Journal of Quantitative Description: Digital Media 3(2023), 1 – 35 10.51685/jqd.2023.006

Let's report our rivals: how Chinese fandoms game content moderation to restrain opposing voices

Andy Zhao

Cornell University, USA

Zhaodi Chen

Indiana University, USA

While crowdsourcing approaches in content moderation systems increase the governance capacity of social media, they also offer a loophole for malicious users to report and restrict disliked content massively. To fill the knowledge gap about large-scale, bottom-up attempts at restraining online expressions, we focus on a type of public and institutionalized mass reporting: anti-smear (反黑) campaigns within Chinese online fandom communities, where fans coordinate together and collectively report content they perceive as inappropriate. Based on detailed data from more than two hundred anti-smear groups collected from Weibo and interviews with active participants, our paper examines the motives and dynamics of anti-smear campaigns, the coordination strategies used to game the content moderation system, and the diffusion of anti-smear culture among fandom networks. We argue that anti-smear is essentially a practice of information control and reflects an intolerant mindset of social media users towards dissidents. This paper also points out the vulnerability of community-based content moderation systems to be weaponized in a polarized age, which brings great challenges to platform governance.

Keywords: fandom, mass reporting, content moderation, social media

Andy Zhao (corresponding author): dz352@cornell.edu
Zhaodi Chen: zc20@iu.edu
Authors contributed equally.
Date submitted: 2022-05-27

Introduction

Due to very high volumes of potentially harmful content, social media companies usually embed crowdsourcing characteristics in their discovery mechanism of content moderation. Users may report any offensive content to the platform with just a few clicks. This mechanism can help social media authorities quickly identify and moderate countless problematic pieces of information every day. However, this power is a double-edged sword, as the opportunity to curb unappealing content is appealing. Users may abuse their power to report any content in their own interests, and malicious agents may even coordinate en masse to falsify a mass dissatisfaction with a large number of reports in a short time, thus attempting to restrain a certain type of information on social media. For example, Vietnam dissidents were repressed by mass reporting on Facebook (Gleicher, 2021), and Russian accounts also attempted to silence Ukrainians with this adversarial strategy in 2022 (Nimmo and Agranovich, 2022).

In Chinese, *Fan Hei* (Anti-Smear thereafter) is a coordinated online campaign to weaponize the reporting system and attempt to impede negative expressions on social media, but it is more public and institutionalized than general coordinated inauthentic behaviors. In this study, we focus on the online anti-smear groups in fandom communities on Weibo, one of the largest Twitter-like social media platforms in China. These anti-smear groups are fan-driven organizations on social media where coordinators regularly collect and publish links to content they perceive to reflect poorly on their idols, and then mobilize other fans to report these targets together.

We use anti-smear in the Chinese fandom community as the lens to understand online mass reporting activities for two reasons. First, even though mass reporting is not unique to fandom (Crawford and Gillespie, 2016), fandom communities increasingly have more leverage in shaping the tide of online political activism during recent years (Dodson, 2020). Second, contrary to covert and occasional mass reporting behaviors, anti-smear campaigns within Chinese fandom communities are public and routinized, which provides us with a unique opportunity to observe the changes in size, scope, and strategies of mass reporting. However, despite the growing size and impact of fandom communities and the public attention on fandom anti-smear(Tan, 2020; LaiFu, 2020; Jian, 2021), we still lack an empirical understanding of this large-scale social phenomenon and how it games the content moderation system on social media.

Our paper aims to fill the knowledge gap about anti-smear with interviews with participants and quantitative analysis. By interviewing seven anti-smear participants, we suggest that fans join anti-smear to avoid being accused of free-riding and maintain their status within the community. Institutionalized anti-smear accounts play an essential role in guiding newcomers and sustaining long-term actions.

With data collected from anti-smear accounts on Weibo, we notice that anti-smear in China originated in 2015 from peripheral celebrities and has diffused to more fan communities since 2019. It expanded rapidly after the outbreak of COVID-19 pandemic. We suggest that mass reporting behaviors in Chinese fandom may render more than seventy million reports on Weibo and result in millions of posts being suspended in the most active months of anti-smear.

We then indicate that anti-smear groups tend to lower the action costs by providing the most simplified instructions and asking fans to use the most generic reasons when reporting. We also suggest that anti-smear groups pursue greater engagement by requiring fan participants to "check in" on daily tasks and more participants are associated with a higher chance of successfully suspending the targeted content.

Background

The phenomenon of anti-smear connects two strands of studies: Internet reporting and online fandom communities. Before introducing our data, we present a brief review of the literature on online reporting and weaponized content moderation and discuss how online fandom communities, with their increasing size and impact, may affect content moderation via mass reporting.

Online Reporting and Weaponized Content Moderation

Online content moderation is one of the central processes through which public discourse is negotiated among different actors including users, social media platforms, and governmental regulators. In many content moderation systems, the most common and feasible way for ordinary internet users to moderate content is online reporting (Buni and Chemaly, 2016; Seering, 2020).

While the specific operation of moderation varies from case to case, most platforms have features that support user reporting, such as the "flag" feature that allows users to report content that they believe violates rules or norms (Crawford and Gillespie, 2016). From the platform's perspective, letting users report inappropriate content greatly reduces the burden of platform governance and justifies the removal of content. From a user's perspective, reporting is a way to directly participate in the content moderation process by bringing issues to the moderators' attention.

However, reporting features may be utilized by users in more tactical and even abusive ways (Crawford and Gillespie, 2016; Fiore-Silfvast, 2012). It is generally difficult to account for the many and often complex reasons why people might choose to report. For example, a user may report another user as a form of personal attack rather than genuinely being offended. As a result, the reporting mechanism offers a loophole for users to maliciously report disliked content and restrict the dissemination of the reported targets' content. The potentially abusive use of reporting undermines its value as a gauge of what the community considers as "proper" content. Moreover, reporting undesirable content can be a collective act.

Recent years have witnessed examples of organized, strategic flagging occurring in a wide range of contexts, especially in online debates on contentious topics such as racial and gender-based contention (Matamoros-Fernández, 2017). Despite the evidence showing the widespread occurrence of organized reporting, more studies are needed to examine the operation of these collective actions, the diffusion of organized reporting as a strategy for users to handle undesirable content, and the motivations of users who participate in organized reporting.

Reporting in the Context of China

Mass reporting was perceived as an effective means of collaborative governance to combat corruption in China for a long time (Rosenbloom and Gong, 2013). Referring to mass reporting as "whistleblowing," Gong (2000) tends to emphasize how such a system empowers ordinary people to fight against malfeasance and misfeasance. The dominant interpretation of reporting as a mechanism mediating between the party-state and the society is justifiable given the particular political environment and the history of mass reporting in China. However, this perspective may overlook the other logic and motivations that may drive mass reporting, especially when it comes to the context of online reporting.

In the internet age, reporting has become an institutionalized strategy to take advantage of mass power to suppress dissenting content (Jiang, 2021; Staff, 2022). By promoting the "official version of morality or ethics," the state can encourage self-purification and self-discipline to strengthen the legitimacy of the state in a populist way. Huang (2021) has shown a tendency of increasing online reporting cases in Chinese cyberspace, and the social media site, Weibo, has become the most commonly used platform for reporting. The switch of reporting arena from state-backed institutions to commercial platforms suggests a need for more studies on spontaneous online mass reporting that are not directly promoted by the state.

Online Fandom Communities

In this study, we pay special attention to the mass reporting efforts of Chinese online fandom communities. Existing literature recognizes fans as active producers and consumers of online content in the digital age and argues for fandom communities' role in constituting an online participatory culture (Jenkins, 2006; Earl and Kimport, 2009; Kahne et al., 2015). Moreover, online fandom communities allow fans to forge common identities, gain media and digital literacy and develop hierarchies and organizational structures (Zhang, 2016). As a result, they provide fertile soil for collaborative and collective actions. For example, the high frequency of debates between competing online fandom groups requires fans to organize themselves in order to win online battles. In some cases, these collective actions may even go beyond the initial purpose of the fandom community. Park et al. (2021) shows that the BTS fan community on Twitter successfully organized the #MatchAMillion campaign to raise money for the Black Lives Matter movement. Similarly, scholars of the Chinese Internet point out that fan groups on Weibo have served as active participants in online nationalistic activism (Liu, 2019; Shan and Chen, 2021). The organized battles fans had within their communities on a daily basis were the reason they were able to mobilize and organize immediately on occasions of nationalistic activism (Wu et al., 2019).

With the organizing and mobilizing capacity developed from everyday fandom activities, online fandom communities can also effectively intervene in the content moderation process. Recently, the anti-smear campaigns within fan communities have attracted much public attention in China through media coverage (Tan, 2020; LaiFu, 2020; Jian, 2021), as well as a considerable amount of scholarly attention. Mostly based on interviews and ethnographic observations of one or a few fan groups, these studies provide important insights into the micro-level operations of anti-smear campaigns (Qin and Chen, 2021; Zhang and Hu, 2021). However, little systematic and quantitative analysis has been provided regarding the size, scale, motivations, strategies, and diffusion of anti-smear campaigns on the Chinese Internet. Our study intends to address these limitations by interviewing participants and examining all posts and relevant data from over two hundred anti-smear accounts.

Data

Anti-smear data collection

Anti-smear campaigns are coordinated by the anti-smear accounts of fan groups. These anti-smear accounts are initiated and maintained voluntarily by fans and are independent from the official accounts of the celebrities.[1] As shown in Figure 1, a typical anti-smear campaign involves four steps. First, once ordinary fans encounter undesirable content, they collect and send the content to the anti-smear account. Second, the anti-smear account, often administered by veteran fans, organizes the content and decides what strategies to use

[1] We do not exclude the possibility that some celebrities or agencies may covertly cooperate with or even sponsor these anti-smear accounts. However, our qualitative and quantitative investigation provides no substantial evidence of this hypothesis.

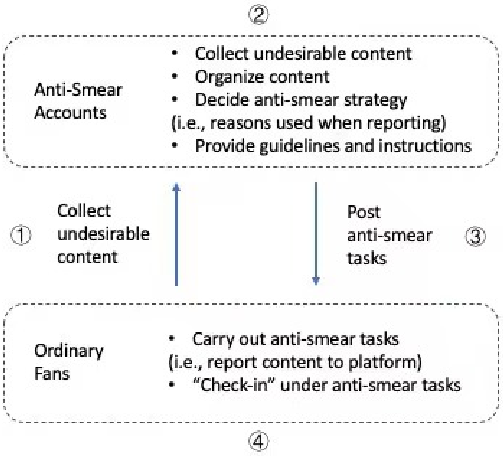
②
Anti-Smear Accounts
• Collect undesirable content
• Organize content
• Decide anti-smear strategy
(i.e., reasons used when reporting)
• Provide guidelines and instructions
① Collect undesirable content
Post anti-smear tasks ③
Ordinary Fans
• Carry out anti-smear tasks
(i.e., report content to platform)
• "Check-in" under anti-smear tasks
④

Figure 1. Illustration of the anti-smear process on Weibo

Note. Fans and anti-smear accounts play different roles and collaborate to report undesirable content massively.

when reporting it. Third, the anti-smear account posts anti-smear tasks on its homepage, with instructions on how to report the content. Fourth, ordinary fans collectively report the content based on the anti-smear account's instructions. After reporting the content, ordinary fans often "check-in" by leaving a comment under the anti-smear task. Due to the public nature of anti-smear activities, social media data about user behaviors and reporting results can be used to understand the dynamics of this Internet phenomenon.

The first step of our data collection is to identify a group of anti-smear accounts that can effectively capture anti-smear activities within the Weibo fandom community. To begin with, we created a list of more than six hundred Chinese celebrities who appeared on two celebrity popularity ranking tables between 2014 and 2021. One such table is the Star Table, which was operated by Sina Weibo and calculated with their social media data. Another is the Internet Powerstar table maintained by a private business organization, which has a long history and more detailed classification. [2]

[2] Both tables were either taken off or discontinued as one of the influential and widely recognized ranking tables during a rectification campaign on entertainment industries and fandom activities in August 2021.

After finalizing the celebrity candidate list, we manually searched for the anti-smear account for each celebrity. We kept anti-smear accounts with at least 200 posts and 1,000 followers and dropped unpopular or inactive accounts. Eventually, we identified 230 popular and active anti-smear accounts. Then, we collected all public account information and public post information of these accounts, including send time, post content like text and links, and post feedback (e.g., like, retweet, comment) from their creation dates to July 2021.

Given the rich text in anti-smear posts, we can identify the report links and corresponding targets. Therefore, we also used the accessible status of reported targets to infer the outcomes of anti-smear reports. We randomly sampled 100 reported links for each day from all anti-smear accounts since 2016 and checked their current statuses by attempting to access them via a crawler. If targeted users or posts were suspended, we would treat the corresponding reporting as effective. This approach is not perfect because the users or posts might have been removed later for other reasons. However, we assumed that the bias is random among all targets and this strategy would successfully approximate the trend and the scale of anti-smear outcomes.

In addition to anti-smear data, we also identified 224 out of 230 celebrities who have an active anti-smear fan group as well as an active personal account on Weibo. Then, for each celebrity, we randomly sampled 50 fans who had at least 100 followings and 100 followers. For each fan account, we then randomly looped their following pages ten times and collected as many followings as possible (the number is usually between one hundred to two hundred). By cross-matching the user ID of celebrities in the following lists, these data showed the relative popularity of celebrities among the selected fans and how fans perceived the similarities between the two celebrities. Meanwhile, we also retrieved 100 days' search index on Baidu (Chinese search engine) before their anti-smear groups used the anti-smear language for the first time (this part will be explained later). This data reflected the Internet popularity of a celebrity at the moment fans started to engage in anti-smear.

Coding

Given the heterogeneity of celebrities, we manually coded the celebrities in our sample into different categories to see if the scale and strategies of anti-smear campaigns varied by celebrities' characteristics. Specifically, we considered three dimensions that may have an impact on fans' anti-smear activities: the celebrity's gender, work style, and character. By work style, we refer to the celebrity's status as working solo ("Solo"), working in a specific boy/girl group ("In-group"), or as previously working in a group but is currently performing individually ("Grouped").

By celebrity character, we asked whether the celebrity is an "Ai-Dou" or not. "Ai-Dou," which is a transliteration of "idol", specifically refers to a category of celebrities that are mostly observed in the entertainment industry in East Asian countries such as China, Japan, and South Korea. Different from professional actors, singers, or comedians, the work of an "Ai-Dou" usually involves a combination of dancing, singing, and possibly acting and hosting TV shows. They are typically young, good-looking, and have huge fan bases. Members of pop groups like BTS and Blackpink are examples of the so-called "Ai-Dou".

Interviews

We also interviewed seven fans who have engaged in an anti-smear activity to supplement our analysis of online data. We recruited our informants using snowball sampling. As members of anti-smear groups tend to be highly cautious about an inquiry from outsiders, we selected our seed interviewee from one author's personal connections. In this way, we are able to quickly establish trust and rapport with interviewees. While we are not aiming for a representative sample, we selected interviewees from different fan groups to diversify our information sources. All interviewees in our sample share similar demographic characteristics: young, female, urban, middle-class, and Internet-savvy. These characteristics are consistent with previous scholarly and media portraits of the Weibo fan population. All interviews were conducted in 2020. The length of each interview was approximately 1-1.5 hours.

Result

Interviews with Participants

We start our analysis with a qualitative interrogation into fans' interpretation of anti-smear activities and aim to explore how ordinary fans understand anti-smear and why they engage in collective efforts of reporting. In general, our results clarify the underlying logic of anti-smear and the motivations to participate in such activities. They also point out the importance of fan communities in socializing fans and sustaining fans' persistent participation in anti-smear. Finally, they suggest that community-level, institutional anti-smear accounts are the key site for the initiation and operation of anti-smear campaigns.

As fans always expect the success of their idols and sometimes take it as their own responsibility in the context of the Chinese entertainment market, fan communities have developed different measures to increase the popularity and prove the commercial values of their idols to entertainment companies, producers, and other stakeholders. Such measures include, but are not limited to, collectively purchasing the celebrity's products (e.g. albums, concert tickets), boosting online video view numbers, and managing the celebrity's online image and volume. As one interviewee commented:

> *"In the old times, we bought albums or went to the signing events together. Celebrities' sales statistics were generated by actual money. Nowadays, the statistics are also decided by the online traffic brought by celebrities. When a TV program or a brand is looking for business partners, the first thing they will look at is the performance of fans – whether we are active enough - of course, they know this celebrity may have no real talents, but the data and the volume created by fans are real. (I02)"*

This excerpt shows why the management of online information has become one of the most important activities in Chinese online fan communities. Fans' practices of information manipulation, or in their words *"making data,"* consists of two general aspects that are similar to propaganda and censorship: promoting positive content about their idols and stifling criticism or negative comments. While fans do not interpret or explicitly describe their practices as mirroring propaganda and censorship, they recognize the nature of *"making data"* as a struggle over the power of dominating the online discourse, a zero-sum game of public attention between positive and negative content. An interviewee explained how the logic of *"making data"* is embedded in the nature of Weibo as an open public sphere:

> *"There were boundaries between different sub-forums on Tieba and Douban (both are Reddit-like forums). You only join this sub-forum if you are interested in the same thing as me. But Weibo has no such boundaries: groups have to compete with each other in public spaces, for example, in the comment sections under entertainment accounts and commercial accounts. When a post says something about a celebrity or releases a rumor, fans have to occupy its comment section. So when ordinary users see this post, they can only find positive things about the celebrity. (I07)"*

Therefore, we can understand the underlying logic of anti-smear as essentially a strategy to stifle criticism and negative content. By collectively reporting undesired content to the platform, fans are able to intervene in the content moderation process to achieve their own goals of dominating the online discourse and maintaining their idols' online image.

However, sustaining the long-term participation of fans in anti-smear requires making a substantial commitment and socializing with other fans. Fan communities play a central role in socializing new fans to the norms of *"how to be a good fan"*. On one hand, formal and informal networks of fans allow new fans to learn about the importance of anti-smear and help familiarize them with anti-smear languages and duties. On the other hand, fans who neither purchase idol-related products nor participate in daily routines would be blamed for free-riding since they enjoy the pleasures provided by their idols and other fans for free. In other words, voluntary participation in anti-smear becomes a norm within the fandom community to the point where non-compliers would face moral judgments from their peers. One interviewee explained the pressure of participating in anti-smear as follows, *"I usually do not even look at what is in the anti-smear tasks...And I only do the tasks when they [administrators] are going to kick free-riders out of the chat rooms. (I02)"*

The mention of *"chat rooms"* in this excerpt shows that fans not only face moral pressures but also actual punishment for not participating. In fact, fan communities have developed an organizational structure to ensure fans' persistent participation in anti-smear. Many fan communities established their institutional anti-smear accounts and numerous anti-smear chat rooms on Weibo. Usually, the anti-smear account is the core agent that initiates and coordinates anti-smear activities on a daily basis, while the chat rooms are the vessels guaranteeing the fans' participation in every day's anti-smear tasks. This organizational structure allows fan communities to monitor the participation and contribution of their members.

Moreover, the efforts individuals have put into anti-smear became an important indicator in deciding their hierarchy within the community: fans who are more active in participating in anti-smear tasks will reach a higher rank. Most interviewees mentioned that they have joined one or more chat rooms for anti-smear. They are required to engage in anti-smear tasks on a regular basis; otherwise, they will be *"kicked out"* of the room. Higher-ranked fans will be prioritized for opportunities such as attending signing events and getting early bird tickets for concerts. As one interviewee explained, *"they [fan community] have a record of your anti-smear participation. Some events are only open to people who have participated enough and reached a certain level in the record. (I02)"* This hierarchical system has connected anti-smear with the distribution of resources within the community, thus creating persisting incentives for fans to participate in anti-smear activities.

One consequence of this system is that fans care less and less about the content they report and instead focus only on fulfilling these tasks. All interviewees acknowledged that they began to *"not care much about the content"* they reported after they had participated in anti-smear for a long time. With clear instructions in each task, fans can complete their missions without actually reading and evaluating the content for themselves. Anti-smear, as a result, becomes mainly a weapon for the fan community to achieve its goal - the domination of online discourse:

> *"The things fans report may not be talks that are actually mean to the idol, they can just be someone with some objective criticism. Fans organize as a whole to get rid of these contents because they want to silence this criticism. It's a struggle over the power of discourse. But some anti-smear practices are*

> *understandable, like reporting those contents that curse your family members or attack you personally. Overall, I think anti-smear is necessary. (I07)"*

As shown in this excerpt, fans sometimes are able to distinguish between content that is genuinely harmful (e.g. trolling) and those simply expressing objective opinions or criticism. However, they chose to report these remarks regardless, as the purpose of anti-smear is to dominate the public discourse.

Scale of Anti-Smear Campaign

The prevalence of anti-smear campaigns varies across the fan communities of different types of celebrities. In general, junior celebrities who are less established in the entertainment industry, those who belong to, or used to belong to, boy/girl groups, and those from Mainland China are more likely to have an anti-smear account. We found that 75.2% of the celebrities who attained fame after 2015 have an anti-smear account, compared to 27.4% of those who were already famous prior to 2015. Among all celebrities in groups and those who used to be in groups, 70% have an anti-smear account, while only 27.8% of the solo stars do. Also, anti-smear accounts are found for 53.8% of the celebrities who come from Mainland China, while only 9.2% of outside-Mainland stars have an anti-smear account.

Table 1: Anti-smear accounts summary

	Celebrity Gender		Work Style			Character	
	M	F	Solo	In-group	Grouped	not Ai-Dou	Ai-Dou
Observations	138	92	138	78	14	129	101
Median followers	9156	5002	5881	16000	5019	5753.5	12000
Mean likes	279.5	168.6	148.5	403.8	140.0	197.5	356.5
Mean reposts	363.4	124.9	193.6	429.6	88.8	5753.5	12000
Mean check-ins	411.3	265.0	229.3	588.3	244.7	224.8	514.5
Mean report links	8.0	8.4	8.0	8.8	6.7	8.0	8.4

Note. This table shows the average statistics of anti-smear accounts categorized by different types of celebrities with references to Section 3.2. For "median followers", it shows the median number of followers of anti-smear accounts. For "mean likes", " mean reposts", and "mean check-ins (comments)", they are indicators of the average reaction a post would receive. For "mean report links", it shows the average number of report links a post would contain.

In our 230 anti-smear accounts, each with at least 1,000 followers, the median number of followers is 7,314. On average, an anti-smear post contained 8.2 links to report, received 234.9 likes and 267.6 reposts, and 352.5 comments as “check-ins”.

Table 1 shows the basic descriptive statistics of anti-smear accounts in different categories with references to Section 3.2. It suggests that a male “Ai-Dou” in group would probably have the most active anti-smear group.

Figure 2a shows the likes and reposts received by anti-smear accounts in total. It suggests that the large-scale participation of fans in anti-smear started in early 2020 around the outbreak of COVID-19 pandemic. The spike in the repost trend in May 2021 was driven by controversy over Weibo’s bug in “like” and a celebrity who won the final in a show. Figure 2b demonstrates the likes and reposts received by all celebrities during the same time period. Their reaction changes are different from Figure 2a, which suggests that the participation in anti-smear campaigns was not solely driven by the popularity of celebrities or increasing online activities in general, and is also not a representation of activities of general fans on Weibo.

By counting the reporting links in anti-smear posts and check-in (comment) numbers under each call, we tracked the scale of anti-smear campaigns since 2016. Figure 3a shows the total reporting trend of 230 anti-smear accounts since 2017, characterized by the number of links that were reported (red line) and the number of reports (blue line) that were made in anti-smear campaigns. As one link usually suffers from multiple reports, we use different scales for the two indicators in Figure 3a.

Overall, the scale of reporting driven by anti-smear campaigns has increased from 2016 to 2021. The momentum of anti-smear campaigns was weak in 2016 but reached its first small peak around July 2017. This small peak could be related to some events of a popular singer and a patriotic movie in that period. After that, the momentum dropped to the previous level and then entered a stage of flat growth from October 2017 to January 2019. During this period, anti-smear accounts gradually expanded their workloads and received consistent support from fans.

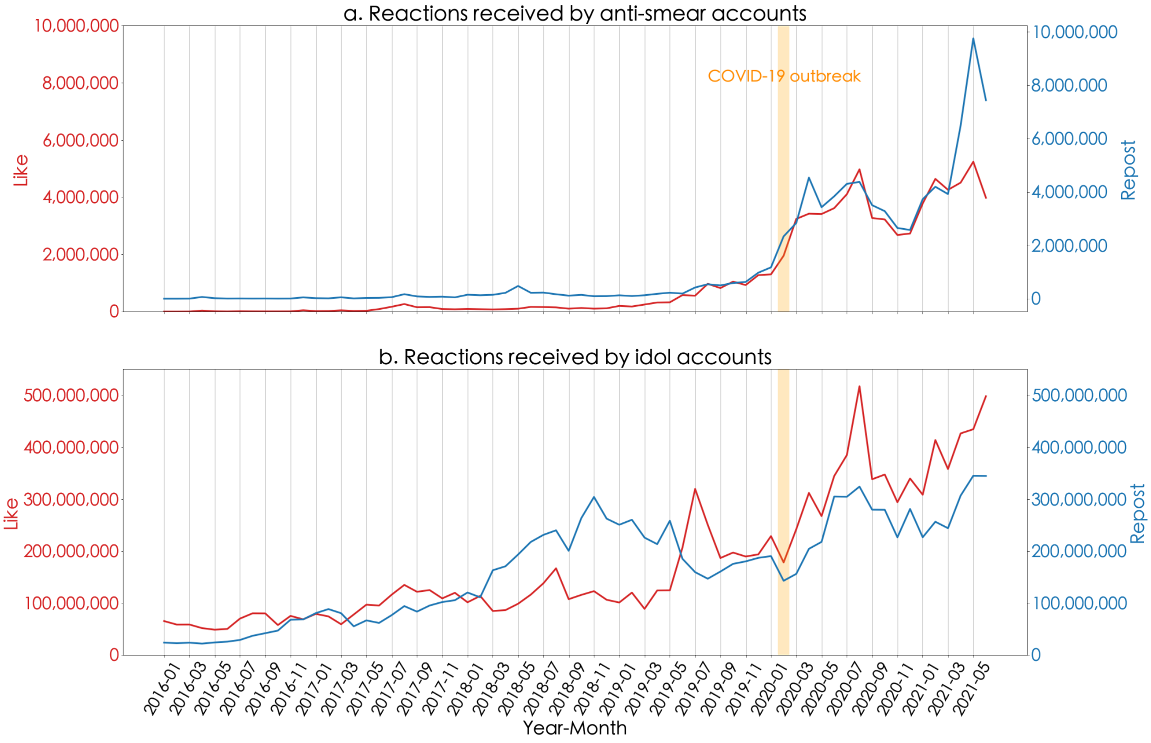

Figure 2. Likes and Reposts received by anti-smear accounts and celebrity accounts since 2016

Note. Above two figures show the reactions received by fans-operated anti-smear accounts and celebrity accounts on Weibo. The red line and the left Y-axis represent the "Like," the blue line and the right Y-axis represent the "Repost." The orange vertical bar represents the time period of the first COVID-19 outbreak and national lockdown in China.

In 2019, anti-smear activities started to take off: the number of report links and estimated reports have both increased significantly. One interesting pattern is that, between July 2017 and June 2020, a gap exists between the number of reporting links (red line) and the number of reports made by fans (blue line), meaning that the former increased at a faster rate than the latter. If we understand anti-smear campaigns from a demand-supply perspective, the gap suggests that reporting demands grew faster than reporting supplies from 2017 to 2020. In other words, anti-smear accounts posted more and more links for reporting, but fans who make reports could not catch up with their pace.

Similar to Figure 2 , Figure 3a also demonstrates a sharp increase in 2020, especially after the outbreak of COVID-19 pandemic. It is worth noting that while the report links increased by roughly 25% between January and April, the number of reports made by fans doubled in this time period. In other words, while anti-smear accounts and ordinary fans both became more active during this period, the increase in reports made by fans is proportionally higher than the increase in reporting links from anti-smear accounts. This observation is consistent with the fact that people spent more time on social media during the pandemic (Sun et al., 2020; Huang et al., 2021; Zhao and Zhou, 2021). As shown in Figure 3a, the gap between the red and blue lines disappeared and even twisted in July 2020, suggesting that reporting supplies started to grow faster than reporting demands.

In 2021, the scale of anti-smear campaigns reached its historical peak. In the heyday of anti-smear, more than 70 million reports could be driven by these 230 accounts on Weibo in a single month. To examine whether this soaring reporting is merely a parallel to the overall increase in social media activities, we use the reaction received by all celebrity accounts in Figure 2b as a baseline of the fan community's activities on Weibo. We find that the reporting trend shown in Figure 3a is not always consistent with the trend of social media activities. Specifically, while reactions received by celebrity accounts fluctuated during 2018 and 2019, reporting activities during the same period increased steadily. Also, while reactions to celebrity accounts trended ups and downs greatly between 2020 and 2021, reporting activities generally remained at high levels since 2020 (despite some minor fluctuations) and reached a peak around January 2021.

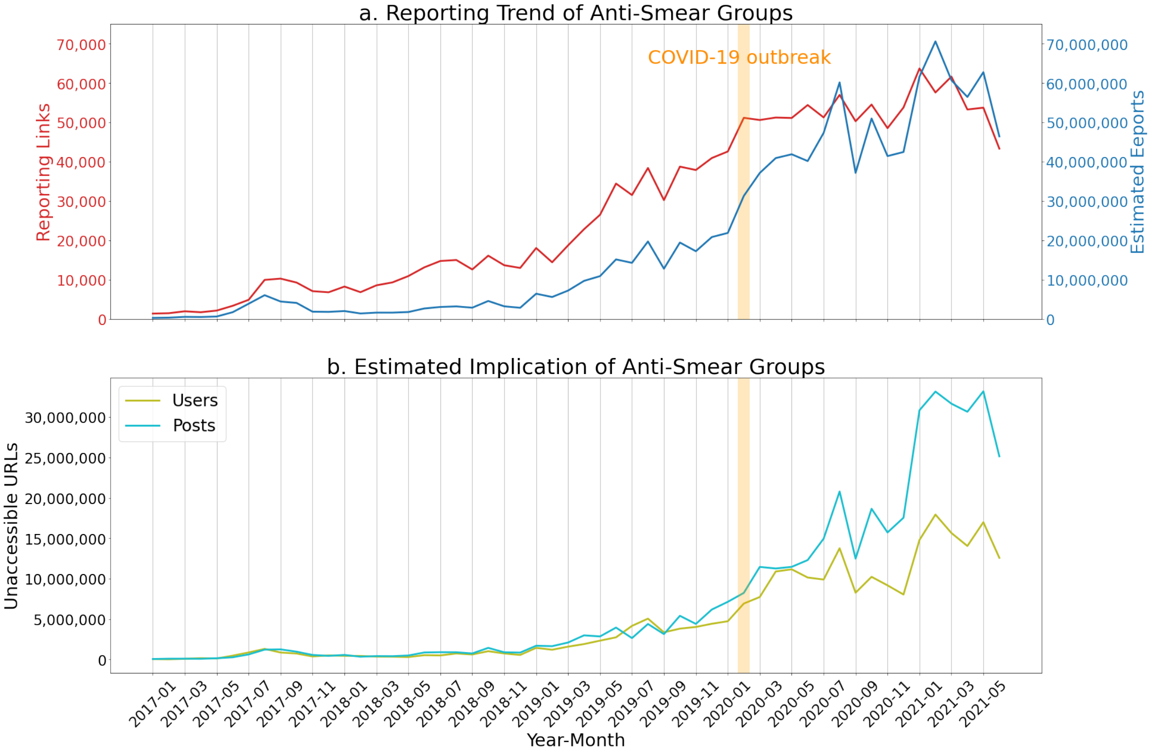

Figure 3. Reporting Trend of Anti-smear

Notes. In Figure 3a, the red line represents the number of links reported in anti-smear campaigns, estimated based on the number of links posted by anti-smear accounts as reporting targets. The blue line represents the number of reports made by fans in anti-smear campaigns, estimated by counting the number of comments that fans left under posts about anti-smear tasks. In Figure 3b, the two lines represent the number of reported links and reported users that were inaccessible in July 2021. These numbers are estimated by the method introduced in Section 3.1.

Meanwhile, we used the current status of the reported target as a proxy for reporting results to estimate the effectiveness of an anti-smear campaign, as explained in the Data section. By calculating the proportion of reports on users and posts, and reaccessing the links in the anti-smear posts, we estimated the scale of contents removed due to anti-smear reports. Since users may also hide their posts for privacy concerns, this estimation is an upper bound of the total restrain effect from all anti-smear accounts we identified. As Figure 3b shows, while anti-smear groups started to be active many years ago, only since 2019 has it begun to show their prowess. These anti-smear groups may successfully result in the inaccessibility of at most 30 million posts and 15 million accounts in one month.

We acknowledge that the comments under anti-smear posts do not always accurately represent fans' actual reporting behaviors. It is possible that a fan did not actually report but still left a comment; or in the opposite case, a fan chose not to comment after reporting. Despite these individual variations, we consider the number of comments as a legitimate estimation of the overall scale of reports made by fans, because fans have a shared understanding that the comment area under each anti-smear post is reserved for tracking the progress of anti-smear tasks. In their daily practices, anti-smear groups would use the comment number as an indicator to evaluate the completion of anti-smear tasks. Also, in our interviews with anti-smear participants, we did not observe the behaviors or tendencies of false-claiming or non-claiming. Another potential bias is that some comments may be simply commenting about the post itself rather than the ”check-in” of reporting. Due to the large amount of text in our data, we cannot examine the specific content of all comments. However, our observations suggest that these situations should be occasional and would not introduce systematic bias to the overall estimation.

Anti-Smear Strategies

The most prominent strategy in anti-smear is to coordinate mass reporting rather than just letting a few fans file individual reports. It is puzzling that fans chose a strategy which clearly cost them considerable energy and resources. Our interview results indicate that fan communities may play an important role in convincing individual fans to join this collective action and ensuring their persistent participation. However, this process-oriented interpretation does not clear the doubt about why fans tend to recognize anti-smear as an effective strategy to manipulate online discourse.

An objective-oriented hypothesis is that filing more reports is associated with a better chance of suspending reported targets. To explore this question, we again took advantage of the success rates estimated by the current status of reported targets. As shown in Figure 4a, we compared the average check-in numbers of won and lost cases since 2016. Figure 4b shows the estimated success rate of all anti-smear reports on posts and users, and both types of anti-smear activities demonstrate a clear declining chance of success over the years. Except for the beginning year, we notice that the mean check-in number of successful reports is always significantly higher ($p<0.01$) than that of failed reports, though the effect size is small (usually below 0.3) and the standard deviation is usually as big as the mean (since accounts vary in size greatly).

Combining these two patterns, we observe a positive correlation between successful reporting and the number of fans who were mobilized into mass reporting. However, the size of mobilization is clearly not the only factor driving the success of reporting; we can see a declining trend of success rates despite the increasing check-in numbers over the years that we observed in Figure 3. That is, even though anti-smear campaigns have created more reports and have been associated with more content being removed from Weibo over the years, this relation was actually weakened.

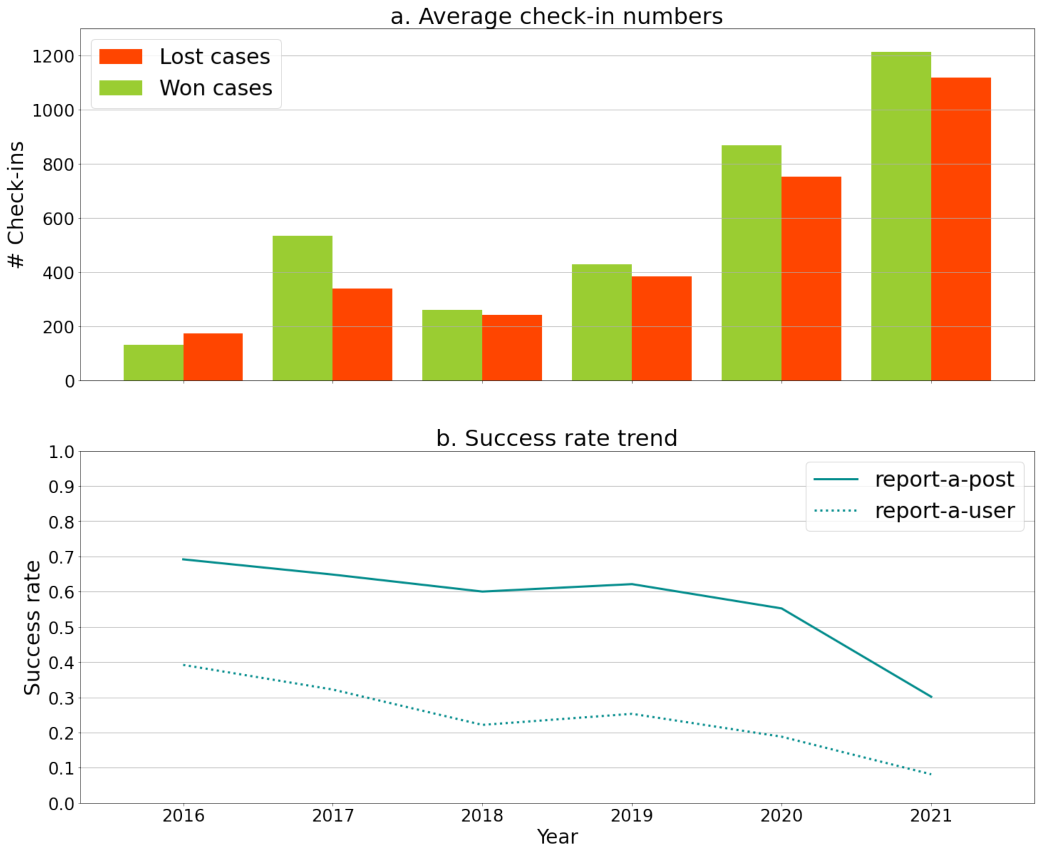

Figure 4. Results of Reporting Cases

*Note.*In Figure 4a, bars represent the average check-in numbers under anti-smear posts since 2016 for both lost and won reporting cases. In Figure 4b, the two lines represent the success rate of anti-smear reports targeting a post or a user respectively, estimated by the method introduced in Section 3.1.

Anti-smear activities have a standard workflow among all anti-smear accounts. Usually, disliked content was sent by ordinary fans to anti-smear accounts or directly collected by anti-smear accounts. Then, anti-smear accounts would post the reporting targets regularly, with simple and fixed language to instruct fans on what reporting reasons to use. These reporting reasons are pre-defined by Weibo as users can choose from different categories of reasons when reporting a post (e.g., "spam marketing"). Sometimes they did not directly spell out the reasons; rather, they used the abbreviations of reporting categories to represent the reasons. Besides, anti-smear accounts would set up a number as the reporting goal every time, and fans who finished the reporting task could tacitly comment below the anti-smear post with almost identical praise to their idols as a check-in to help with performance monitoring. These standard processes largely lower the barriers to reporting and make anti-smear an easy routine for fans.

While we cannot attribute the reporting reason for each reported link as a result of the unstructured text in anti-smear posts, we are able to track the number of reporting categories mentioned in anti-smear posts. Figure 5 demonstrates the top five frequent reporting categories and how many times they were mentioned in an anti-smear post on average since 2016. It shows that the custom of clearly indicating the reporting reasons was not prevalent in the beginning but was quickly adopted by more anti-smear accounts in the following years. Moreover, the distribution of mentioned categories suggests that fans were inclined to use ambiguous reasons like "harmful information" and "unclassified". These two categories significantly outnumber the other categories for more specific reasons such as "spam marketing", "pornographic info", and "trolling". The "unclassified" category, in particular, had a huge popularity in reporting since 2019. This selective use of reporting categories may be a strategy to maximize the likelihood of successful reporting. While there are clearer standards to determine whether the content is spam or pornography, nearly anything can be virtually considered "harmful information" or "unclassified". Because of these campaigns, it is likely that much reported content in anti-smear campaigns is "normal" content that would not otherwise be moderated. As this content does not violate any specific regulatory rules, it can only be attributed as "harmful information" when reporting.

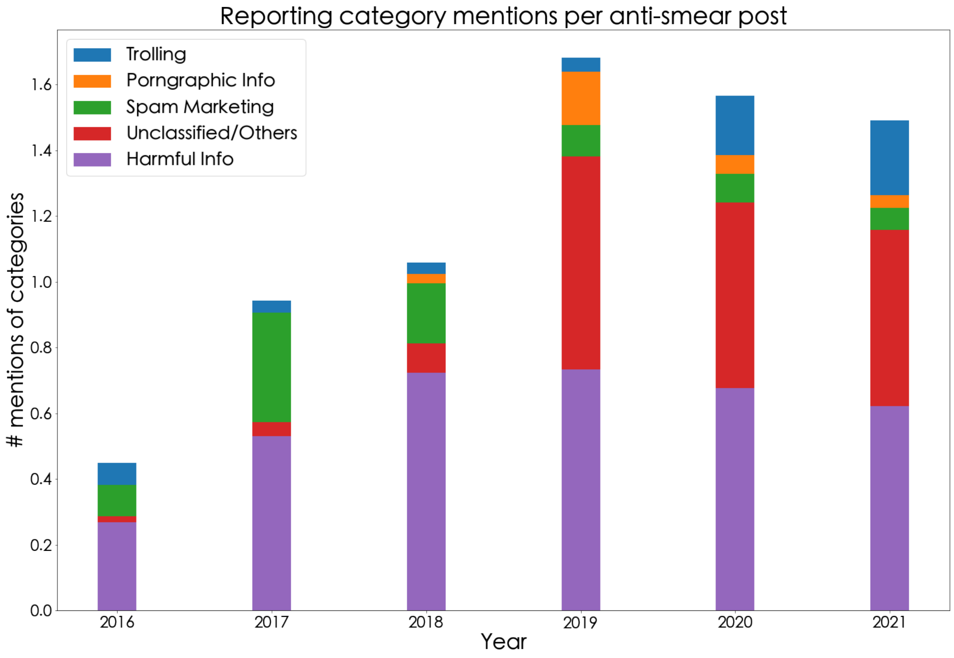

Figure 5. How Anti-smear accounts mentioned different reporting categories

Note. The colors of bars represent different reporting categories that users need to select during the reporting process. The height of each bar represents how many reporting categories were resorted to on average in one anti-smear post each year.

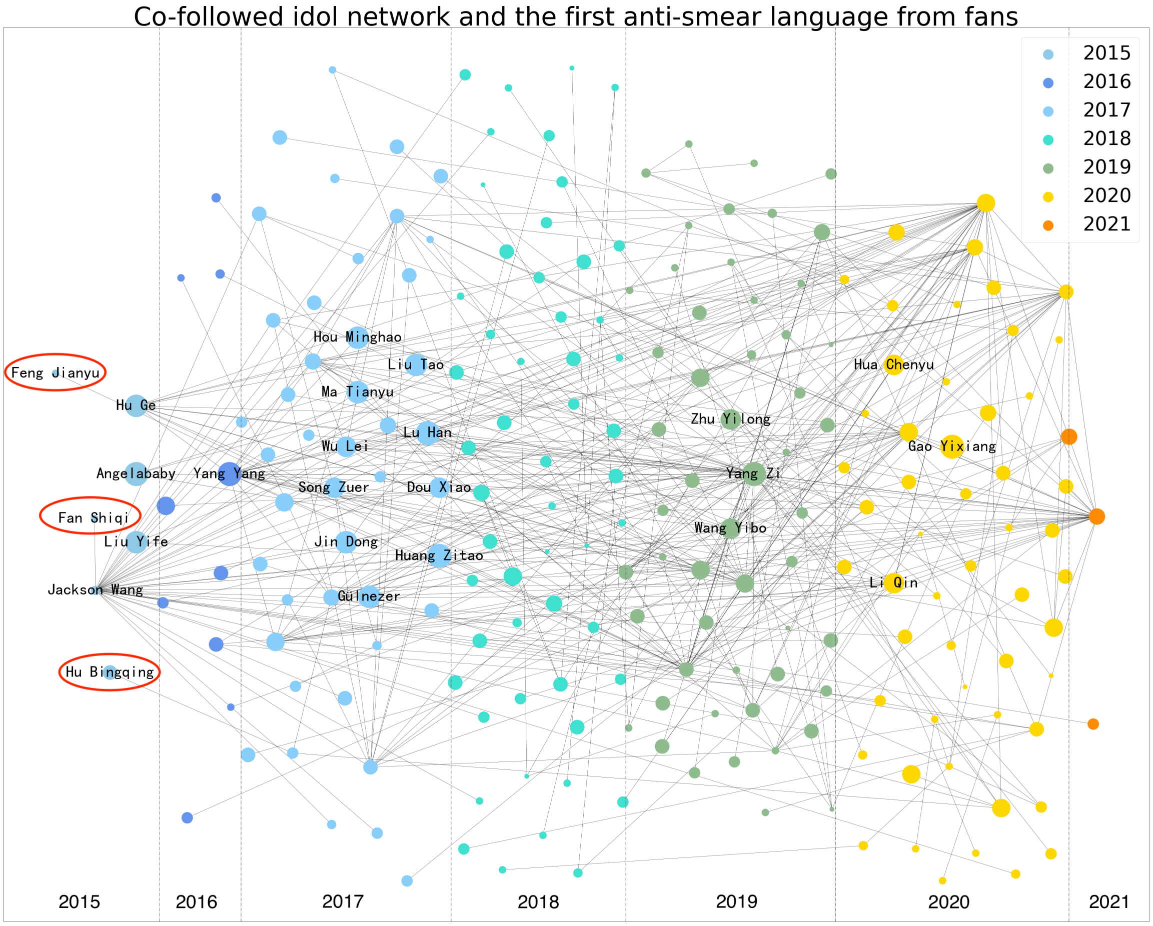

Figure 6. Diffusion of anti-smear among fans-idol network

Note. In this network, each node represents an idol and each edge represents the perceived proximity derived from co-followed relations. The node color indicates the first year that anti-smear groups of this idol started to use anti-smear language. Only nodes in 2015 or with significant popularity in our data are labeled with celebrities' names. Our data shows that anti-smear culture (language) was first adopted by relatively unpopular fan communities (nodes in red circles on the left) in 2015 and then diffused from left to right to other popular fan communities. More and more fan groups have chosen to join anti-smear campaigns in recent years.

The diffusion of anti-smear culture

The last question we explore in this paper is how anti-smear language diffuses across fan communities on Weibo. The development of a culture is accompanied by the adoption of similar language, behavior, and symbols (Spillman, 2020). Thus, describing the diffusion of anti-smear language allows us to better understand the emergence and development of the anti-smear culture within the Weibo fan population.

To portray the diffusion of the anti-smear language across fan communities, we mapped a network of celebrities with the fans following data as explained in Section 3.1. Figure 6 demonstrates the popularity and the co-followed relation of celebrities within the fan's community. The node size represents the mean search index of a certain celebrity on Baidu within one hundred days before the anti-smear group started to use anti-smear language, which is a proxy of Internet popularity at that time. The edge suggests a co-followed relationship, where two celebrities are followed by one fan at the same time. This relation implies the perceived similarity between two celebrities from the perspective of fans. The weight of the edge is the number of co-followed relations. For the sake of simplicity, we only include the edges with a weight of more than 400 or the most significant edge for relatively smaller nodes.

We demonstrate how anti-smear languages were adopted by anti-smear groups of celebrities from 2015 to 2021 in Figure 6 by node colors. Specifically, the node color represents the year in which the celebrity's anti-smear group started to adopt any of the following keywords: "净化"(cleanse), "反黑"(anti-smear), "打卡"(check-in), "目标"(goal), "教程"(instruction). Also, the locations of nodes are partially manually adjusted to emphasize the core nodes and show the timeline of the relationship in the horizontal direction.

Overall, we observe an increasing rate of adopting anti-smear language across fan communities: the number of celebrities whose anti-smear group started to use anti-smear language has increased every year since 2015. In the beginning, only seven groups were using anti-smear language, and this number increased at a relatively slow rate from 2016 to 2018. However, since 2019, the adoption rate has become faster. More than half of the celebrities' fans started to use anti-smear language in 2019 and 2020. For rising celebrities who debuted in or after 2019, their fans almost immediately adopted anti-smear language after their debut.

We paid special attention to the first few earliest celebrities whose fans started systematically using anti-smear terms, who are red-circled on the left side in Figure 6. The very earliest observable anti-smear action in our dataset came from a fan account of Fan Shiqi, a not well-known singer. While this account called for anti-smear actions in early 2015, it was not institutionalized as an anti-smear account until 2021, given the public records.

The second anti-smear account was created for a somewhat popular idol Hu Bingqing. This account started with standard anti-smear language and called her fans to report other accounts that made negative comments.

Four days later, a celebrity named Feng Jianyu had his anti-smear account created, even though he did not even have his debut and had almost zero popularity at that time. This account immediately started to call for “purifying” search pages and reporting malicious accounts. In theory, as the growth of popularity requires a certain amount of time, one should expect to see a time difference between a celebrity’s debut time and the emergence of anti-smear activities and anti-smear languages within this celebrity’s fan community. But this is certainly not the case for this anti-smear pioneer.

While our analysis suggested that the very beginning of observable anti-smear actions started with relatively lesser-known celebrities, we also noticed that the celebrities whose fans engaged in anti-smear campaigns in the early years were usually more popular. In Figure 6, the average Baidu search indexes of nodes in the first three years are 44,714, 17,770, and 30,186, respectively. However, for the latter four years between 2018 and 2021, their average Baidu search indexes are 7,164, 11,782, 14,486, and 18,121, respectively.

Moreover, our data show that 16.4% of all anti-smear accounts were created before 2015, and 20% of them were created between 2015 and 2016, which indicates that many accounts were repurposed in the following years. Because only 9% of all anti-smear accounts ever used these special terms in 2015 or 2016, we noticed that many accounts were used as ordinary fan accounts or silent users.

Discussion

Online anti-smear marks the intersection of two important, yet understudied, components of the social media landscape: Internet reporting and fandom communities. By systematically examining the activities of anti-smear accounts on Weibo, we conclude that the scale of anti-smear campaigns is too large to be ignored by scholars who are interested in the Chinese Internet. This collective action drags thousands of users into reporting others on a day-to-day basis and has created more than 70 million reporting cases during its heyday. A substantial portion of this reported content, as shown in Figure 4, was permanently removed from the Weibo sphere due to fans' anti-smear efforts. Meanwhile, anti-smear campaigns also contribute to the temporary or permanent banning of a significant number of accounts from posting content on Weibo.

The "reporting" feature on Weibo, similar to the flagging feature on other social media platforms worldwide, is designed for users to report content that violates community norms or rules. However, our data suggest that much of the reported content may not necessarily be violations of the platform's regulations. Specifically, anti-smear accounts tend to instruct fans to use ambiguous reporting reasons when using the reporting system. This strategy possibly indicates that such content does not violate any specific regulations. Consequently, such content can only be described as "harmful information" or "unclassified" when fans are eager to restrain them. This finding is further confirmed by our interview data, as an interviewee pointed out that some reported content was simply "*objective criticism*". In other words, fans' anti-smear efforts have gone beyond the platform-designated scope of content moderation. By using Weibo's content moderation system as a weapon to restrain the online speech of their rivals, fan groups are able to influence the online discourse on the Weibo sphere.

The immediate logic underlying fans' collective reporting efforts is that repeated reports may increase the chance of success of reporting (i.e., if the reported content is removed by the platform). However, our data suggest mixed results. On one hand, we find that the mean check-in number of successful reports is always significantly higher than that of failed reports. While we know little about the causality between these variables, this trend may create an impression to fans that more reports can lead to higher success rates, especially in 2017, as fans cannot control other hidden variables like us. On the other hand, we also observe that the increased total number of anti-smear campaigns is correlated with the decreasing success rate in general over the years, indicating that the potential correlation between check-in numbers and the report success rate has weakened in recent years, which could be the result of overwhelming reports in recent years (Figure 3) (Liu et al., 2022). Overall, it is likely that the strategy of collective-based reporting started to work in 2017 and proved to be an effective strategy for fans, which led to their continued enthusiastic participation in these collective actions.

Our interview data reveal other motivations driving fans' mass reporting: fan groups may intentionally use anti-smear as a way to socialize fans, enhance a sense of belonging, and build hierarchies within the community. Fans become more connected with each other and develop a shared identity of “us” through the construction of shared “rivals” and the collective efforts of reporting them. The set of practices that come with anti-smear, such as check-ins in chat rooms and anti-smear pages, serves as a unique ritual fans perform that ultimately allows for the growth of a sense of connectivity. The jargon and internal language used for anti-smear also contribute to the group's solidarity and reinforce the boundaries between the overall Weibo fandom population and outside groups. Anti-smear also functions as an indicator for distributing resources and establishing hierarchies among fans. As a result, fans face pressure to keep participating in anti-smear in order to stay informed and maintain their statuses within the community.

We also delineate how anti-smear emerged and diffused among fan groups. Our network in Figure 6 and closer observations suggest that anti-smear culture did not originate from and was not adapted by the fan groups of the most popular celebrities. Rather, the early practitioners were fans of several peripheral celebrities. While we do not have a record of the accurate origin of anti-smear campaigns, our results demonstrate that anti-smear was not adopted by many fans in its early years. However, it gradually accumulated attention beginning in 2017 and finally reached unprecedented popularity during the early pandemic.

Spillover of Anti-Smear and the State Response

The impact of anti-smear may go beyond the entertainment realm and spillover to other domains. Anti-smear, at its core, is a mass reporting strategy to call for censorship of disliked content regardless of its substance. In other words, it is essentially a strategy to manipulate information and a mindset that makes people believe that it is necessary to restrain opposing opinions. Once internalized by fans, anti-smear can be easily reproduced elsewhere, and the reporting feature on social media would be treated by participants as a resource that should not be wasted. For instance, Weibo users have applied the logic of anti-smear to endeavors beyond fandom battles: a Weibo page named "Anti-Smear for Motherland" has organized patriotic support for "Bro China" and received 7.7 billion views since 2019. In this case, Weibo users reproduced the practices of anti-smear and voluntarily helped the party-state to curb anti-regime sentiments online. The implication of anti-smear illustrates the important role of fandom communities in shaping the forms and repertoire of online activism in China.

It is also worthy noting that the fandom community demonstrates a huge potential of coordinating millions of ordinary people into completing specific tasks, such as massive reporting and crowdfunding. Moreover, fandom communities' well-structured organizations and extensive informal networks allow fans to pursue such tasks in a quick and effective manner. Such mobilizing potential and action capability are particularly unusual in China where the state has heavy-handed control over bottom-up online and offline collective actions (King et al., 2013).

Unsurprisingly, the high profile and popularity of fandom brought unintended attention from the government, which is concerned about the wild growth of fandom communities and their practices despite the apolitical or pro-regime attitudes among fans. In 2021, China started a "cleanse operation" to strike fandom activities, including high consumption, language abuse, and astroturfing (CAC, 2021). In the official discourse, fandom activities were portrayed as irrational, uncivilized, and bringing "negative energy" to the online environment. Such a discourse resonates with Yang (2018)'s finding that the state uses the ideology of "civility" to moderate online speech. While we do not know the precise intention of the cleansing operation, the fate of anti-smear campaigns seems to mirror other types of online collective actions: many anti-smear accounts became less active or even completely silent.

Reporting Culture and Weaponized Content Moderation

Anti-smear complicates our understanding of the motivation and operation of mass reporting in the digital age because anti-smear is driven by a distinctly different set of motivations and models compared to traditional mass reporting. Unlike traditional mass reporting in China, where state actors play a more important role and the activities tend to be a simple aggregation of non-organized individuals(Gong, 2000; Jiang, 2021), anti-smear operates as well-organized and routinized collective actions. Market actors, rather than state actors, play a bigger role in facilitating anti-smear. Meanwhile, fans participate in anti-smear because they consider it as a way to maintain their idols' public image and, ultimately, to prove the celebrity's commercial value to the entertainment market. Also, rather than looking for monetary rewards, fans participate in anti-smear for symbolic rewards; anti-smear legitimizes their affection to the idol, and non-participants often face strong criticism from fellow fans for being free-riders.

Despite these differences, anti-smear also resonates with traditional mass reporting in certain aspects. They both rely on the intervention of higher powers to strike undesired content or behaviors. Once successful, participants would be greatly encouraged and feel endorsed by the authority (Qin and Chen, 2021; Jiang, 2021). As getting more used to taking advantage of the reporting system to settle disagreements, users may become less tolerant of dissents, which is a common phenomenon now in China (Jiang, 2021). Overall, the similarities and divergences between traditional mass reporting and anti-smear provide a window to observe the increasingly intertwining and complex interactions among the political authority, the market, and ordinary Internet users in China.

Finally, anti-smear, as a specific form of mass reporting, points out the vulnerability of social media content moderation with crowdsourcing features in the case of the collection process. Especially in a polarized sphere, users may coordinate to impede the distribution of all content they find disturbing, which brings a great burden to the content moderation system. While our results suggest that anti-smear accounts act as key opinion leaders, the fact that fandom communities are usually active across platforms may also undermine countermeasures targeting the core coordinators. Chinese cyberspace administration suggests that Weibo accepts roughly ten million cases per month [3], and our data shows that anti-smear groups can file more than seventy million reports in a single month. While social media companies may use some technologies to minimize the negative effects of overwhelming duplicate reports, the cost of anti-smear campaigns on content moderation is still significant. A moderator from a popular social media platform once complained that the reports from anti-smear “really consume manpower” (Liu et al., 2022).

Limitation and Future Works

An undeniable flaw of this study is that we lack the perspective of the social media platform itself. Future works can explore how social media platforms, as well as the internet administration, are dealing with mass reporting in fandom or in other domains. Interviews with internal personnel would be valuable to understand the incentives and the concerns they have for platform governance.

Another limitation in our result is the review of history from the present, rather than a process tracing. As a result, we may inaccurately estimate the implication of some factors in the development of anti-smear. For example, our network (Figure 6) only captures the relations between celebrities at this moment, rather than at the time they debut. Therefore, we are likely to overestimate the influence of some idols in the celebrity network.

While our study solely relies on Weibo data, online anti-smear is by no means limited to Weibo or to the Chinese online space. We observed similar forms of reporting groups on social media platforms outside Weibo. [4] Previous studies also observed mass reporting efforts in other countries and cultural contexts (Gleicher, 2021; Nimmo and Agranovich, 2022). Such actions may also expand to troll dissidents and spread biased information, like right-wing raids from 4chan (Hine et al., 2017). Future studies are needed to understand how the operation and diffusion of anti-smear may vary across platforms or cultural contexts.

We also recognize that some of the patterns presented in this study may be particu-

[3] https://www.12377.cn/tzgg/list1.html

[4] For example, this is an anti-smear Twitter account of BTS fans: https://twitter.com/pjm_report

larly pertinent to the user base and organizational norms of the fandom community. As fans on Weibo tend to be younger generations who generally spent longer time on social media than other segments of population (Auxier and Anderson, 2021), they may be more familiar with social media platforms' functions and reporting features. Moreover, the fandom community provides a unique context where reporting behaviors are coordinated in public while reporting in other contexts is mostly unobservable. As we did not examine whether and how people may change their reporting behaviors and tactics in other settings (Alrwais and Alhodaib, 2019), we are cautious in generalizing the behavior and motivation patterns presented in this study.

Despite this limitation, we want to highlight two insights that are potentially generalizable to understand reporting behaviors in general. First, users' reporting behaviors do not occur in a vacuum but in the ongoing relationships between users (Crawford and Gillespie, 2016). As shown by the Weibo fandom, the momentum of anti-smear is rooted in the competition between celebrities over visibility and fans' embeddedness in the community. Second, users' reporting tactics depend on their interpretations of how the moderation system works, even though these interpretations are not always accurate. For example, fans increasingly engaged in anti-smear despite the decrease of anti-smear success rate. We encourage future work to yield empirical data on reporting activities from other settings and populations to test and extend the insights we find in this study.

Anti-smear is a changing and prominent phenomenon that reflects an intolerant mindset of netizens on polarized social media, as well as reshapes the design of content moderation systems and platform governance methods. Our paper can at most illustrate a small part of the landscape about how online communities may take advantage of moderation design for their own purposes. Our findings call for more attention to the massive coordinated online actions about reporting and the vulnerability of community-based moderation systems.

Acknowledgments

We would like to thank the reviewers and editors for their valuable feedback. We also thank Will Hobbs, Byungkyu Lee, and Michael Schultz for their comments on earlier versions of this paper. We are grateful to Fang Zhao for her assistance with the interviews, as well as all of our interviewees who have made this research possible.

References

Alrwais, O. and Alhodaib, E. (2019). What derives people to use reporting functions on social networks? *International Journal of Applied Information Systems*, 12(25):10–16.

Auxier, B. and Anderson, M. (2021). Social media use in 2021. https://pewresearch-org-preprod.go-vip.co/internet/2021/04/07/social-media-use-in-2021/.

Buni, C. and Chemaly, S. (2016). The secret rules of the internet. https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

CAC (2021). Cyberspace administration of china initiated “cleanse: Remediation of fandom chaos” operation (in chinese). http://politics.people.com.cn/n1/2021/0615/c1001-32130750.html.

Crawford, K. and Gillespie, T. (2016). What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428.

Dodson, C. (2020). On k-pop fans, political activism, and the necessity of nuance. https://www.teenvogue.com/story/k-pop-fans-political-activism-necessity-of-nuance.

Earl, J. and Kimport, K. (2009). Movement societies and digital protest: Fan activism and other nonpolitical protest online. *Sociological Theory*, 27(3):220–243.

Fiore-Silfvast, B. (2012). User-generated warfare: A case of converging wartime information networks and coproductive regulation on youtube. *International Journal of Communication (19328036)*, 6.

Gleicher, N. (2021). Meta’s adversarial threat report. https://about.fb.com/news/2021/12/metas-adversarial-threat-report/.

Gong, T. (2000). Whistleblowing: what does it mean in china? *International Journal of Public Administration*, 23(11):1899–1923.

Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., and Blackburn, J. (2017). Kek, cucks, and god emperor trump: A measurement study of 4chan’ s politically incorrect forum and its effects on the web. In *Eleventh International AAAI Conference on Web and Social Media*.

Huang, Q. (2021). The mediated and mediatised justice-seeking: Chinese digital vigilantism from 2006 to 2018. *Internet Histories*, pages 1–19.

Huang, Q., Chen, X., Huang, S., Shao, T., Liao, Z., Lin, S., Li, Y., Qi, J., Cai, Y., and Shen, H. (2021). Substance and internet use during the covid-19 pandemic in china. *Translational psychiatry*, 11(1):1–8.

Jenkins, H. (2006). Convergence culture. In *Convergence Culture*. new york university press.

Jian, G. (2021). Story of a fan who quit the fandom community: When fandom became “internal affairs”. https://www.jfdaily.com/news/detail?id=401074.

Jiang, J. (2021). The eyes and ears of the authoritarian regime: Mass reporting in china. *Journal of Contemporary Asia*, 51(5):828–847.

Kahne, J., Middaugh, E., and Allen, D. (2015). Youth, new media, and the rise of participatory politics. *From voice to influence: Understanding citizenship in a digital age*, pages 35–55.

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

LaiFu (2020). Fans, idols, nation: Who was destroyed after the fandom war? https://theinitium.com/article/20200310-opinion-xiaozhan-fanquan-fans/.

Liu, H. (2019). *From cyber-nationalism to fandom nationalism.* Abingdon, Oxon.

Liu, L., Zhu, L., and Yao, Y. (2022). The world of moderators: indispensable, but nobody cares (in chinese). https://chinadigitaltimes.net/chinese/676761.html.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an australian race-based controversy on twitter, facebook and youtube. *Information, Communication & Society*, 20(6):930–946.

Nimmo, B. and Agranovich, D. (2022). Meta's adversarial threat report, first quarter 2022. https://about.fb.com/news/2022/04/metas-adversarial-threat-report-q1-2022/.

Park, S. Y., Santero, N. K., Kaneshiro, B., and Lee, J. H. (2021). Armed in army: A case study of how bts fans successfully collaborated to# matchamillion for black lives matter. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Qin, X. and Chen, X. (2021). Idol disqualification, group irrationality and moral panic: Reporting strategies and incentives in fan group attacks (in chinese). *Shanghai Journalism Review*, 10.

Rosenbloom, D. H. and Gong, T. (2013). Coproducing "clean" collaborative governance: Examples from the united states and china. *Public Performance & Management Review*, 36(4):544–561.

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.

Shan, W. and Chen, J. (2021). The little pinks: Self-mobilized nationalism and state allies in chinese cyberspace. *International Journal of China Studies*, 12(1):25–46.

Spillman, L. (2020). *What is Cultural Sociology?* John Wiley & Sons.

Staff, C. (2022). Empowering china's digital informants. https://chinamediaproject.org/2022/02/07/empowering-chinas-digital-informants/.

Sun, Y., Li, Y., Bao, Y., Meng, S., Sun, Y., Schumann, G., Kosten, T., Strang, J., Lu, L., and Shi, J. (2020). Brief report: increased addictive internet and substance use behavior during the covid-19 pandemic in china. *The American journal on addictions*, 29(4):268–270.

Tan, E. (2020). Internet observation: The battle between fans of xiao zhan and the rest of fandom community under the culture of mass reporting. https://theinitium.com/article/20200302-internet-observation-xiaozhan-fans-ao3/.

Wu, J., Li, S., and Wang, H. (2019). From fans to "little pink" : The production and mobilization mechanism of national identity under new media commercial culture. In *From cyber-nationalism to fandom nationalism*, pages 32–52. Routledge.

Yang, G. (2018). Demobilizing the emotions of online activism in china: A civilizing process. *International Journal of Communication*, 12:21.

Zhang, S. and Hu, C. (2021). Fans, platform, capital and state: fans anti-smear and its governance under a multi-interaction perspective (in chinese). *Study and Practice*, pages 132–140.

Zhang, W. (2016). *The Internet and new social formation in China: Fandom publics in the making*. Routledge.

Zhao, N. and Zhou, G. (2021). Covid-19 stress and addictive social media use (smu): Mediating role of active use and social media flow. *Frontiers in Psychiatry*, 12:85.