Social Media Moderations, User Ban, and Content Generation: Evidence from Zhihu

Xiaohui Zhang
Arizona State University
XiaohuiZhang@asu.edu

Zaiyan Wei
Purdue University
zaiyan@purdue.edu

Qianzhou Du
Nanjing University
qianzhou@nju.edu.cn

Zhongju Zhang
Arizona State University
Zhongju.Zhang@asu.edu

Abstract

Social media platforms have evolved as major outlets for many entities to distribute and consume information. The content on social media sites, however, are often considered inaccurate, misleading, or even harmful. To deal with such challenges, the platforms have developed rules and guidelines to moderate and regulate the content on their sites. In this study, we explore user banning as a moderation strategy that restricts, suspends, or bans a user who the platform deems as violating community rules from further participation on the platform for a predetermined period of time. We examine the impact of such moderation strategy using data from a major Q&A platform. Our analyses indicate that user banning increases a user's contribution after the platform lifts the ban. The magnitude of the impact, however, depends on the user's engagement level with the platform. We find that the increase in contributions is smaller for a more engaged user. Additionally, we find that the quality of the user-generated content (UGC) decreases after the user ban is lifted. Our research is among the first to empirically evaluate the effectiveness of platform moderations. The findings have important implications for platform owners in managing the content on their sites.

1. Introduction

Over the last decade or so, social media and other digital platforms have evolved as major outlets for many entities to distribute and consume news and information. According to Pew Research Center, as of 2020, 71% of Americans rely on social media platforms to keep them informed. The literature has also documented that the content on social media platforms significantly affects people's political viewpoints [1], career choices [2], financial decisions [3], etc. Despite the extensive impacts, the content on social media sites are often seen by many as inaccurate, misleading, or even harmful. Additionally, such content often diffuses rapidly and can distort other users' absorption of the true information, which lead to non-negligible economic losses and negative consequences to the society [4].

With the increasing volume of misinformation and other harmful content, social media platforms have developed specific rules and guidelines to moderate and regulate their users and the content on their sites. Social media moderations refer to a series of actions employed by platforms to maintain community norms and reduce anti-social behaviors [5]. Such actions can take several forms. For example, platforms can monitor the content and/or their users either manually or through algorithm-based procedures [6]. Content monitoring helps identify potentially harmful content. Platforms can either alter or completely remove the content, which would prevent other users from being misled by the particular content [7]. User moderations, on the other hand, refer to the regulation of user behaviors on the platform, e.g., by requiring them to learn the norms of the online community [8]. In this study, we focus on one of such tools that restricts, suspends, or bans a user, who the company deems violating community rules, from further participations on the platform for a certain period of time [9]. Major social media platforms, such as Twitter, Facebook and Wikipedia, have adopted this strategy and would ban certain users from time to time.

Despite the wide adoption of user banning, our understanding of this moderation strategy is still limited. On the one hand, banning users dramatically increases their costs of inappropriate behaviors on the platforms, often times leading to their compliance with the platform norms [10]. On the other hand, banning users may result in perceptions of unfairness and discourage them from future participations, e.g., stop contributing content on the platforms [9]. In other words, by banning users, platforms face a tradeoff between maintaining the platform norms and discouraging user participation. Our goal is, therefore, to examine the effectiveness of the user banning policy on social media platforms. We aim to empirically examine whether and how temporary user

URI: https://hdl.handle.net/10125/79877
978-0-9981331-5-7

ban/restriction influences a user's behavior on the social media platform. Specifically, we ask:

Q1. Whether and how does user ban influence individual users' content generation on social media platforms?

Q2. Does user ban influence different users (high vs. low engagement) differently?

To empirically answer these research questions, we collected data from Zhihu.com, the largest questions and answers (Q&A) platform in China. Our data contain roughly 35,000 active Zhihu users with their weekly activities (e.g., questions posted, answers offered, banned by the platform or not, etc.) between September 2020 and January 2021. On Zhihu.com, the users who violate platform rules, such as posting spam-type ads and offensive content, may face temporary ban by the platform for a period of one day to two weeks.

Our findings are mainly threefold. First, we find that the users who experienced banning from the platform (for the first time) contributed more answers than others who had not been subject to any banning (Q1). Numerically, after recovering from the first banning, a typical user posts about 13% more answers (than users without any banning records). This positive impact on content generation may last at least four weeks since the lift of the ban. Second, per our research question Q2, we find that the users who were subject to a banning and had a higher level of engagement did not increase their content generation. In contrast, the banned users with low engagement (prior to banning) increased significantly the number of answers after the lift. Lastly, we find that user banning decreased the quality of answers a user offers. Specifically, the average number of words in an answer reduces after the lifting of the first ban, and the answers are less subjective.

The rest of the paper is organized as follows. After reviewing relevant literature, we develop our hypotheses in Section 2. Section 3 describes the research context and the data. We detail our econometric specifications, discuss the findings as well as their implications in Section 4. Finally, Section 5 concludes.

2. Literature Review and Hypotheses

2.1. Social Media and User-Generated Content

A large strand of the literature on social media has focused on the value of the content generated by their users [11]. For example, past research shows that user-generated content (UGC) conveys rich product information in e-commerce websites [12], helps predict the performance of private firms or stocks [13], helps detect financial fraud [3], and serves as signals of job switching [2]. Social media platforms often face the challenge of encouraging their users to keep contributing content. The literature has explored different aspects of individuals' motivations and mechanism designs to spur UGC. Wasko and Faraj [14] point out that enjoyment, gaining reputation, and social capital are important drivers of individuals' motivations to contribute UGC. Burtch, Hong, Bapna and Griskevicius [15] show that financial incentives, social norms, and connectedness affect both the quantity and quality of UGC interactively. Besides those intrinsic and extrinsic motivations, the literature has also shown that some design features, e.g., gamification, goals ladder, and performance feedback, can effectively spur contributions of UGC on social media platforms [16, 17].

Among the extant literature on UGC, studies focusing on the diffusion and consequences of misinformation on social media platforms are closely related to ours. Vosoughi, Roy and Aral [18] show that the diffusion of misinformation is often faster and broader than legitimate information. Other studies find that misinformation covers a wide variety of political, financial, and health-related topics and distorts individuals' beliefs [4, 19]. Besides misinformation, other "negative" consequences can arise from the content generated by the users. For example, Lowry, Zhang, Wang and Siponen [20] show that cyberbullying is common on many social media platforms. Levy [1] demonstrates that hate speech and spam-type content, facilitated by the "echo chamber" effect of social media, can lead to ideological polarizations.

As summarized above, the previous studies focus on either the value or the "negative" consequences of UGC on social media platforms. We contribute to the literature by studying the tradeoff between promoting content generation and maintaining the rules and social norms (from the platforms' perspective). Our current findings show that by temporarily banning users who violate the community rules, the users may not be discouraged from contributing content at least in the short run.

2.2. Social Media Moderations

As discussed in the introduction, social media platforms adopt various strategies to moderate the content generated by their users. The literature on social media moderations has shown that appropriate "proactive" strategies (to prevent users from posting misinformation and harmful content) help online communities maintain desirable environments and

promote more contributions from their users. The effects of such proactive strategies are shown to depend on the response rate and interactiveness of the moderation [5]. Slow-interactive and fast-non-interactive moderations are more effective in encouraging participation than other strategies. In addition, making community rules visible to newcomers can effectively reduce online harassment and unruly behaviors. Matias [8] shows that this proactive moderation strategy increases newcomers' compliance rate (to community rules) by more than 8% and leads to a drastic 70% increase in their participation rate.

On the flip side, the literature has also explored "reactive" strategies that platforms' moderations take place quickly after the posting of misinformation or other harmful content. Srinivasan, Danescu-Niculescu-Mizil, Lee and Tan [7] show that removing such content can reduce the immediate non-compliance rates. However, as we argued, there exists the tradeoff between enforcing community rules and encouraging user participations. For example, past studies have shown that after their content being removed, most users do not accept the deletion and express frustrations [21]. The same study also finds that asking the users to read community guidelines and providing explanations about the deletion can effectively increase the perception of fairness. Similar to our research, [22] also study moderation strategies that focus on individual users. They find that although a "blocklist" of individual users can prevent harmful interactions, such strategies cannot adequately protect users from harassment while making the blocked users feel unfair.

A recent stream of the literature has focused on user banning, which is the closest in spirit to ours [9]. However, unlike us, most previous studies examine the impact of user banning at the group or community level. For example, Chandrasekharan, Pavalanathan, Srinivasan, Glynn, Eisenstein and Gilbert [10] show that after banning two subreddits on reddit.com, a large portion of active members of the banned subreddits became inactive or deleted their accounts. Some members joined other subreddits and reduced their hate speech usage. Hobbs and Roberts [23], on the other hand, document that after the banning of Instagram in China, users elicit more information-seeking behaviors (e.g., joining other censored communities). At the individual user level, Chang and Danescu-Niculescu-Mizil [9] conduct a descriptive analysis and find that 18% of the users banned by Wikipedia violated the community rules again, and 30% of them left the community completely. Compared with these existing studies, we are the first systematic research exploring the impact of user banning on individual users' behaviors.

2.3 Research Hypotheses

We rely on the reactance theory to develop our research hypotheses. Brehm and Brehm [24] argue that if an individual perceives her freedom being threatened or lost, the reactance state of her emotion will be aroused. The authors continue to state that the reactance state drives individuals to engage in behaviors to restore freedom. Several studies in information systems (IS) have documented empirical evidence consistent with the reactance theory. For example, Murray and Häubl [25] find that perceived constraints on the freedom of choice would alter individuals' preferences over different designs of user and computer system interfaces. In particular, the lack of freedom will sabotage market leaders' advantages and drive people to competing designs. In traditional organizations, Feng and Wang [26] find that when an employee perceives inappropriate supervisions, she is less likely to share knowledge with other employees in the organization.

In our context, banning a user causes the loss of her freedom to participate on the platform (including posting questions or answers and all other activities). In such cases, the reactance theory predicts that a banned user is desired to restore the freedom of participation. Chancellor, Pater, Clear, Gilbert and De Choudhury [27] find that removing harmful content on Instagram did not reduce the production of harmful content from the same user but encouraged her to create different versions of the same (harmful) content and post other toxic content. Chancellor, Pater, Clear, Gilbert and De Choudhury [27] note that these unintended consequences underscore the possibility that removing user posts may lead to more toxic content. In our scenario, we argue that after the lifting of a temporary ban, the associated user may contribute to more content in the short run (because of the desire to restore the freedom of posting content on the platform). Therefore, our first hypothesis is:

H1. A user contributes more content to the social media platform after the ban is lifted.

Brehm and Brehm [24] continue to note that the magnitude of reactance is determined by the importance of the associated freedom and the non-compliance cost. In particular, a higher cost of non-compliance will lead to a lower intensity of reactance. Empirical findings are largely consistent with this prediction. For example, Haselhuhn, Pope, Schweitzer and Fishman [28] find that in video rentals, experiencing a larger fine boosts a customer's compliance (i.e., less reactance) more than a smaller

fine. In our context, a user who is more engaged and contributes more content on the platform arguably incurs higher costs of non-compliance (with social norms or platform rules). Therefore, the reactance theory suggests that a banning would spur less reactance from a more engaged user (because of the higher non-compliance costs). Our second hypothesis, hence, predicts that:

H2. After the lift of the user ban, the increase in user contribution to the social media platform is smaller for a more engaged user.

3. Research Context and Data

3.1. Zhihu.com and User Banning

Our research context is Zhihu.com, the largest Q&A platform and the equivalent of Quora in China. Zhihu was established in 2011 and has since evolved as the largest peer-to-peer network of knowledge, expertise, and insights in China. We choose to examine Zhihu for a number of reasons. First, Zhihu is one of the largest social media platforms with over 220 million registered users by the end of 2019. Its daily active users had reached 34 million, and the number of answers provided by its users is over 130 million. Second, user banning is widely implemented on the Zhihu platform. In our data, a random sample of active users, about 18% had been banned on the platform (for various periods of time ranging from one day to two weeks). Last but not least, unlike other major social media platforms such as Twitter and Weibo, Zhihu API offers researchers to reach a much wider range of information on its site. We take advantage of this data availability and obtain rich information from Zhihu.

On its official page about user banning, Zhihu specifies four groups of activities on the platform that can lead to user banning. They are (1) violations of national laws, (2) unfriendly activities such as intimidate speech and racism, (3) spam advertisements including those generated by machines, and (4) other malicious actives such as posting "fish content" to bait other users. Under each group, there are more specific items that Zhihu regards as violating its community rules. After a site administrator identifies one of the above activities or a user was reported (by other users) to have convicted one of those violations, Zhihu would first review the case. Then, if the violation is confirmed, the account would be banned from all activities except browsing the content on the site. Depending on the severity of the violation, an account can be banned for one day to being permanently deleted. Figure 1 shows the profile page of a user being banned (or suspended) for three days. We can tell, on the top of the page, that the user was under suspension due to violations of Zhihu's community rules (not specified how), and the banning would be lifted after three days.

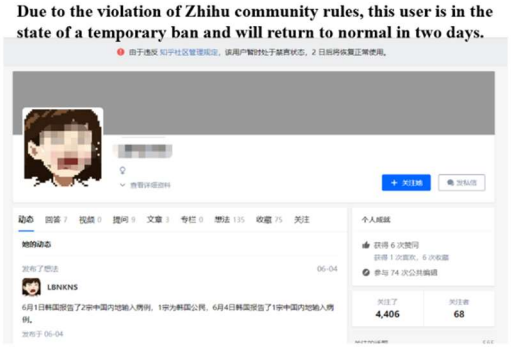

Figure 1. The Screenshot of a Profile Page During Banning

In addition to allowing users to post and answer questions, Zhihu.com also facilitates a social network among its users. That is, a Zhihu user can follow and be followed by other users. In this research, we name the user who follows another user in a following relation as the "follower" and the other party the "followee" [29]. The followee's activities on the site, e.g., posting and answering questions, will show up on the follower's news feed. Although most of the content on Zhihu is questions and answers, users can also write blogs. In addition, Zhihu users can comment on and "like" (known as "vote ups" on other platforms) others' content (blogs, answers, etc.). Unlike other social media platforms such as Twitter and Facebook, the content and interactions among users on Zhihu are centered on knowledge, expertise, and insights.

3.1. Data

Our data collection period was between September 17, 2020 and January 21, 2021. The sample, generated from a standard snowballing algorithm, contains 167,900 users and their daily records during the study period. Among all sampled users, 34,258 made at least one contribution on the platform in this period (e.g., providing answers or posting blogs). Therefore, we rely on this subset of active users in our analyses. Out of all active users, 7,376 (or 21.3%) had banning records during our data collection period. To test our research hypotheses, we rely on the first banning (ever in our study period) for a user who had experienced some.

We obtain a rich set of information for each user. For example, we collected all questions and answers posted by a user on the platform (including all posted before and during the data collection period). We also kept recording the number of followers and the number of followees a user has, so that we can

calculate the number of new followers and new followees she gains in each period (daily, weekly, etc.). In addition, we scraped other time-variant information such as user tenure (lengths of time since registration), "Excellent User" badge status (to indicate whether a user has made high quality contributions on the platform), the number of answers highlighted by the platform (because of the high quality of those answers), and the number of vote ups she obtained. Lastly, we collected a plethora of time-invariant variables, e.g., (self-reported) gender and location.

From the data, we construct an individual user by week panel for our empirical analyses. For a user in a week, we first determine her banning status and use the dummy variable, 1(After lifting) $_{it}$, to indicate the weeks after the lifting of user i's first banning (ever in our study period). Because such users are prohibited from having any activities on the platform, we remove the observations for the weeks during banning. For example, suppose user A was first banned for three days during the first week of January 2021, we delete her observation associated with that week but keep all her other observations (including the weeks after the lifting of the banning). The indicator takes the value of zero for all users who never experienced banning in our study period. Other variables are constructed accordingly. For example, # New followers $_{it}$ records the number of followers user i gains in week t. Table 1 provides the summary statistics of all variables used in our empirical analyses.

Table 1. Summary Statistics

Variable	Definition	Mean	SD	Min.	Max.
1(Ever banned)$_i$	A dummy indicating whether a user has banning records	0.21	0.38	0	1
1(After lifting)$_{it}$	A dummy indicating whether the time period is after the lifting of the first banning	0.11	0.31	0	1
# New answers$_{it}$	The number of new answers posted by user *i* in week *t*	0.63	5.39	0	1,043
# New followers$_{it}$	The number of new followers user i gains in week *t*	99.09	4,727.01	0	817,244
# New followees$_{it}$	The number of new followees user *i* gains in week *t*	1.37	20.23	0	3,608
# New vote ups$_{it}$	The number of new vote ups user *i* gains in week *t*	100.66	1,430.67	0	220,917
1(Excellent User)$_{it}$	A dummy indicating whether user *i* has the badge by week *t*	0.06	0.23	0	1
User tenure$_{it}$	Days since user *i*'s registration by week *t*	1,326.74	794.37	0	3,680
# Answers$_{it}$	Cumulative number of answers posted by user *i* by week *t*	100.80	515.44	0	47,823
# Followers$_{it}$	Cumulative number of followers user *i* gains by week *t*	11,169.89	305,482.51	0	58,060,687
# Followees$_{it}$	Cumulative number of followees user *i* gains by week *t*	376.88	762.45	0	58,655
# Voteups$_{it}$	Cumulative number of vote ups user *i* gains by week *t*	15,577.69	127,500.78	0	7,700,009

4. Empirical Models and Findings

4.1 Regression Models

To test our research hypotheses, we rely on our panel data and leverage the events of first banning by the platform to construct a differences-in-differences (DID) design. Specifically, our sample contains some users (18% of the whole sample) who had experienced banning at some point in time during our study period, while others had never been banned by the platform for the entire period. We include individual user fixed effects and week fixed effects to capture any time-invariant user unobservables (e.g., personal habits in posting content on social media platforms) and those that affect all users on the platform but can change over time (such as a change in website design at the platform level). We also include a set of user-specific, time-variant control variables such as the number of new followers a user gains in a week and her tenure on the platform. Equation (1) below is our main regression specification:

$$Y_{it} = \nu_i + \beta_1 \cdot 1(\text{After lifting})_{it} + \gamma' \mathbf{Z}_{it} + \mu_t + \varepsilon_{it}, \quad (1)$$

where ν_i are user fixed effects and μ_t are week fixed effects. The coefficient β_1 is associated with the dummy variable indicating the lifting from the first banning, $1(\text{After lifting})_{it}$, and therefore captures the effects of our interest on some outcome variable Y_{it}. Governed by our research hypotheses and the nature of a Q&A platform that Zhihu is known for, our main outcome variable is the weekly number of answers a user offers on the platform. Depending on the specific regression model, we take log transformations of the count in OLS regressions and maintain its original form (i.e., not taking logs) in models of count data, e.g., negative binomial regressions. $\mathbf{Z}_{it}$ is the vector of time-variant control variables.

Our second hypothesis, H2, formulates the idea that the effect of user banning may vary over users with different levels of engagement on the platform. To this end, we use the badge status of "Excellent User," $1(\text{Excellent User})_{it}$, and the number of answers highlighted by the platform, $\#\text{ Highlighted answers}_{it}$, as two measures of user engagement and explore their moderating roles in the impacts of user banning. Arguably, obtaining the badge of "Excellent User" and having more highlighted answers indicate higher levels of engagement with the platform. Equation (1) can be adjusted as follows to capture the moderation effects:

$$Y_{it} = \nu_i + \beta_1 \cdot 1(\text{After lifting})_{it} \cdot \text{Moderator}_{it} + \beta_2 \cdot 1(\text{After lifting})_{it} + \beta_1 \cdot \text{Moderator}_{it} + \gamma' \mathbf{Z}_{it} + \mu_t + \varepsilon_{it}. \quad (2)$$

Note that both the lifting indicator and our two measures of user engagement (as moderators) are user and week specific. We keep them in the regressions. The coefficient of the interaction term, β_1, in Equation (2) captures the moderating effect.

4.2 Findings

4.2.1 Impact of User Banning on Contributions. Main findings: Table 2 reports the impacts of user banning on individual users' subsequent contributions to answers upon lifting. Columns (1) and (2) report the coefficients from Probit and Logit regressions of the dummy variable indicating whether a user posted new answers in a particular week. We can see that the coefficients associated with the lifting dummy are positive and statistically significant in both regressions, suggesting that a user is more likely to provide new answers after the lifting of the first banning, relative to other users never experienced a banning. Quantitatively, user banning increases the likelihood by about 3% to 7% depending on the specification. Consistent with these results, we find that user banning significantly increases the quantity of a user's new answers after the lifting. The estimate in column (3) suggests that the number of new answers increases by about 2% after the lifting (relative to users who have never experienced banning). These findings all lend support to our hypothesis H1.

Table 2. Impacts of User Banning on the Likelihood and Quantity of New Answers

	DV: $1(\#\text{ New answers} > 0)_{it}$		DV: $\log(\#\text{ New answers})_{it}$	DV: $\#\text{ New answers}_{it}$	DV: $\#\text{ New answers}_{it}$
	(1) Probit	(2) Logit	(3) OLS	(4) PPML	(5) NB
$1(\text{After lifting})_{it}$	0.038*	0.072**	0.019***	0.120**	0.112***
	(0.020)	(0.035)	(0.005)	(0.057)	(0.027)
$\#\text{ New vote ups}_{it}$	0.000	0.001**	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$\#\text{ New followees}_{it}$	0.001	0.005**	0.000***	0.002***	0.002**
	(0.001)	(0.002)	(0.000)	(0.001)	(0.001)

# New followers$_{it}$	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
User tenure$_{it}$	-0.082**	-0.128*	-0.007	0.000	-0.005
	(0.041)	(0.072)	(0.012)	(0.043)	(0.043)
# Answers$_{i,t-1}$	-0.134***	-0.172***	-0.016	-0.308*	-0.311***
	(0.034)	(0.060)	(0.010)	(0.167)	(0.063)
# Voteups$_{i,t-1}$	-0.010	-0.036	0.003	0.193**	0.059*
	(0.021)	(0.036)	(0.005)	(0.083)	(0.034)
# Followers$_{i,t-1}$	-0.024	-0.044	-0.030**	-0.184***	-0.137***
	(0.037)	(0.064)	(0.012)	(0.062)	(0.052)
# Followees$_{i,t-1}$	0.110***	0.201***	0.013	0.030	0.096
	(0.041)	(0.077)	(0.010)	(0.081)	(0.062)
User FE	Y	Y	Y	Y	Y
Week FE	Y	Y	Y	Y	Y
Num.Obs.	181,755	181,755	181,755	181,755	181,755

*p<0.1; **p<0.05; ***p<0.01

Note:Robust standard errors clustered at the individual level are shown in parentheses. (Default in the following tables)

Table 2 also reports some interesting patterns in the relationships with the control variables. For example, we find that the number of new vote ups and the number of new followees are positively correlated with the dependent variables. In contrast, the longer a user has been on the platform, the fewer the answers she would provide. Lastly, the associations with the cumulative number of followers and followees (by the week before the current one) take opposite signs.

Other empirical specifications: In columns (4) and (5), we estimate two nonliner regressions that are dedicated to model count data, i.e., a Poisson pseudo-maximum-likelihood (PPML) model and a Negative Binomial model. Because our contribution measure is the number of new answers provided, one would naturally suggest estimating models of count data. The results are highly consistent with our main findings (column (3) of Table 2) and confirm that a user would indeed provide more answers after the lifting of the first banning.

An important assumption of a valid DID design is that there existed relatively parallel trends (in providing answers) between the users who had experienced at least one user banning and those who had not. Table 3 reports the coefficients of time dummies from estimating a relative time model. Specifically, we include dummies indicating five weeks before the first banning and seven weeks after (with two additional dummies indicating 6+ weeks before and 8+ weeks after respectively). First, most of the pre-banning coefficients are not statistically significant, suggesting that the two groups of users were indeed comparable in terms of providing answers before the first banning. Second, the coefficients of post-banning dummies are positive and statistically significant immediately after the lifting, which indicates that a user increased her contributions right after being freed to post content on the platform. Lastly, we also notice that the positive effect is insignificant after seven weeks (including the seventh-week dummy and 8+ week dummy), suggesting that the user's contribution level drops back to its pre-banning level after about two months.

Table 3. Relative Time Model

DV:	log(# New answers)$_{it}$
6+ week before banning	-0.070
	(0.123)
5 week before banning	-0.128**
	(0.057)
4 week before banning	-0.049
	(0.030)
3 week before banning	0.018
	(0.064)
2 weeks before banning	0.183
	(0.124)
1 week before banning omitted as baseline	
1 week after banning release	0.214***
	(0.051)
2 weeks after banning release	0.093***
	(0.036)
3 weeks after banning release	0.178***
	(0.046)
4 weeks after banning release	0.269***
	(0.103)

5 weeks after banning release	0.088**
	(0.042)
6 weeks after banning release	0.219**
	(0.086)
7 weeks after banning release	0.159
	(0.149)
8+ weeks after banning release	0.016
	(0.062)
Control variables	Y
User FE	Y
Week FE	Y
Num.Obs.	187,356

*p<0.1; **p<0.05; ***p<0.01

Although we control for both time-variant and time-invariant factors in our main specification (Equation (1)), one may still worry about endogeneity. To further alleviate the concern and also test the robustness of our findings, we follow the literature [30] and estimate our DID specification using a sample generated by a look-ahead propensity score matching method. Table 4 reports the regression results. The coefficient of the lifting dummy continues to be positive and statistically significant.

Table 4. LA-PSM

DV:	log(# New answers)$_{it}$
1(After lifting)$_{it}$	0.276***
	(0.087)
Control variables	Y
User FE	Y
Week FE	Y
Num.Obs.	17,355

*p<0.1; **p<0.05; ***p<0.01

4.2.2 Moderating Effects Our second hypothesis predicts that more engaged users should have experienced a smaller increase (or even a decrease) in content contributions after the lifting. As we discussed in the previous section, we explore the moderating role of "Excellent User" badge status and the number of highlighted answers to test the hypothesis. Table 5 reports the coefficients from estimating Equation (2). The negative and statistically significant coefficients of the interaction terms (between the lifting dummy and the moderators) clearly imply that more engaged users indeed provided fewer answers than less engaged users after the lifting of the first banning. In particular, a user who has a large number of highlighted answers may in fact provide fewer answers on the platform after the lifting of the first banning. These findings are consistent with H2. Further, we explore the impacts of banning reasons. In the data, most observations' banning reasons are labeled as "OTHER" and a small portion of observations are labeled as "Advertisements" and "Unfriendly". We category the following two reasons as malicious reason and explore the moderation effect. The result is shown in Table 5, column (3). We can observe that after lift banning, the users banned due to malicious reasons are actually more active than other reasons. This indicates that banning is not suit to moderate maliciously intended behaviors.

Table 5. Moderation Effects

	DV: log(# New answers)$_{it}$		
	(1)	(2)	(3)
1(After lifting)$_{it}$ x 1(Excellent User)$_{it}$	-0.278*		
	(0.144)		
1(After lifting)$_{it}$ x # Highlighted answers$_{it}$		-0.793***	
		(0.267)	
1(After lifting)$_{it}$ x Malicious Reason			8.753***
			(0.126)
1(After lifting)$_{it}$	0.128**	0.126**	0.117**
	(0.058)	(0.057)	(0.057)
1(Excellent User)$_{it}$	0.021		
	(0.147)		
# Highlighted answers$_{it}$		-0.210	
		(0.204)	
Malicious Reason			-8.515***
			(0.042)
Control variables	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Num.Obs.	187,356	187,356	187,356

*p<0.1; **p<0.05; ***p<0.01

4.2.3. Further Analysis: Impact of User Banning on Qualitative Attributes of Answers We

also study the impact of user banning on the "qualitative" aspects of contributed answers. Table 6 reports the impacts on a series of such attributes (all generated by Linguistic Inquiry and Word Count, or LIWC). Interestingly, we note that the average number of words in an answer decreased after the lifting of the first banning (column (1)). This finding indicates that a user exhibits less effort in providing answers after experiencing a banning. In terms of the subjectivity of answers, we find that the answers provided by a user after the lifting of the first banning are more subjective (i.e., less objective), as in column (2) of Table 6. Similarly, column (3) shows that after the first banning, a user uses less cause words (e.g., because, therefore) which usually correlate with logical expressions. Further, from column (4) and column (5), we can see that users after the first banning would utilize more negative emotional words but less concrete numerical data. These findings indicate that the banning experience turns a user away from effort consuming objective knowledge output and promotes weak logic emotional expressions.

Table 6. Impact of User Banning on Qualitative Attributes of New Answers

	(1)	(2)	(3)	(4)	(5)
DV:	Avg. # words	Subjectivity	Cause	Negative	Number
1(After lifting)$_{it}$	-26.353***	0.004**	-0.350***	0.323**	-0.333**
	(7.262)	(0.002)	(0.087)	(0.144)	(0.137)
Control variables	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes
Num.Obs.	181,755	181,755	181,755	181,755	181,755

*p<0.1; **p<0.05; ***p<0.01

5. Conclusions

As social media becomes increasingly popular, many rely on such platforms to consume and distribute news and other information. To mitigate the effects of inaccurate, misleading, or harmful information on their sites, social media platforms have adopted specific tools to moderate and regulate the content and their users. Specifically, user ban is becoming increasingly popular on social media platforms such as Twitter and Instagram. A user would be banned from further participation for a certain period of time if she is deemed as violating community rules. We examine this moderation strategy and how user ban affects individual users' future contributions to UGC (both quantitatively and qualitatively).

Our research context is the largest Q&A platform in China, Zhihu.com. Relying on a rich set of information obtained from the site, we find that a typical user would increase her contributions (by providing more answers specifically) upon the lifting of the first ban. The magnitude of the increase, however, is moderated by the engagement level of the user with the platform. The increase is smaller, or even non-existent, for a more engaged user. Equally interesting, we find that the quality of answers decreases after the user ban if lifted. Specifically, the average word count in an answer decreases after the user ban, and banning a user drives up the subjectivity of her answers. These findings have both theoretical and practical implications. To the best of our knowledge, this study is the first to establish empirical evidence about how user ban affects individual users' subsequent behaviors. Practically speaking, our findings offer important insights for platform owners to consider implementing similar moderation strategies.

6. References

[1] Levy, R. e. Social media, news consumption, and polarization: Evidence from a field experiment. American Economic Review, 111, 3 (2021), 831-870.

[2] Huang, P. and Zhang, Z. Participation in Open Knowledge Communities and Job-Hopping: Evidence from Enterprise Software. Mis Q., 40, 3 (2016), 785-806.

[3] Dong, W., Liao, S. and Zhang, Z. Leveraging Financial Social Media Data for Corporate Fraud Detection. Journal of Management Information Systems, 35, 2 (2018), 461-487.

[4] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. and Lazer, D. Fake news on Twitter during the 2016 US presidential election. Science, 363, 6425 (2019), 374-378.

[5] Wise, K., Hamman, B. and Thorson, K. Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. Journal of Computer-Mediated Communication, 12, 1 (2006), 24-41.

[6] He, Q., Hong, Y. and Raghu, T. The Effects of Machine-powered Platform Governance: An Empirical Study of Content Moderation. Available at SSRN (2021).
[7] Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L. and Tan, C. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. Proceedings of the ACM on Human-Computer Interaction, 3, CSCW (2019), 1-21.
[8] Matias, J. N. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. Proceedings of the National Academy of Sciences, 116, 20 (2019), 9785-9789.
[9] Chang, J. and Danescu-Niculescu-Mizil, C. Trajectories of blocked community members: Redemption, recidivism and departure. City, 2019.
[10] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J. and Gilbert, E. J. P. o. t. A. o. H.-C. I. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, 1, CSCW (2017), 1-22.
[11] Zhang, K., Evgeniou, T., Padmanabhan, V. and Richard, E. Content contributor management and network effects in a UGC environment. Marketing Science, 31, 3 (2012), 433-447.
[12] Chen, P.-Y., Hong, Y. and Liu, Y. The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. Management Science, 64, 10 (2018), 4629-4647.
[13] Sul, H. K., Dennis, A. R. and Yuan, L. Trading on twitter: Using social media sentiment to predict stock returns. Decision Sciences, 48, 3 (2017), 454-488.
[14] Wasko, M. M. and Faraj, S. Why Should I Share? Examing Social Capital And Knowledge Contribution In Electronic Networks Of Practice. Mis Q., 29, 1 (2005), 35-57.
[15] Burtch, G., Hong, Y., Bapna, R. and Griskevicius, V. Stimulating online reviews by combining financial incentives and social norms. Management Science, 64, 5 (2018), 2065-2082.
[16] Goes, P. B., Guo, C. and Lin, M. Do Incentive Hierarchies Induce User Effort? Evidence from an Online Knowledge Exchange. Information Systems Research, 27, 3 (Sep 2016), 497-516.
[17] Huang, N., Burtch, G., Gu, B., Hong, Y., Liang, C., Wang, K., Fu, D. and Yang, B. Motivating user-generated content with performance feedback: Evidence from randomized field experiments. Management Science, 65, 1 (2018), 327-345.
[18] Vosoughi, S., Roy, D. and Aral, S. The spread of true and false news online. Science, 359, 6380 (2018), 1146-1151.
[19] Zhang, X., Du, Q. and Zhang, Z. An explainable machine learning framework for fake financial news detection. In Proceedings of the ICIS 2020 Proceedings (2020), [insert City of Publication],[insert 2020 of Publication].
[20] Lowry, P. B., Zhang, J., Wang, C. and Siponen, M. Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model. Information Systems Research, 27, 4 (Dec 2016), 962-986.
[21] Jhaver, S., Appling, D. S., Gilbert, E. and Bruckman, A. " Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit. Proceedings of the ACM on human-computer interaction, 3, CSCW (2019), 1-33.
[22] Jhaver, S., Ghoshal, S., Bruckman, A. and Gilbert, E. Online harassment and content moderation: The case of blocklists. ACM Transactions on Computer-Human Interaction, 25, 2 (2018), 1-33.
[23] Hobbs, W. R. and Roberts, M. E. How sudden censorship can increase access to information. American Political Science Review, 112, 3 (2018), 621-636.
[24] Brehm, S. S. and Brehm, J. W. Psychological reactance: A theory of freedom and control. Academic Press, 2013.
[25] Murray, K. B. and Häubl, G. Freedom of choice, ease of use, and the formation of interface preferences. Mis Q., 35, 4 (2011), 955-976.
[26] Feng, J. and Wang, C. Does abusive supervision always promote employees to hide knowledge? From both reactance and COR perspectives. Journal of Knowledge Management (2019).
[27] Chancellor, S., Pater, J. A., Clear, T., Gilbert, E. and De Choudhury, M. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. City, 2016.
[28] Haselhuhn, M. P., Pope, D. G., Schweitzer, M. E. and Fishman, P. The impact of personal experience on behavior: Evidence from video-rental fines. Management Science, 58, 1 (2012), 52-61.
[29] Wei, Z., Xiao, M. and Rong, R. Network Size and Content Generation on Social Media Platforms. Production Operations Management, 30, 5 (2020), 1406-1426.
[30] Bapna, R., Ramaprasad, J. and Umyarov, A. Monetizing freemium communities: does paying for premium increase social engagement? Mis Q., 42, 3 (2018), 719-736.