *New Media & Society*, page 14614448241263933.

Li, M. (2023). Promote diligently and censor politely: how sina weibo intervenes in online activism in china. *Information, Communication & Society*, 26(4):730–745.

Li, Y. (2018). Electoral system effects re-examined using the largest vote share variable. *Democratization*, 25(1):58–77.

Liu, L. (2024). *From click to boom: The political economy of e-commerce in China*. Princeton University Press.

MacKinnon, R. (2008). Flatter world and thicker walls? blogs, censorship and civic discourse in china. *Public Choice*, 134(1):31–46.

Matias, J. N. (2016). Going dark: Social factors in collective action against platform operators in the reddit blackout. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 1138–1151.

McMahon, L., Kleinman, Z., and Subramanian, C. (2025). Meta is abandoning the use of independent fact checkers on facebook and instagram. https://www.bbc.com/news/articles/cly74mpy8klo. Accessed 13 July 2025.

Miller, B. (2018). The limits of commercialized censorship in china.

Miller, C. (2022). Elon musk embraces twitter's radical fact-checking experiment.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

Pan, C. A., Yakhmi, S., Iyer, T. P., Strasnick, E., Zhang, A. X., and Bernstein, M. S. (2022). Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–31.

Pröllochs, N. (2022). Community-based fact-checking on twitter's birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 794–805.

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

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

Roberts, S. (2019). *Behind the screen*. Yale University Press.

Scharlach, R., Hallinan, B., and Shifman, L. (2024). Governing principles: Articulating values in social media platform policies. *new media & society*, 26(11):6658–6677.

Seering, J. (2020). Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact*, 3.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Van Dijck, J., Poell, T., and De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford university press.

Wang, R., Zhang, Y., Suk, J., and Levin, S. H. (2024). Empowered or constrained in platform governance? an analysis of twitter users' responses to elon musk's takeover. *Social Media+ Society*, 10(3):20563051241277606.

Zhao, A. and Chen, Z. (2023). Let's report our rivals: how chinese fandoms game content moderation to restrain opposing voices. *Journal of Quantitative Description*.

Zittrain, J. and Edelman, B. (2003). Internet filtering in china. *IEEE Internet Computing*, 7(2):70–77.

The Evolution of Weibo Community-driven Content Moderation System

In this section, we process traced the evolution of Weibo's community-driven content moderation system with policy documents published online by Weibo or the historical documents retrieved from Internet archives.

The early initiators of Weibo's content moderation experiment were very ambitious and aimed to draft a constitution, modeled after the judicial system, for this prominent social media platform (Lei et al., 2013). In the experience of Internet practitioners, self-discipline is an effective management method. Modeled on the real judicial system, the idea of building a 'Weibo court' was quickly adopted, 'because it is easy to operate.'

Thanks to internet archives, we investigated dozens of historical documents published by Weibo regarding its Community Committee system and traced how this experimental system

was built and then reformed. The cornerstone of this system was the Weibo Community Convention first published in 2012, which was rooted in the aspiration for a society governed by laws and the delicate balancing of the need for order and the protection of freedom of speech. This first convention emphasized the rights and norms of users and Weibo being moderated by both the platform and community committee:

In order to build a harmonious and healthy network environment with the rule of law, maintain the order of the Weibo community, and better protect the legitimate rights and interests of users, Weibo and users jointly formulated this Convention.

This convention also set up a credit system as a new penalty system, in which each user starts with a certain amount of credit points, and misbehaviors may deduct user credits and then restrict the user functions. However, later versions of the Community Convention revealed a drastic change in Weibo's understanding of self-orientation and Internet governance since 2017. They gradually downgraded the role of users from a collaborative partner with rights to a more passive and submissive player with obligation and emphasized how Weibo would comply with current laws and regulations to regulate users and expressions online. It details various regulation rules and forbidden ranges of speech, which makes this convention less like a constitutional document but a code of conduct for users.

Next, along with the community convention are the policy documents to establish a community committee system. The first document explicitly states that the Weibo authority would obey the verdict of committee members and act accordingly. However, Weibo's expert committee was de facto suspended silently in October 2014 and no longer judged misinformation cases. We do not know the exact reasons behind this shutdown as it requires abundant insider knowledge, but we speculate this change was associated with the increased regulation of Internet in China.

In early 2021, Weibo announced that it brought this community committee system back to life (but most new cases were no longer public on the original website). By positioning user-based groups as self-disciplined organizations in the new system, Weibo again degraded the role of users in its governance model and focused on the management of these volunteers. Weibo emphasized that '(the committee should) accept the unified management and instructions from

Weibo'. The regular committee keeps its original function of handling user disputes with votes, but the expert committee becomes a group of invited experts with advisory roles on content moderation. A new role of community volunteers was introduced to detect inappropriate content and the Weibo Community Committee system, a self-governance model based on the jury system, has now become an almost voluntary group serving Weibo.

Last, Weibo set clear rules in their regulation documents, which detailed specific criteria for inappropriate content and how moderation works. The first regulation document functions as a manual for reporting inappropriate content and handling policy violations, and the later versions added more terms about harmful and illegal content, misinformation, spam content, and user disputes. It elaborated on how to handle appealing cases and aggravated punishments. The latest regulation document in 2022 stressed again how Weibo would comply with relevant laws and regulations and is committed to strictly moderating all content.

Criteria to Join User Committee (May 2012)

- Must be a personal user who has completed real-name identity verification.
- Must be at least 18 years old.
- Account must be registered for more than 180 days, and the user must have posted more than 100 Weibo posts.
- Must have more than 50 followers.
- Must have logged in on at least 15 days within the most recent calendar month.
- Selection is based on the order of application.

Spam Ads Cases as Mistrials

A large part (56.33%) of our case data from the content management center is mistrial cases that still surprisingly remained unanswered. 99.73% of these mistrial cases, or more than six million cases we collected, are categorized as 'spam ads'. Most of these cases only contained a simple statement: 'Spam Ads.' There are still more than 65 thousand spam ads that received a verdict from the Weibo platform or the jury, and most of them were reviewed between 2014

and 2017, and a few cases before or after. Meanwhile, only 302 reports about spam ads were dismissed by Weibo. This fact suggests that ignoring spam ad cases is a systematic trend in Weibo moderation rather than a technical error. However, it is uncertain if Weibo chose to deal with this problem via another channel other than the content management center.

Contentious Cases and Opinion Conflicts

Surprisingly, most jurors have the experience of being a true minority, regardless of which side they took. We selected the cases with at least five jurors and eighty percent of all votes were on one side, either for or against the plaintiff. In these one-sided cases, there are 33,737 unique users, which is 71.4% of all jurors who reviewed at least five cases during their careers, ever voted against the majority tides. 23,819 of them even had minority votes more than once. Only less than seven percent of all jurors with at least five votes have no experience being a minority. The fact that most people cannot always be the majority even on the social media moderation voting resonates with what Li (2018) discussed as an unstable majority in the electoral systems, where people have different interests and opinions on diverse issues and cannot always keep a majority alliance. This result suggests that majority and minority identities are fluid in vote recordings, revealing moderation cases' diversity.

Given this unstable majority pattern, we examined the consensus and conflicts of public opinions revealed by jury votes. Contentious cases are defined by authors as cases where the vote difference between the plaintiff and the defendant does not exceed 20% of the total votes, and all other cases are unanimous. Then, we identified 83,456 unanimous cases under this definition where all jurors have a consensus on the reported content. 65,734 unanimous cases fell into the verbal abuse category (39.30% out of the total number of jury-verdict verbal abuse cases), 16,337 in harassment (18.99%), 1,248 in plagiarism (24.82%), and 132 in misinformation (41.90%). There are only 101 (0.12%) cases where jurors unanimously voted for the defendants.

We also identified 15,071 contentious cases, 5.83% of all jury-verdict cases, have been identified as contentious cases, with 8,223 in harassment (9.56%), 6,448 in verbal abuse (3.86%), and 36 in misinformation (11.43%). Proportionally speaking, harassment cases were more likely to have conflicting opinions, as the vote records show.

Supplementary Materials: Figures and Tables

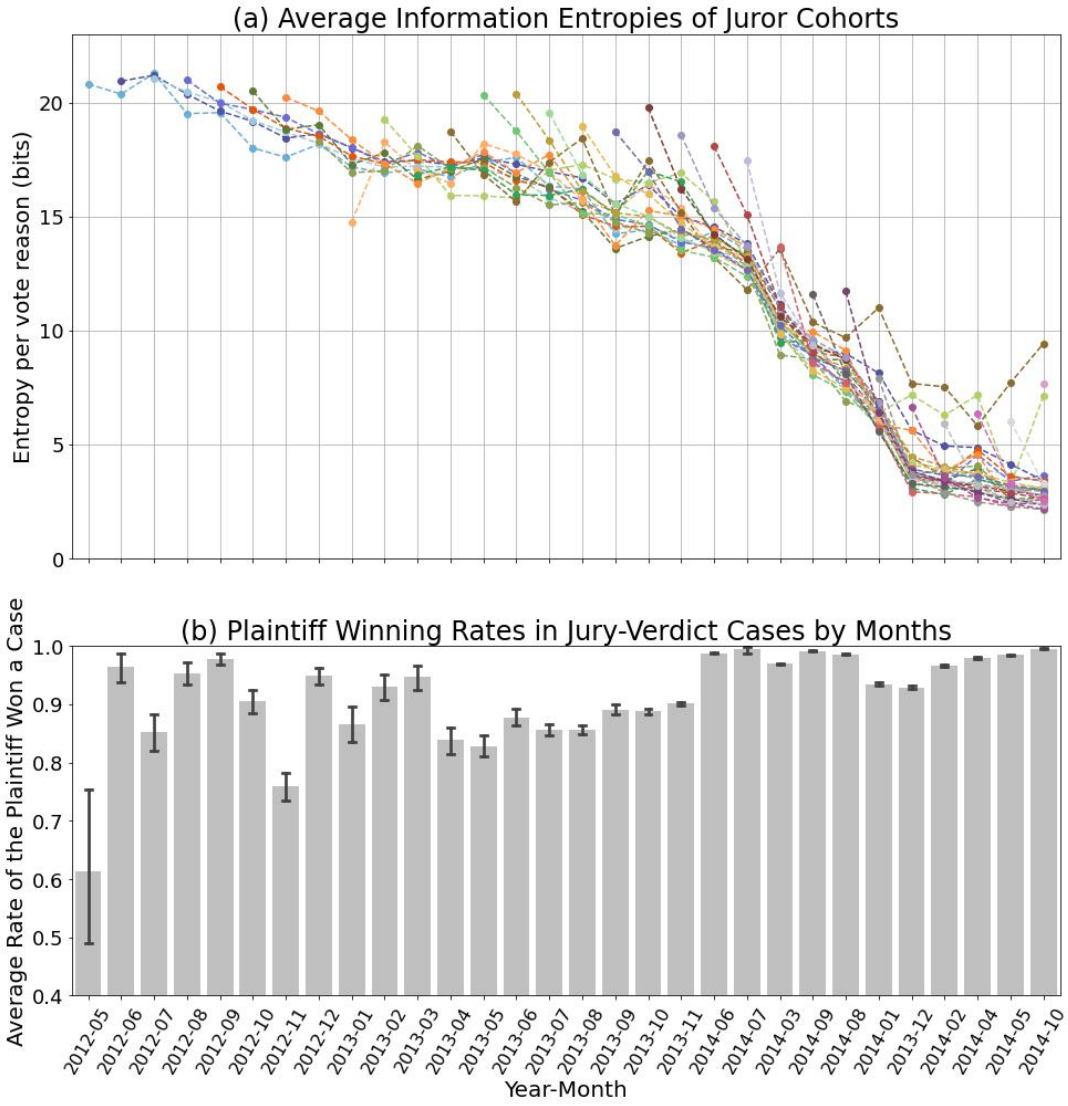

Figure 3. (a) The change of information entropy per vote reason in voter cohorts. Each line represents a voter cohort who cast their first votes in the same month and therefore has a different starting point. Some cohorts/lines may have similar colors but they started at different times and could be easily separated in their beginning months. (b) The winning rates of the plaintiffs by time, and the error bars indicate the 95% confidence intervals

Table 7: Nine Types of Penalties on Weibo

Label	**Description**
(A)	Account Temporary Suspension
(B)	Banned from being followed
(C)	Weibo Credit Deduction
(D)	Case Dismissal
(E)	Not a Violation
(F)	Content Flagged
(G)	Content Deletion
(H)	Account Permanent Cancel
(I)	User Information Deletion

Table 8: Percentage of credit points deducted in four categories. While we cannot validate whether such a penalty was consistent with Weibo's regulation at that time, web archives of Weibo policies indicate that user conflicts usually receive lighter penalties than cases with public interests. Account suspension data tells a similar story about misinformation being a big concern for Weibo. 59.21% of violators in misinformation cases got fifteen days suspensions, and the rest got seven days suspensions. But almost all violators in verbal abuse and plagiarism cases only got seven days of suspension.

Points	**Abuse**	**Plagiarism**	**Harassment**	**Misinfo**
1	76.84%	55.84%	99.94%	0.00%
2	23.12%	33.86%	0.06%	13.58%
3	0.01%	0.00%	0.00%	0.00%
5	0.03%	10.25%	0.00%	35.47%
10	0.00%	0.05%	0.00%	50.57%
20	0.00%	0.00%	0.00%	0.38%

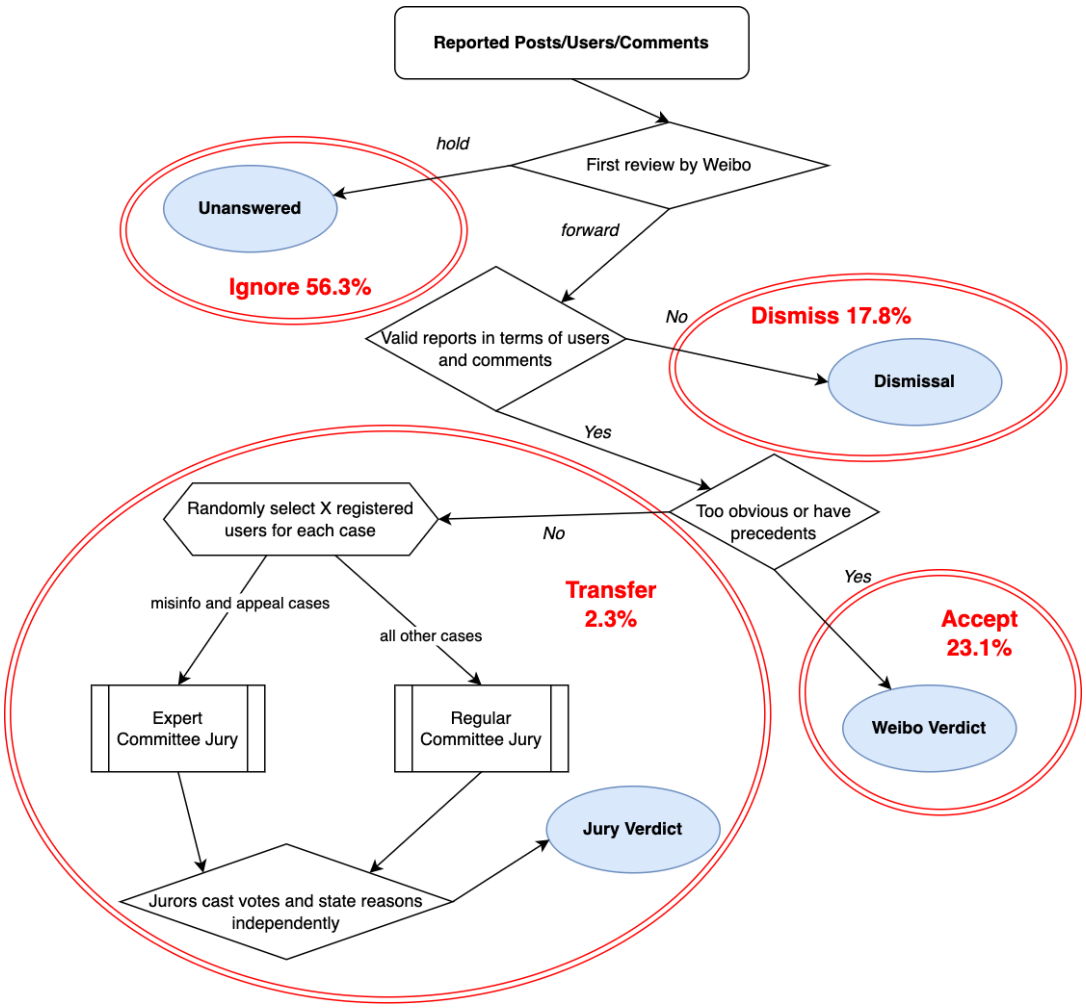

Figure 4. How a user-reported moderation case is processed on Weibo

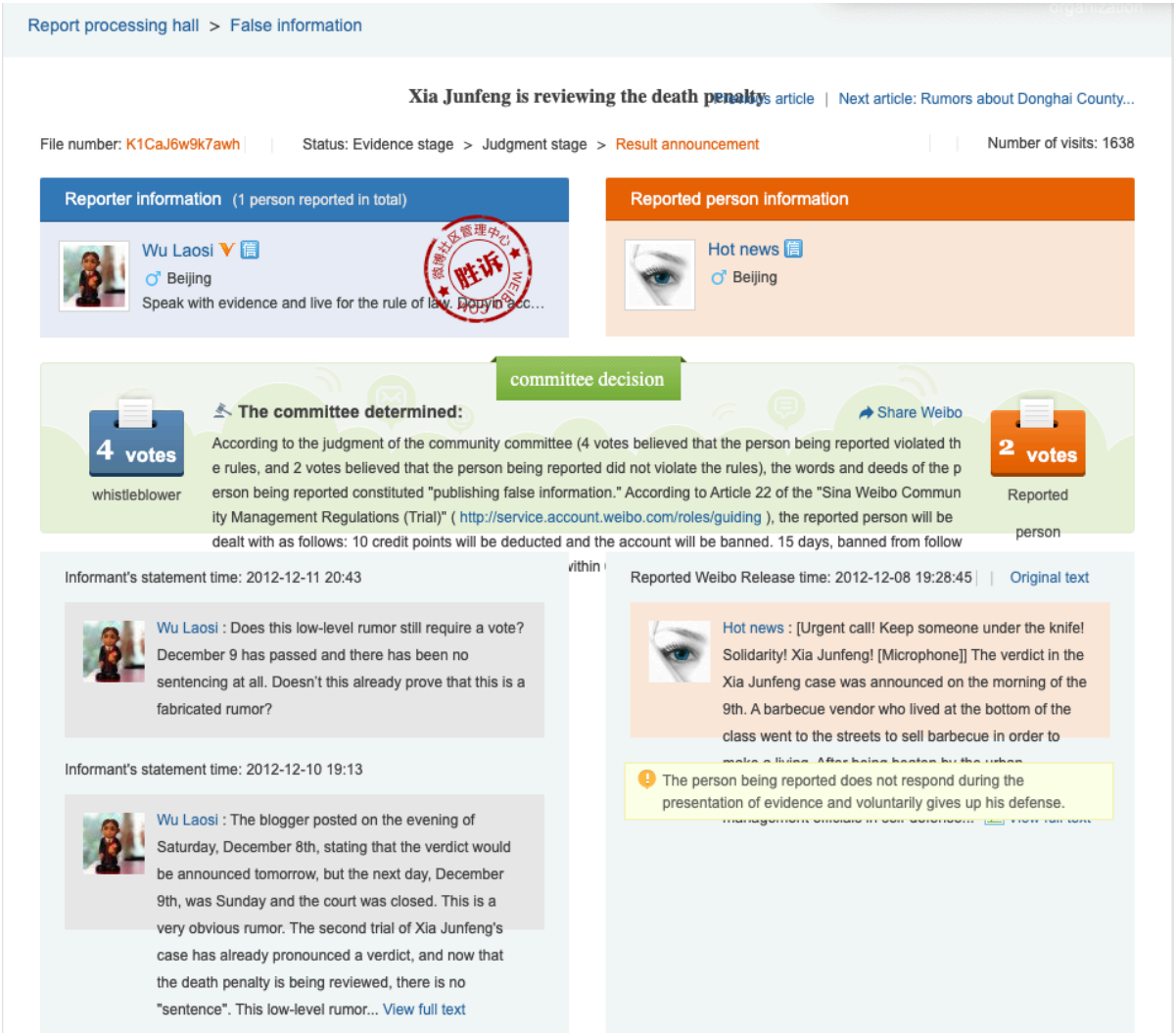

Figure 5. A screenshot of a moderation case on Weibo, auto-translated by Google

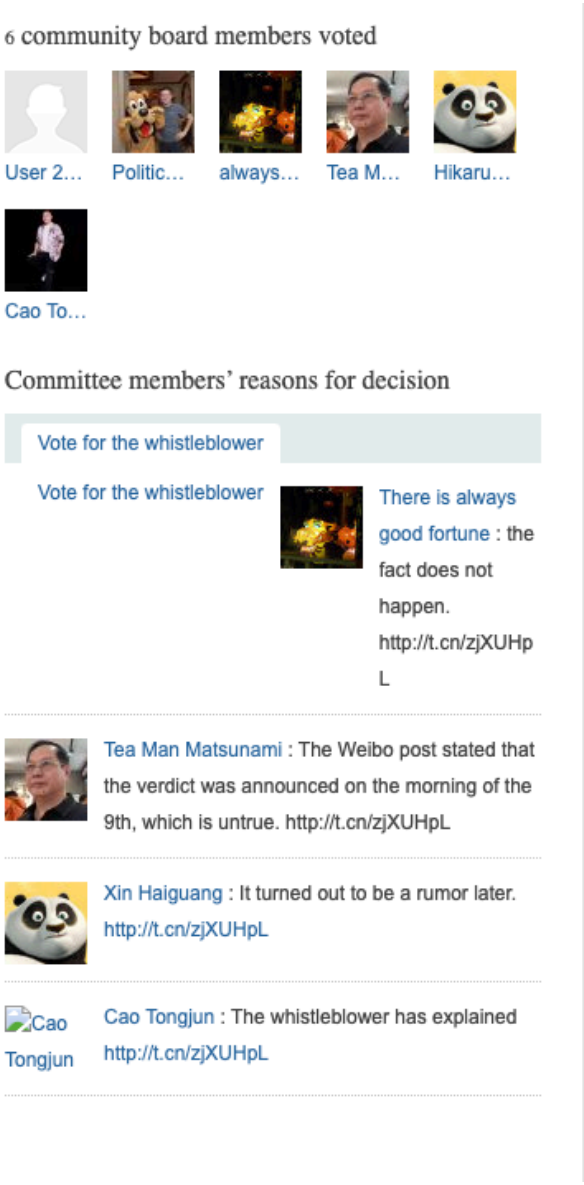

Figure 6. A screenshot of jury votes on the previous moderation case, auto-translated by Google

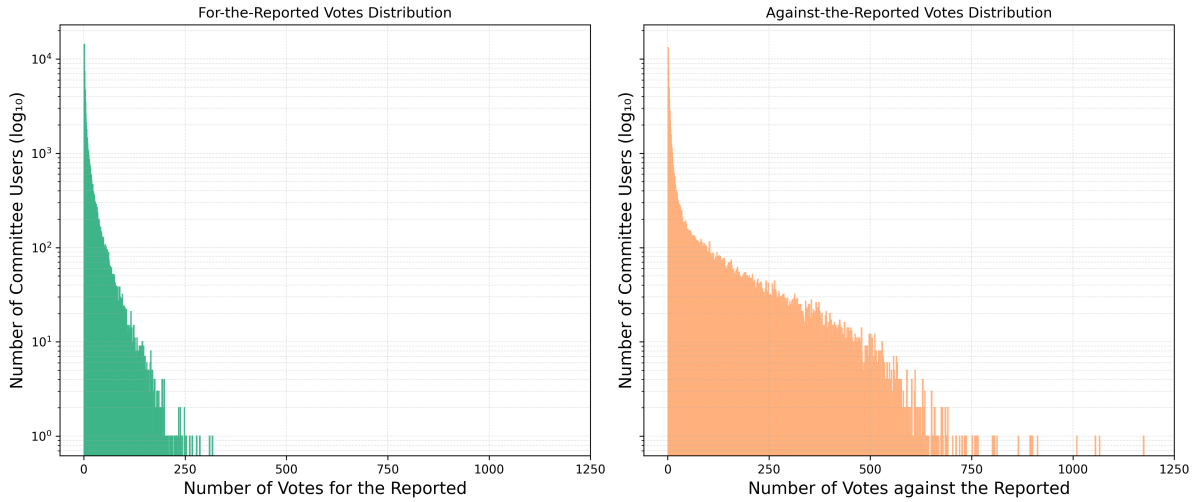

Figure 7. Voting activity of Committee Users. Each bar represents how many users (y-axis) have casted how many numbers of votes (x-axis). The left figure with the mint bars represents the votes that supported the reported users or posts. The right figure with the peach bars represents the votes that were against the reported users or posts. Most members contribute only a handful of "for" votes, while "against" votes are spread across a wider, heavier-tailed range.

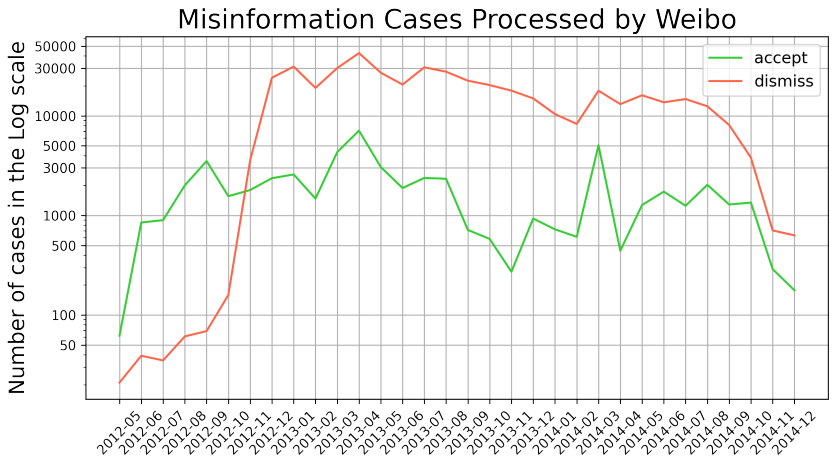

Figure 8. A line plot of misinformation cases being processed by Weibo between 2012 and 2014. Since there was no substantial change in the reported contents of misinformation cases, the sharp increase of dismissal rate likely came from the governance strategy change from Weibo towards misinformation disputes.

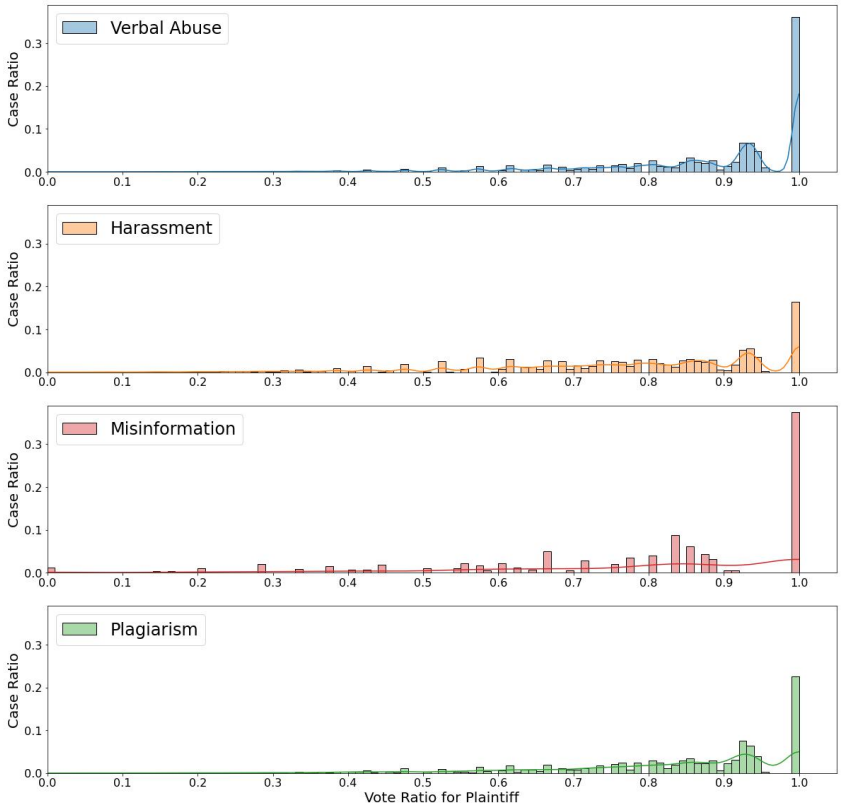

Figure 9. Voting Pattern by Case. The x-axis represents the plaintiffs' vote ratio, while the y-axis represents the case ratio for each category. The vote ratio refers to the proportion of votes received by the plaintiff in a given case, while the case ratio indicates the proportion of cases that fall within a specific vote ratio range.

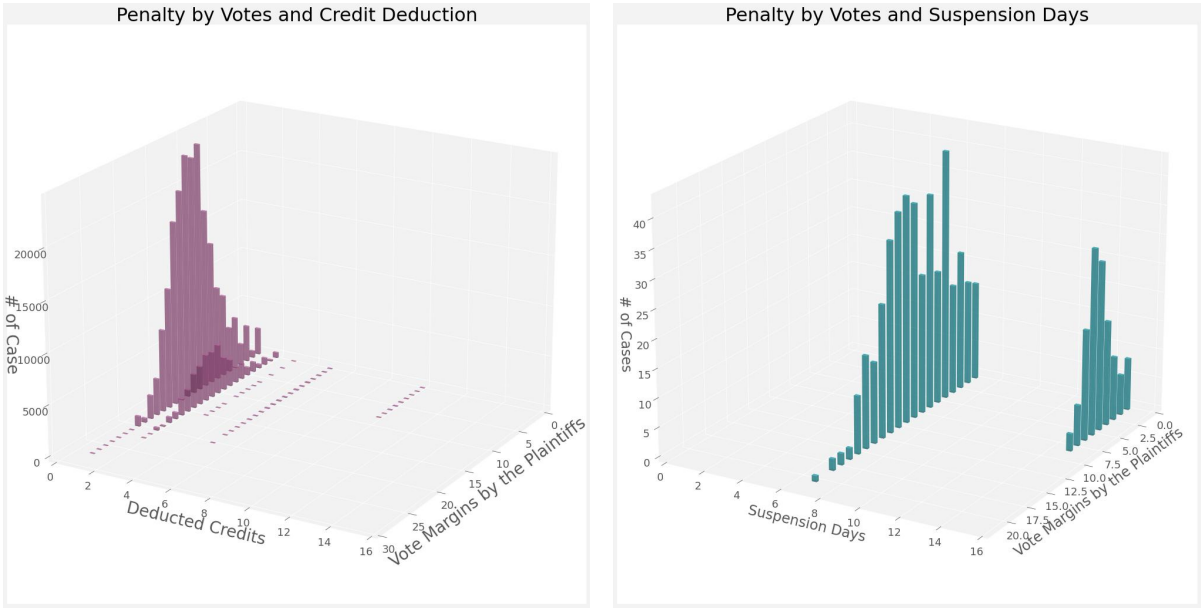

Figure 10. Left: Vote and Credit Point Deduction. Right: Vote and Account Suspension. These figures describe the distributions of the victory margin of the plaintiffs and the penalties the defendants received. There is no cluster of data points of high penalties and big vote margins (which is supposed to appear in the upper right corner of the bottom plane). We use the Shapiro–Wilk test—not to assume normality in penalties, but as a diagnostic tool—to examine the distributional patterns of vote margins at specific penalty levels. In the left plot, vote margins at credit deduction levels 1, 2, and 5 follow normal distributions ($p<0.05$); in the right plot, vote margins at suspension lengths of 7 and 15 days have p-values of 0.16 and 0.27, respectively. This analysis supports a conservative conclusion that the public distaste revealed in votes has no clear relation with the sentencing levels.

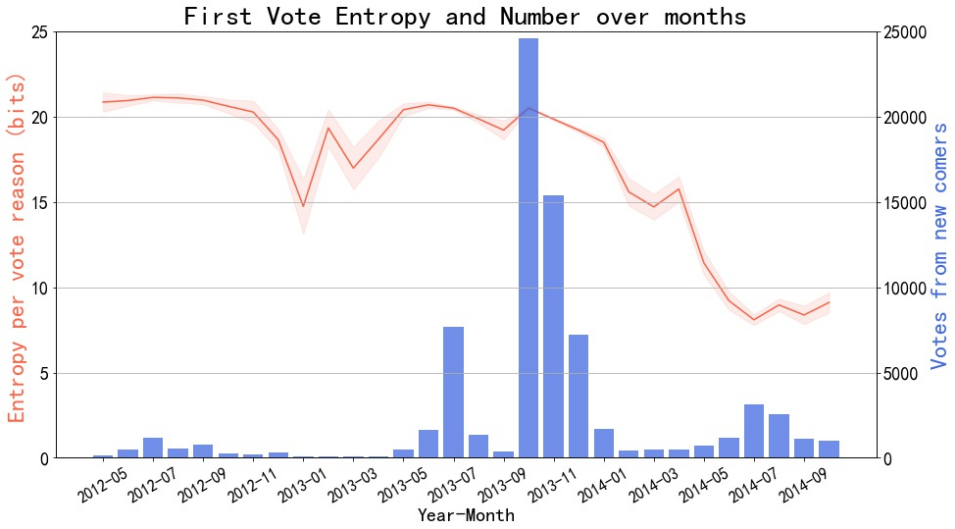

Figure 11. Information entropy (left in orange, as the line) and numbers (right in blue, as bars) of first votes over months between 2012 to 2014. The shadow around the line indicates the confidence intervals. Except for an up-and-down during late 2012 and early 2013, the entropy trend of first votes is generally consistent with the overall entropy of all cohorts. This result suggests that with decreased engagement, newcomers initially adapted to a low-motivated atmosphere and started with a standard and uninformative vote reason.