Post Approvals in Online Communities

Manoel Horta Ribeiro,[1]* **Justin Cheng,**[2] **Robert West**[1]

[1]EPFL, [2]Meta
manoel.hortaribeiro@epfl.ch, jcheng@fb.com, robert.west@epfl.ch

Abstract

In many online communities, community leaders (i.e. moderators and administrators) can proactively filter undesired content by requiring posts to be approved before publication. But although many communities adopt post approvals, there has been little research on its impact on community behavior. Through a longitudinal analysis of 233,402 Facebook Groups, we examined 1) the factors that led to a community adopting post approvals and 2) how the setting shaped subsequent user activity and moderation in the group. We find that communities that adopted post approvals tended to do so following sudden increases in user activity (e.g. comments) and moderation (e.g. reported posts). This adoption of post approvals led to fewer but higher-quality posts. Though fewer posts were shared after adoption, not only did community members write more comments, use more reactions, and spend more time on the posts that were shared, they also reported these posts less. Further, post approvals did not significantly increase the average time leaders spent in the group, though groups that enabled the setting tended to appoint more leaders. Last, the impact of post approvals varied with both group size and how the setting was used, e.g., group size mediates whether leaders spent more or less time in the group following the adoption of the setting. Our findings suggest ways that proactive content moderation may be improved to better support online communities.

1 Introduction

Online communities are partially shaped by the design affordances of the platforms they inhabit (Bucher and Helmond 2018). In Facebook Groups, administrators can turn on "post approvals," a setting that requires members' posts to be accepted by community leaders (i.e., administrators and moderators) before others in the group can see and interact with them (Meta 2022). This setting changes *when* norms are enforced in a community or group, as illustrated in Figure 1. If the setting is turned off, community leaders must reactively moderate the posts in the community, e.g., by browsing posts in the group as they appear or by responding to reports from other members or the platform. If the setting is turned on, leaders can proactively moderate the community, prescreening posts that are low quality or that break the rules.

*Work done mostly while interning at Meta.

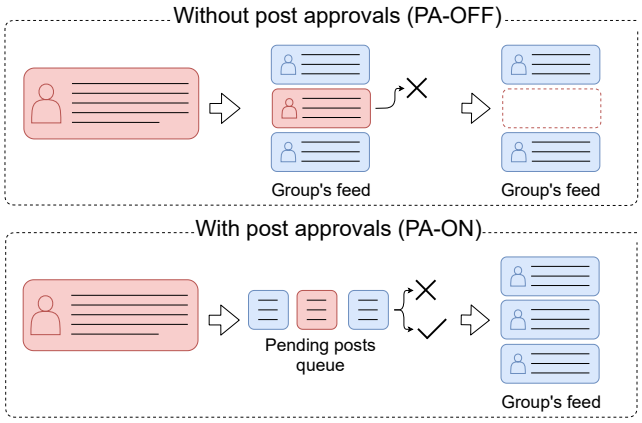

Figure 1: Post approvals allow posts that violate a community's guidelines (in red) to be filtered *before* other members in the community see them. Without post approvals, these posts can only be moderated *after* they are posted in the group.

Well-moderated spaces are more attractive to users (Wise, Hamman, and Thorson 2006) and can improve the quality of users' contributions (Cosley et al. 2005). However, overenforcement of rules can discourage participation (Jhaver, Bruckman, and Gilbert 2019; Kiene, Monroy-Hernández, and Hill 2016), and moderation creates more work for leaders (Lo 2018; Dosono and Semaan 2019). Thus, post approvals, a proactive moderation strategy, involves several trade-offs. On the one hand, it may prevent harm caused by violations of a community's guidelines and improve members' overall experience. On the other hand, it introduces participation friction and may increase leaders' workloads.

Present work. This paper presents an observational study of the adoption and the impact of post approvals in online communities. We ask:

- **RQ1** What leads communities to adopt post approvals?
- **RQ2** How do post approvals shape user activity and moderation in online communities?
- **RQ3** Does the impact of post approvals depend on community properties and on how the setting is used?

Using a longitudinal dataset of user activity- and moderation-related traces from 233,402 Facebook Groups from March to July 2021, we compared communities that enabled post approvals (PA-ON; $n = 8{,}767$) to communities that did not change any moderation-related settings (PA-OFF; $n = 224{,}635$).

To examine the factors that led to the adoption of post approvals **(RQ1)**, we studied activity in PA-ON and PA-OFF communities in the 4 weeks before the former enabled the setting. During this period, PA-ON communities experienced greater growth in user activity (e.g., *Comments*) and moderation (e.g., *Posts reported*) compared to PA-OFF communities. Further, right before PA-ON communities enabled post approvals, they experienced a sudden increase in moderation, which may have been the final straw that led administrators to turn on the setting. These findings continue to hold when using propensity score matching to control for initial baseline user and moderation activity, and are further confirmed when examining how user activity or moderation predicts if post approvals will be turned on in future weeks.

To study how post approvals shape online communities **(RQ2)**, we matched PA-ON and PA-OFF communities on user activity and moderation traces in the 4 weeks prior to post approvals being turned on and compared differences in their subsequent activity. We found that, while fewer posts were shared in groups that enabled post approvals, the posts that were shared received more comments, more reactions, more time spent, and fewer reports, suggesting improvements in the quality of content being posted. Further, post approvals did not significantly increase the average time leaders spent in their groups, though groups that enabled the setting tended to increase their moderation team.

Last, post approvals may differently impact a community depending on its properties and on how post approvals are used in practice **(RQ3)**. To understand these differences, we studied how the effects of post approvals varied with group size (i.e., how many members there were in a group), leaders' response time for submitted posts (i.e., how much time did it take for a post to be approved) and the post approval rate (i.e., what fraction of posts submitted in a given group were approved).

For all three factors, we found significant interactions with time spent by leaders in the group and with changes in activity in the group following the adoption of the setting. Leaders spent significantly more time after adopting post approvals in larger groups, groups with higher post approval rates, and groups with faster response times. There were sharper decreases in the number of posts and increases in the number of comments, reactions, and time spent per post in larger groups, groups with lower post approval rates, and groups with slower response times. Still, other changes persisted across different communities. Independent of group size, response time, or approval rate, the fraction of posts reported decreased significantly after post approvals was adopted. This suggests that regardless of how post approvals was enforced, the setting nonetheless reduced content perceived by members as problematic or rule-breaking.

Overall, our findings suggest that post approvals substantially change how online communities work and that the setting creates communities centered around fewer, higher-quality posts. These insights may guide improvements to community-level moderation processes and the quasi-experimental approach we adopted can be easily extended to analyze other opt-in features provided by social media platforms.

2 Related Work

Moderation in online communities increases their attractiveness to newcomers (Wise, Hamman, and Thorson 2006), improves the quality of contributions (Cosley et al. 2005), and decreases anti-social behavior (Seering, Kraut, and Dabbish 2017).

Nonetheless, effective moderation is difficult – platforms experience several challenges related to the scale, the legitimacy, and the contextual nature of content moderation (Gillespie 2018; Filgueiras and Almeida 2021). Beyond work to better predict when content may violate community guidelines (Schmidt and Wiegand 2019; Papadamou et al. 2020), community-oriented social media (and moderation) has been suggested as part of the solution to these challenges because community leaders may better incorporate local and cultural context into moderation decisions (Seering 2020) and because the decisions taken would be considered more legitimate (Filgueiras and Almeida 2021). To this end, some research has examined how moderators engage and regulate their communities (Seering et al. 2019) by developing specific design guidance from fundamental theories in the social sciences (Kraut and Resnick 2012).

Most relevant to the present work are existing studies that explored how technological affordances provided by platforms shape content moderation. For example, participation controls (Kraut and Resnick 2012) limit what specific users are allowed to see or do within a social media platform or a specific online community. In the development of open-source software, collaborators receive "commit rights" as they offer evidence of their technical expertise (Ducheneaut 2005); on PalTalk (an early video group chat service), moderators could impose "activity quotas" to chat room users, limiting their participation (Kraut and Resnick 2012); on Twitch, moderation includes chat "modes" that change how users can participate, for instance allowing only emotes to be sent (Seering, Kraut, and Dabbish 2017); on Reddit, Jhaver et al. (2019) studied the usage of *AutoModerator*, a system that allows moderators to define "rules" to be automatically applied to posts in their communities.

This work examines post approvals, a participation control that is central to community-level moderation in Facebook Groups but whose specific effects have not yet been systematically studied. Post approvals change the dynamics of content moderation by allowing community leaders to *proactively* moderate posts before they ever land in the communities' feeds. Moreover, when someone attempts to contribute to a community with post approvals turned on, posts may take hours or even days to get published (if they do). Communities could thrive in the better-moderated spaces enabled by post approvals (Wise, Hamman, and Thorson 2006) and the participation friction could disrupt mindless interactions (Mejtoft, Hale, and Söderström 2019). However, the

setting could also discourage participation (Kiene, Monroy-Hernández, and Hill 2016) and create unnecessary work for leaders (Lo 2018; Dosono and Semaan 2019).

Studying the impact of post approvals (and other participation controls) in online communities can help create better governance practices and further our understanding of how participation friction and proactive moderation can improve online spaces.

3 Data

Between March 28, 2021, and July 11, 2021, we collected data on 1) communities that turned on post approvals and did not change other moderation-related settings (PA-ON; $n = 8{,}767$); and 2) a random sample (50%) of communities that did not change *any* moderation-related setting (PA-OFF; $n = 224{,}635$). For PA-ON groups, we considered only communities that enabled post approvals at least 28 days after the start of the study period and at least 28 days before its end. For both PA-ON and PA-OFF groups, we considered only communities with 128 or more members and at least one comment and one post over any 7-day window.

We analyzed PA-ON communities relative to when they turned on post approvals, referring to the day when they enabled the setting as day 0. For PA-OFF communities, we randomly assigned a pseudo-intervention date drawn from the distribution of dates (day and hour) when PA-ON groups enabled post approvals (cf. Appendix C; Fig. 10). We considered the set of variables described in Table 1 in the 28 days before and after each intervention, for a total of 57 days (from -28 to 28). Some variables are 1) time-dependent, capturing group activity and group moderation (e.g., number of posts, number of posts deleted), while others are 2) time-invariant, capturing group topic, demographics, and moderation settings (e.g., group visibility, group category, if a group was a buy-and-sell group, etc.).

All data was de-identified and analyzed in aggregate, and no individual-level data was viewed by the researchers. In the analyses that follow, variables were 95%-winsorized (i.e., the 2.5% smallest and largest values were replaced with the most extreme remaining values (Wilcox 2011)) prior to aggregation unless otherwise stated. This ensured that trends/effects were not dominated by a few large groups. Nonetheless, results were qualitatively similar without winsorization.

Group characteristics (not time-dependent)	
Visibility*	Whether group is private or public.
Join approvals*	Whether leaders have to manually approve new members.
Average age*	Average age of the members in the group.
% women*	The percentage of women in the group.
Buy-&-Sell*	Whether the group is a buy and sell group or not (specified by admin).
Group categories*	Lexical categories obtained from the groups' description and title using Empath (Fast, Chen, and Bernstein 2016). See Appendix B for details.
Moderation-related	
Moderating TS	Average time leaders spent in moderation-related interfaces (e.g., approving posts).
Leader TS	Average time leaders spent in the group.
Members Removed	Number of members removed.
Posts deleted	Number of posts by regular members deleted by leaders.
Posts reported	Number of posts reported in the community by users.
Num leaders	Number of leaders in the community.
Activity-related	
Posts	Number of posts.
Comments	Number of comments.
Time spent	Total time users spent browsing posts in the group (in hours).
Reactions	Number of Likes and of other reactions (Sad, Happy, Wow, Laugh, Angry).
Num members	Number of members in the community.

Table 1: Description of the group-level variables considered in this paper. Variables marked with a star ($\star$) were measured on the day prior to the intervention (for PA-ON groups) or the pseudo-intervention (for PA-OFF groups). They were not analyzed in the result sections of this paper, but were used in the matching to ensure the two sets of communities were comparable (cf., Appendix A).

4 What Leads to the Adoption of Post Approvals?

This section examines *why* communities adopt post approvals to begin with (**RQ1**). We focus on what happens *before* the setting was enabled, contrasting PA-ON and PA-OFF groups. All analyses in this section were done at the group level, with each group weighted equally.

Case–control analysis. Our first analysis follows a case–control design (Schlesselman 1982): we compared user activity and moderation traces of groups that enabled post approvals (PA-ON; the "case") with those that did not change any moderation settings (PA-OFF; the "control"). We considered three variables related to user activity (*Num Members*, *Posts*, and *Comments*) and three variables related to moderation activity (*Posts reported*, *Posts deleted*, *Leader TS*) in the 28 days before the intervention (cf. Table 1 for descriptions). We refer to this scenario as "all" since, in what follows, we examine a subset of this data corresponding to matched pairs of PA-ON and PA-OFF groups.

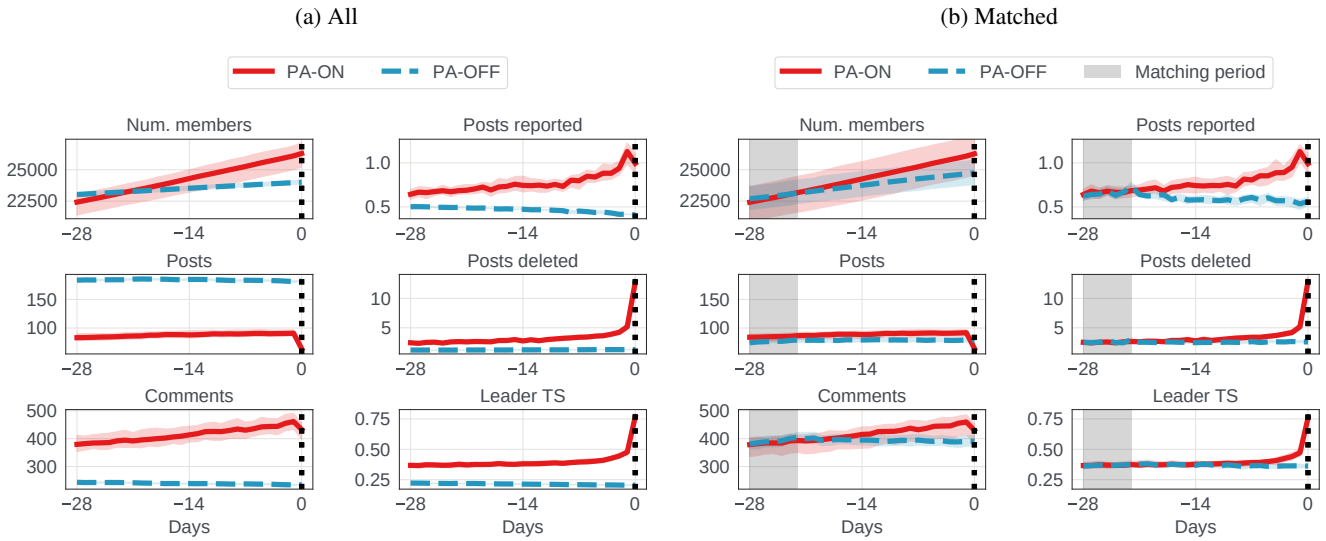

Figure 2: Average values for user activity- and moderation-related variables in the four weeks before communities enabled post approvals. Values for communities that enabled post approvals (PA-ON) are in red and those for communities that did not (PA-OFF) are in blue. For PA-OFF communities, day 0 corresponds to a pseudo-intervention date selected at random. We show trends for all communities in our dataset in *(a)* (PA-ON $n = 8{,}767$; PA-OFF $n = 224{,}635$) and for matched pairs of communities in *(b)* (PA-ON/PA-OFF $n = 8{,}643$). The period when matching was done is marked in gray. Error bars represent 95% CIs.

Fig. 2a shows the average value of each of the variables mentioned above. We found significant differences between PA-ON and PA-OFF groups that are consistent across the 28-day period considered ($p < 10^{-4}$ for independent t-tests conducted each day). PA-ON groups have significantly more reported and deleted posts, more comments, and fewer posts than PA-OFF groups. Leaders also spent more time in PA-ON groups than in PA-OFF groups.

Temporal trends also differed significantly between the two sets of groups: PA-ON groups experienced larger increases in all considered variables in the weeks before enabling post approvals. For moderation-related metrics, we observed a sharp spike on the day (or in the case of posts reported, on the day before) post approvals was turned on. These changes were not observed in PA-OFF groups.

The shifts in user activity (e.g., *Comments*) and in moderation (e.g., *Posts deleted*) before post approvals was turned on suggests that leaders enable the setting in response to new (and perhaps more chaotic) group dynamics. Specifically, the setting was commonly enabled in groups that were quickly growing and that experienced a surge in moderation-related events, which may have been the final straw that led administrators to enable post approvals. This finding is consistent with prior work suggesting that major changes in moderation (e.g., changing settings, creating new rules) happen in reaction to problems that emerge (Seering et al. 2019).

Matched analysis. While indicative, the previous analysis conflates two factors. Not only do PA-ON and PA-OFF communities differ in their baseline user and moderation activity, but they also differ in the way the studied variables change over time. Thus, observed differences in temporal trends may come from the fact that groups that adopt post approvals are different from those that do not. To more fairly compare these communities, we matched PA-ON and PA-OFF groups on user activity and moderation-related metrics between days -28 and -22. This matching ensures that communities were similar in the first week of the study period. Specifically, we performed one-to-one propensity score matching of PA-ON and PA-OFF communities using moderation and activity-related variables, as well as general group characteristics (e.g., *Group categories*). Details of the matching procedure can be found in Appendix A.

After performing this matching, we repeated the same analysis as in the previous subsection (Fig. 2b). Again, we found that PA-ON groups experienced a gradual increase in moderation-related traces which was accentuated right before post approvals was turned on — changes that were not observed in PA-OFF groups. Differences in user activity were subtler. Both PA-ON and PA-OFF groups experienced growth in the number of members, but this growth was higher for PA-ON groups. Moreover, PA-ON groups experienced a significant increase in the number of daily comments received, while comments received remained largely unchanged in PA-OFF groups. Last, in both PA-ON and PA-OFF groups the number of posts increased slightly during the 28 days considered.

Overall, the matched case–control analysis confirms that there are differences in the temporal trends of moderation and user activity of PA-ON and PA-OFF communities. Even when considering communities that were initially similar, PA-ON communities experienced larger increases in user and moderation activity prior to the day when they turned on post approvals.

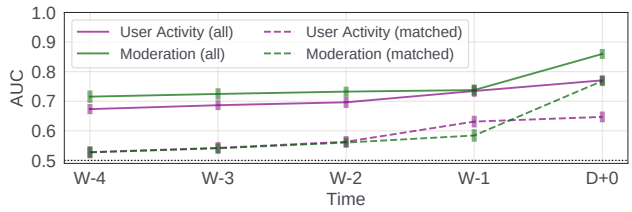

Figure 3: AUCs for classifiers trained to distinguish PA-ON and PA-OFF groups using data from different time spans and different sets of features. Error bars represent 95% CIs obtained through a 20-fold cross validation.

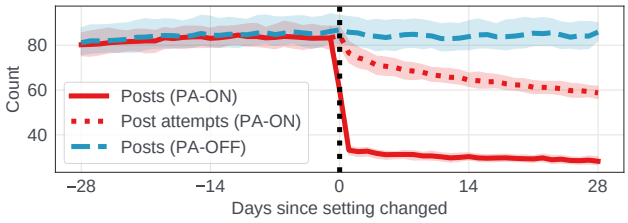

Figure 4: The average number of posts in PA-ON (solid red) and PA-OFF (dashed blue) communities. For PA-ON communities, the average number of post attempts after post approvals was enabled is shown in dotted red. Error bars represent bootstrapped 95% CIs.

Predicting if post approvals will be turned on in future weeks. While findings thus far indicate that both changes in user activity and in moderation precede the use of post approvals, is one a stronger indicator than the other? And how far in advance might they predict the adoption of the setting? To answer these questions, we examined if user activity or moderation can be used to distinguish PA-ON groups from PA-OFF groups.

We created two balanced samples of groups. The first sample (*all*) comprised 10,000 groups — half PA-ON and half PA-OFF. The second (*matched*) comprised the same PA-ON groups as in the *all* sample, while corresponding PA-OFF groups were obtained using one-to-one propensity score matching previously described. We considered two sets of features (group activity-related and moderation-related; cf. Table 1) and five time spans,[1] calculating the value of each feature in each time span by taking its average.

We trained Gradient Boosting classifiers to distinguish PA-ON and PA-OFF groups, varying the feature set and the time span used. Fig. 3 shows the AUC of the classifiers trained in each of the different settings (all vs. matched; moderation vs. activity) considering the features up to the time specified on the x-axis. For instance, the points shown above $x =$ W-3 correspond to the AUC of classifiers trained with features associated with W-4 and W-3.

For the *all* sample, we found that moderation features were more predictive in earlier weeks (W-4 to W-2). However, for $x =$ W-1, there was an increase in the AUC of the classifier trained with activity features. We observed a similar pattern in the *matched* sample: classifiers started with similar AUC values at $x =$ W-4 (user activity: 0.53 AUC vs. moderation: 0.53 AUC), but the classifiers trained with activity features saw a larger increase in performance at $x =$ W-1 (0.58 for activity vs. 0.63 for moderation). Including data up until the time of intervention, $x =$ D+0, the performance of classifiers trained with moderation features increased sharply (e.g., in the *matched* sample: 0.65 activity vs. 0.77 moderation). Overall, these results suggest that the adoption of post approvals was associated with gradual changes in user activity in the weeks before the adoption of the setting and sudden changes in moderation activity on the day the setting was enabled. Repeating this analysis but instead training classifiers using features belonging to each individual time span (e.g., using only W-1 vs. using W-4 to W-1 for prediction) resulted in qualitatively similar findings.

[1]Days -28 to -22 (week -4; W-4), -21 to -15 (week -3; W-3), -14 to -8 (week -2; W-2), -7 to -1 (week -1; W-1), and day 0 (D+0).

5 How do Post Approvals Shape Online Communities?

Having explored changes in user activity- and moderation-related signals that *precede* the adoption of post approvals, we now turn our attention to what happens *after* communities choose to adopt the setting. Here, we examine how user and moderation activity in online communities change following the adoption of the setting (**RQ2**). To do so, we matched communities that turned on post approvals (PA-ON) with similar communities that did not (PA-OFF) and validated the observed differences using regression. To obtain this matching, we performed one-to-one propensity score matching on moderation and activity-related variables as well as general group characteristics. Matching was done across the entire pre-intervention period (day −28 up to day 0 right before the intervention). See Appendix A for details. All analyses in this section were done at the group-level, with each group weighted equally.

Posts. First, we examined how posting behavior changed after the adoption of post approvals. In Fig. 4, we show both the number of posts that were actually published (*Posts*), but also, for PA-ON communities, the number of post attempts, i.e., post requests initiated by regular group members following the adoption of post approvals. For PA-ON communities, we found a significant decrease in the number of posts following the adoption of post approvals. The average number of posts went from roughly 80 posts a day pre-intervention to around 30 posts a day post-intervention, a decrease that was not observed in the matched set of PA-OFF communities.

What explains this decrease in posting? Was it because posts were being filtered? Or were people more hesitant to even post? To understand the relative contribution of these factors, we examined two corresponding quantities that make up the decrease in posting: 1) the difference between the number of posts in the control (PA-OFF) and treatment setting (PA-ON) (blue line vs. dotted red line); and 2)

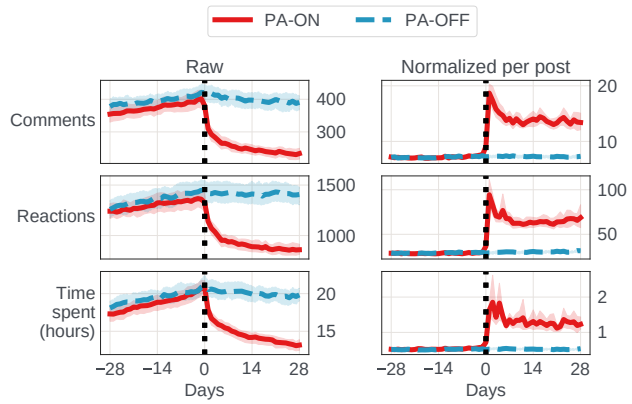

Figure 5: User activity-related signals before and after the adoption of the post approval setting. Signals are shown both in absolute terms (left) and normalized per number of posts (right). Error bars represent bootstrapped 95% CIs.

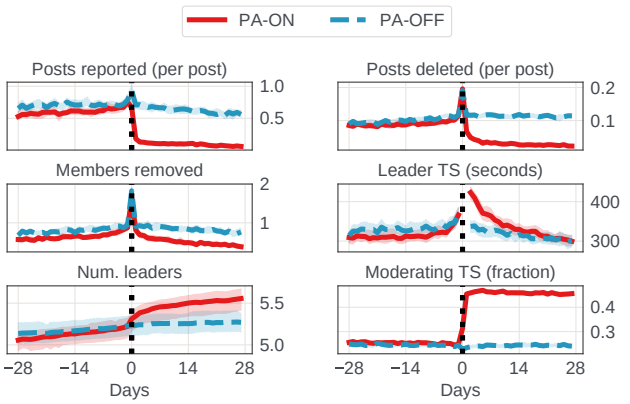

Figure 6: Moderation-related signals before and after the adoption of post approvals. Error bars represent 95% CI. We omit days 0 and 1 from the plot showing the *Leader TS*, as it contains a sharp peak.

the difference between the number of post attempts and actual posts in PA-ON communities (dotted red line vs. solid red line). We observed a gradual decrease in the average number of posts submitted (the first component mentioned above), from 76 posts a day on day 1 to 58 posts a day on day 28. However, the fraction of posts approved per community (the second component) remained largely stable at around 63% of posts (note that this was calculated without winsorization and per group, instead of dividing the overall averages; cf. Fig. 9, Appendix C). These findings suggest that post approvals reduce the number of posts by directly filtering out undesired posts, but also by reducing the likelihood of people to attempt to post in the first place.

Other user activity-related signals. Second, we looked at other user activity-related metrics (*Comments*, *Reactions*, and *Time spent*). These are shown in Fig. 5 both in absolute terms (first column) and normalized per number of posts (second column). Comparing PA-ON with PA-OFF communities, there was an absolute *decrease* but relative *increase* in all of these user activity metrics for PA-ON communities following the enabling of post approvals. In other words, there were fewer posts, but each post received, on average, more comments, reactions, and time spent.

For instance, before the intervention (day -1), PA-ON and PA-OFF communities received an average of around 402 and 421 daily comments and around 7.6 and 7.4 comments per post. (Note that although the two sets of communities are matched, their averages are not perfectly identical.) After day 0, when PA-ON communities enabled post approvals, the number of daily comments declined substantially, reaching an average of 233 daily comments on day 28. Meanwhile, the number of comments per post nearly doubled to around 13.4. This change was not observed in the matched PA-OFF groups.

Moderation-related signals. Third, in Fig. 6, we examined moderation-related metrics – *Posts reported*, *Posts deleted*, *Leader TS* and *Moderating TS* (cf. Table 1 for descriptions). We normalized the number of posts reported and deleted by the total number of posts, and moderating time spent by the total leader time spent. Recall that the moderating time spent encompasses activities such as responding to reported content and, notably, approving posts (if the post approvals setting is turned on).

Following day 0, PA-ON groups had fewer members removed and fewer posts reported/deleted (per post) than PA-OFF communities. For example, the percentage of posts reported decreased from around 0.75% pre-intervention to 0.10% post-intervention for PA-ON groups. This decrease was larger than that for matched PA-OFF groups (from 0.78% to 0.60%). Around the time that post approvals was enabled for PA-ON groups, time spent per admin increased substantially, likely because leaders were getting used to the new moderation style. However, by week 4, time spent per leader in PA-ON groups had returned to pre-intervention levels, though PA-ON groups had also tended to appoint new leaders. Examining the fraction of time spent in admin surfaces, leaders went from spending around 25% of their time using group moderation tools to around 45%, suggesting that leaders spent a substantial fraction of their time approving posts. While this increase appears large, group leaders in groups without post approvals may nonetheless be informally vetting posts by browsing posts in a group as they appear.

Regression analysis. Previously, we compared the user and moderation signals between PA-ON and PA-OFF communities before and after the post approvals are turned on after matching. Here, we performed a more rigorous analysis of the same signals under a regression framework.

We considered the average value of each variable of interest in week -4 (days -28 to -22) and week 4 (days 22 to 28). Then, for the variables in the post-intervention period, we estimate the impact of adopting post approvals using a linear model:

$$y = \boldsymbol{\alpha}\mathbf{X} + \beta\,\mathbf{1}_{[\text{PA-ON = True}]}, \qquad (1)$$

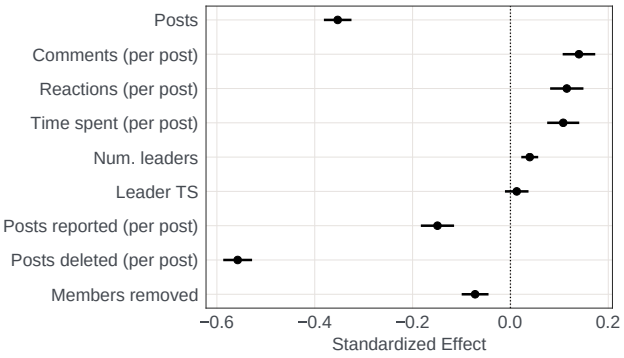

Figure 7: Standardized effect of enabling post approvals on user activity- and moderation-related variables. Error bars represent 95% confidence intervals. Data was not winsorized prior to this analysis.

where y represents the average value of one of the variables we studied in week 4 (i.e., after the intervention), e.g. post approvals, $\mathbf{X}$ represents an array with all the variables we did the matching with in week -4 (i.e., before the intervention), and β represents the coefficient associated with turning on post approvals, as it multiplies an indicator variable that equals 1 for PA-ON, and 0 for PA-OFF, communities.

To facilitate interpretability, we standardized the dependent variable y, so that the coefficients represent (pooled) standard deviations. The coefficient β captures the difference between our treatment and control groups in the matched setting. Since the coefficient is associated with an indicator variable, the effects reported represent the differences between PA-ON and PA-OFF groups in standard deviations. We report β for all outcomes of interest in Fig. 7.

This analysis largely confirms results shown in Fig. 4, Fig. 5, and Fig. 6. The use of post approvals was significantly associated with a reduction in the number of posts (-0.35 standard deviations) but an increase in the number of comments, reactions, and time spent per post (e.g., the number of comments per post increased by 0.14 SDs). Use of the setting was also associated with a decrease in the number of posts reported per post (-0.15 SDs), posts deleted per post (-0.55 SDs) and number of members removed (-0.07 SDs). Taken along with the previous analyses, these results suggest that the setting improves the quality of posts. Further, post approvals do not significantly increase the average time leaders spend in the group, although groups that enable the setting tend to increase the size of their leadership team (around 0.04 SDs).

6 Heterogeneity of Post Approvals

While post approvals change how online communities function, the effect of the setting may vary by group size as well as how it is used. For example, we found that, on average, community leaders do not spend more time in their communities following the adoption of the setting. Yet, this may not be the case for all groups: very large groups (with possibly hundreds of daily post attempts) may actually require more time from leaders after the setting is turned on, while smaller groups may require less time. Thus, we analyzed the impact of post approvals in communities with 1) different member counts, as well as in communities that 2) approved different fractions of the posts submitted (approval rate); and 3) took a different amount of time to approve posts (response time).

For each of the aforementioned variables (number of members; response time and approval rate), we divided PA-ON communities into 4 quartiles.[2] Then, we used the same regression setup depicted in Equation (1), but estimated the effect of post approvals separately for communities in each of the quartiles. This amounts to a linear model of the form:

$$y = \boldsymbol{\alpha}\mathbf{X} + \sum_{i=1}^{4} \beta_i \mathbf{1}_{[\text{PA-ON = True and Quartile = } i]}, \qquad (2)$$

where β_i is the effect for groups in a given quartile. We ran a different regression for each of the three setups described above (group size, approval rate, and response time).

All three factors had noteworthy interactions with user activity- and moderation-related signals (cf. Fig. 8).

Time spent by leaders. Leaders in larger groups (group size Q4), groups with lower approval rates (approval rate Q4), and groups with faster response time (response time Q4) spent more time in their communities following the adoption of post approvals (*Leader TS*: 0.15; 0.10; and 0.15 SDs). These trends were gradual across quartiles, and contrast with the overall null effect reported in Fig. 7. In other words, the moderation burden after enabling post approvals depends on the kind of group and on how community leaders proactively moderate the community.

User activity. Larger groups experienced larger decreases in the number of posts (e.g., Q4: -0.79 SDs) and larger increases in relative activity (e.g., 0.22 SDs for *Time spent* per post). Groups with a higher approval rate (Q3/Q4) experienced smaller decreases in the number of posts and smaller increases in relative activity. The higher the approval rate, the smaller the deviations were from PA-OFF matched communities. To a lesser extent, this was also observed for response time: the faster the response time, the smaller the deviations were. These results suggest that activity-related changes were greater in larger communities and that the strictness and speed of community leaders in approving or rejecting posts mediated changes in user activity.

Moderation. Regardless of group size, response time, or approval rate, the number of posts reported and deleted decreased significantly across quartiles in all three analyses. Other moderation-related metrics such as members removed also decreased in most cases. Overall, this suggests that the decrease in potentially problematic content following the enabling of post approvals is robust and that it holds even when groups are very large (e.g. -0.11 SDs for *Posts reported* per post for group size Q4) or when a vast majority of posts are

[2]Defined by three points: for approval rate: 0.45, 0.69, 0.84; for response time: 1.2, 2.7, 5.75 (hours); for group size: 2850, 8500, 24000 (rounded) Approval rate and response time were measured across the entire post-intervention period. Group size was measured on day 0.

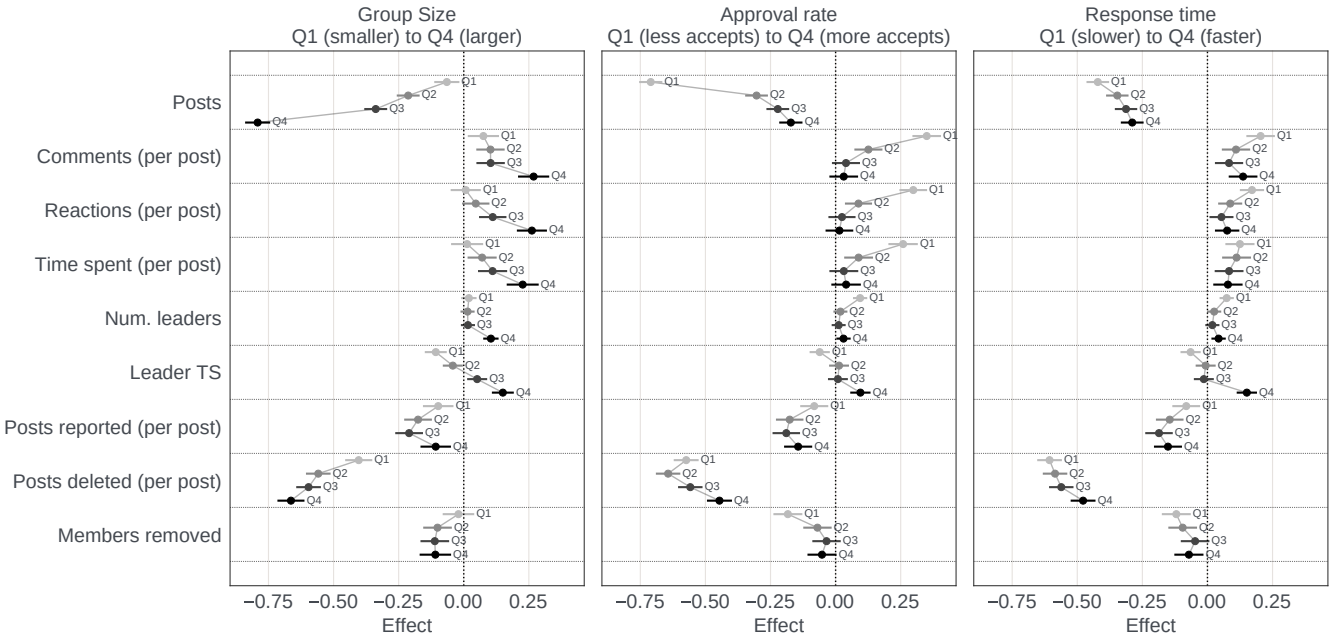

Figure 8: Effects of enabling post approvals for different stratifications of the data. Here, we report quartile-specific effects for group size (i.e., number of members in a group; first column), the post approval rate (i.e., percentage of posts that get approved in a group; second column) and the response time (i.e., average time taken to accept posts in a group; third column). Data was not winsorized prior to this analysis. Error bars represent 95% CIs.

approved (e.g. −0.15 SDs for approval rate Q4). In Fig. 4, we saw that post approvals changed the number of posts through both behavior change and through the filtering done by community leaders. Here, we see evidence that, regardless of the strictness of this filtering, the number of posts reported and deleted (normalized per post) decreases.

7 Discussion and Conclusion

In this work, we presented a large study of post approvals in Facebook Groups, examining both their adoption and their subsequent impact. Post approvals was adopted after changes in the groups' dynamics in the weeks prior: user activity and moderation increased in the weeks before the setting is enabled, and, on the day when the setting was turned on, there was often a surge in moderation activity. After the setting is adopted, communities become, on average, centered around fewer posts that receive more comments and reactions, and which users interact with for longer. These posts were less likely to be reported, and members were less likely to be removed, which suggests an increase in the quality of the discussions happening in the group. However, the strength of these effects varied with group size and with how proactive moderation was carried out – e.g. in larger groups, leaders spent more time in the group after the setting was enabled, while in smaller groups, they spent less time. Overall, the findings provide preliminary insight on how proactive moderation may improve online information ecosystems: by adding participation friction in online communities, post approvals elicit behavior change (cf. Fig. 4).

A limitation of our work is that we focus on a limited set of community-level analyses over a short period without considering spillover effects. This limitation suggests several potential extensions. First, future work could analyze how post approvals impact the participation of different kinds of users (including leaders); for instance, the setting may disproportionately affect highly-active users or may affect newcomers more than veteran members of a group, discouraging the former from participating. Related, future work could examine the reasons for the decrease in post attempts – to what extent do post approvals discourage lower-quality posts vs. all posts? Second, future work could investigate the impact of post approvals in the long run. In other words, do the patterns we observe here continue for months or even years? How do communities evolve with and without the setting? Third, future work might examine spillover effects across different communities. If two communities have many overlapping members and one adopts post approvals (as well as stricter moderation practices), does this influence the behavior of users in the other community which did not adopt the setting? Fourth, future work could explore other community-level variables, such as within-group friendship network properties. This work may include examining if these variables can help explain the adoption of post approvals (as in Sec. 4), if they change suddenly after post approvals are enabled (Sec. 5), or if the effect of post approvals is heterogeneous across these variables (Sec. 6).

Last, we note that our analysis was limited to Facebook Groups. Adapting the methodology here to explore participation controls studied qualitatively in other platforms such

as "chat modes" in live streaming platforms (Seering, Kraut, and Dabbish 2017) or software such as Reddit's *AutoModerator* (Jhaver et al. 2019) remains future work. As argued by Kraut and Resnick (2012), participation controls are "design levers" that shape how people connect with others in online communities. Thus, understanding how they work may inform the design of better-governed online spaces.

A Propensity Score Matching

Throughout the paper, we performed one-to-one propensity score matching (PSM) of PA-ON and PA-OFF communities on the group-level variables. We considered all time-varying variables in Table 1 (under the headers "activity-related" and "moderation-related") as well as all time-invariant variables (under the header "group characteristics"), with the exception of *Group categories*. We did not match on the latter, but found good covariate balance nonetheless, cf. Appendix B. Matching was done using nearest-neighbor matching without replacement, as implemented in Ho et al. (2011). Propensity scores were obtained using a Gradient Boosting Classifier, as implemented in Pedregosa et al. (2011). For both matching procedures and for all continuous variables, we obtained absolute standardized mean differences (SMDs) smaller than the commonly-used 0.1 threshold that indicates imbalance (Austin 2011).

What leads to the adoption of post approvals? For the analyses done in Sec. 4, we performed PSM using an absolute caliper of 0.5, discarding 125 PA-ON groups (1.4%) for which we were not able to find good matches. For time-varying variables (under the headers "activity-related" and "moderation-related" in Table 1), we considered three days of interest (days -22, -25, and -28). In other words, each covariate-day pair corresponds to a distinct feature used by the classifier to obtain the propensity scores (e.g. posts on day -22, posts on day -25, and posts on day -28). Day 0 was when the intervention took place for PA-ON group (i.e., when they enabled the post approvals setting), and it was chosen at random for PA-OFF groups. Covariate balance for this matching is shown in Fig. 12.

How do post approvals shape online communities? For the analyses done in Sec. 5 and Sec. 6, we performed PSM with a caliper of 0.0075, discarding 1426 PA-ON groups (16%) for which we were not able to find good matches. For time-varying variables (under the headers "activity-related" and "moderation-related" in Table 1), we considered five days of interest (days -28, -21, -14, -7 and -1). Additionally, we considered the number of posts, comments, reactions, deleted posts, reported posts and members removed on the day of the intervention (day 0) measured up to the hour when the intervention (or pseudo-intervention) was introduced. These were all variables that we were able to measure hourly. Matching on the day of the intervention was done on an hourly basis as PA-ON communities have periods of activity with and without the post approval setting. Covariate balance for this matching is shown in Fig. 13.

Additional observations. We clarify a couple of decisions regarding the propensity score matching procedure:

- As mentioned in Sec. 3, to obtain candidate PA-OFF groups for matching, we used a random sample of 50% of all communities that had over 128 members and at least one post and one comment over any 7-day period. This reduces the matching space, as we could have used 100% of all communities. Yet, empirically, using more than 50% of the sample harmed the propensity matching score matching capacity to balance the variables of interest as the class imbalance was too extreme. Even with a 50% sample, before matching, the number of PA-OFF groups ($n = 224{,}635$) outnumbered PA-ON groups ($n = 8{,}767$) by approximately 25 to 1.
- Differences between the PSM done for Sec. 4 vs. for Sec. 5 and Sec. 6 are as follows:
  - Different calipers were used, as achieving covariate balance was harder for the matching in Sec. 5/6.
  - Different dates were considered for time-varying variables (-22, -25, -28 vs. -1, -7, -14, -21, -28).
  - The matching used in Sections 5 and 6 additionally included variables from the day when post approvals was turned on, measured up to the very hour when the setting was changed (cf. Fig. 13b)

B Group Topics

To ensure that PA-ON and PA-OFF groups were topically comparable after propensity score matching, we used Empath (Fast, Chen, and Bernstein 2016)'s lexical categories. Lexicons have been used in causal inference by Saha et al. (2019) and by Sridhar and Getoor (2019). We translated group titles and descriptions into English and measured the occurrence of words matching each of the 194 default Empath categories (e.g., *work*, *celebration*, *writing*, etc.). Even without explicitly matching groups by Empath category word frequency, propensity score matching yielded good covariate balance – the standardized mean difference for all 194 categories was below 0.1,[3] suggesting the two sets of groups had similar titles and descriptions. Fig. 12c and Fig. 13d show the covariate balance of the top 20 most common Empath categories before and after PSM.

C Additional Plots

We provide a couple of additional plots as sanity checks:

- Fig. 10 shows the distribution of hour and day of the intervention for the matching done for Sections 5 and 6.
- Fig. 11 reproduces Fig. 4 using the median instead of the (winsorized) mean.
- Complementing Fig. 4, Fig. 9 shows the fraction of posts submitted that were approved.

[3]Except for the categories *communication* and *internet* in the PSM done for Sec. 5, where they had SMDs equals to 0.103/0.107.

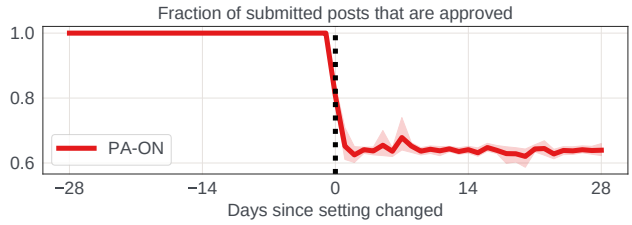

Figure 9: For PA-ON groups, we show the fraction of posts submitted that were approved.

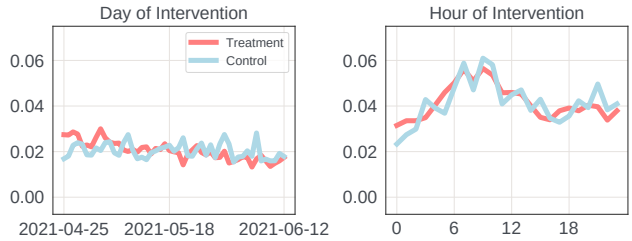

Figure 10: We show the day (left) and hour (right) of the intervention (for PA-ON groups) and pseudo-intervention (for PA-OFF groups) after the matching done in Sections 5 and 6. Note that the hours are shown in UTC+0.

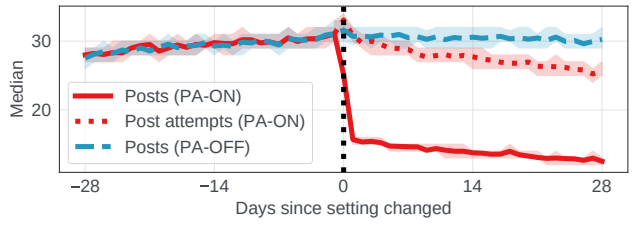

Figure 11: This figure repeats the analysis done in Fig. 4 using the median instead of the winsorized mean.

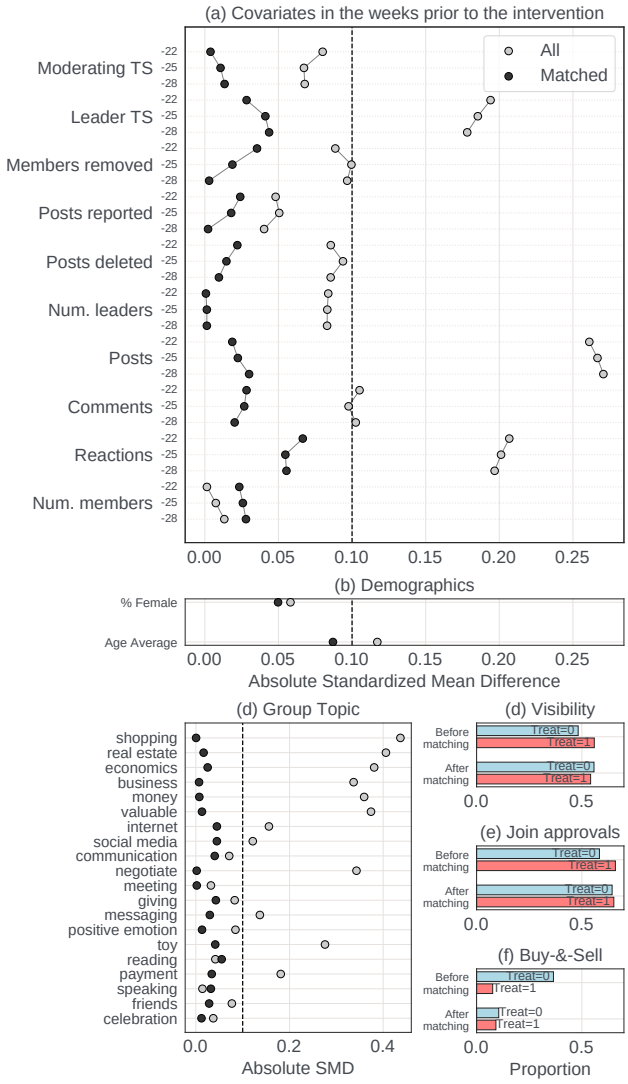

Figure 12: Covariate balance for matching done in Sec. 4. *(a)* Absolute standardized mean difference (SMD) for all time-varying variables considered in days −28, −25 and −22. *(b)* SMD for demographic-related variables. *(c)* SMD for the top 20 most popular group categories (from Empath). *(d-f)* covariate balance pre- and post-matching for the three binary variables considered in the matching.

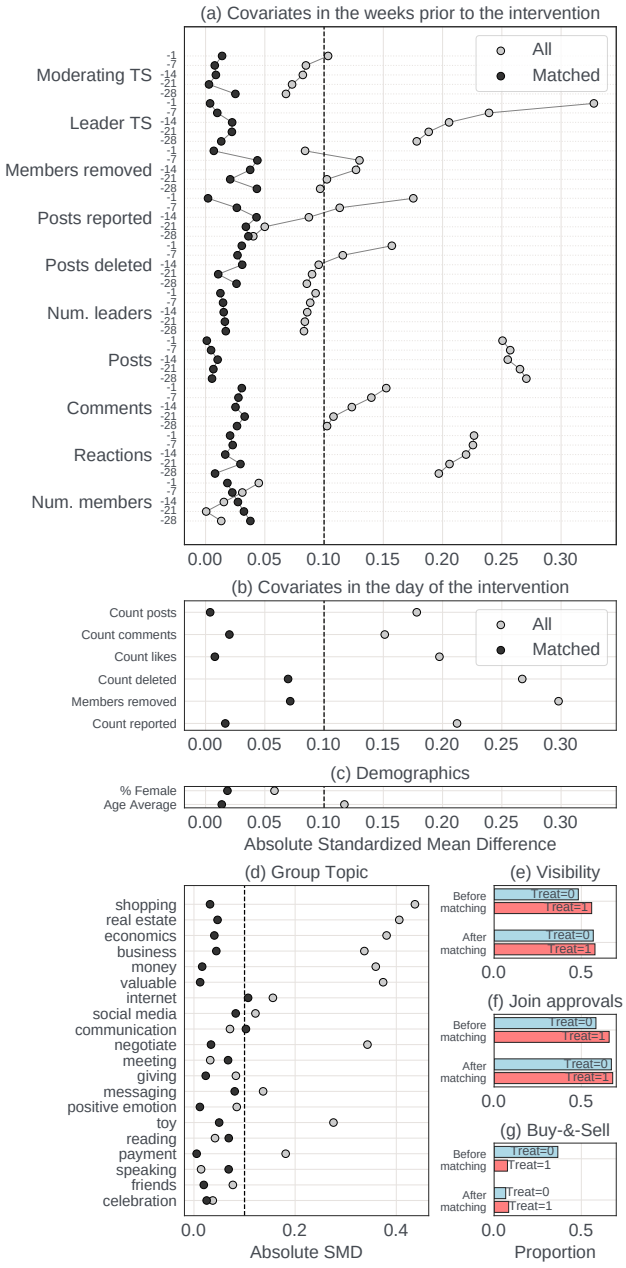

Figure 13: Covariate balance for matching done in Sec. 5. *(a)* Absolute standardized mean difference (SMD) for all time-varying variables considered in days -1, -7, -14, -21, and -28. *(b)* SMD for variables measured in the day of the intervention. *(c)* SMD for demographic-related variables. *(d)* SMD for the top 20 most popular group categories (from Empath). *(e-g)* covariate balance pre- and post-matching for the three binary variables considered in the matching.

References

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3): 399–424.

Bucher, T.; and Helmond, A. 2018. The Affordances of Social Media Platforms. In *The SAGE Handbook of Social Media*, 233–253. Sage Publications.

Cosley, D.; Frankowski, D.; Kiesler, S.; Terveen, L.; and Riedl, J. 2005. How oversight improves member-maintained communities. In *Proceedings of the 2005 CHI Conference on Human Factors in Computing Systems*, 11–20.

Dosono, B.; and Semaan, B. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–13.

Ducheneaut, N. 2005. Socialization in an open source software community: A socio-technical analysis. *Computer Supported Cooperative Work (CSCW)*, 14(4): 323–368.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 4647–4657.

Filgueiras, F.; and Almeida, V. 2021. *The Digital World and Governance Structures*. Springer.

Gillespie, T. 2018. *Custodians of the Internet*. Yale University Press.

Ho, D. E.; Imai, K.; King, G.; and Stuart, E. A. 2011. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42(8): 1–28.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. In *ACM Transactions on Computer-Human Interaction (TOCHI)*, volume 26, 1–35.

Jhaver, S.; Bruckman, A.; and Gilbert, E. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. In *Proceedings of the ACM on Human-Computer Interaction*, CSCW, 1–27.

Kiene, C.; Monroy-Hernández, A.; and Hill, B. M. 2016. Surviving an "Eternal September" How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1152–1156.

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

Lo, C. 2018. *When all you have is a banhammer: the social and communicative work of volunteer moderators*. Ph.D. thesis, Massachusetts Institute of Technology.

Mejtoft, T.; Hale, S.; and Söderström, U. 2019. Design Friction. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, 41–44.

Meta. 2022. Using post approvals in your group. https://www.facebook.com/community/establishing-

membership-and-rules/how-to-use-post-approvals-to-speed-up-moderation-in-facebook-groups/. Accessed: 2022-04-03.

Papadamou, K.; Papasavva, A.; Zannettou, S.; Blackburn, J.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; and Sirivianos, M. 2020. Disturbed Youtube for kids: Characterizing and detecting inappropriate videos targeting young children. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 522–533.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825–2830.

Saha, K.; Sugar, B.; Torous, J.; Abrahao, B.; Kıcıman, E.; and De Choudhury, M. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 440–451.

Schlesselman, J. J. 1982. *Case-control studies: design, conduct, analysis*, volume 2. Oxford University Press.

Schmidt, A.; and Wiegand, M. 2019. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain*, 1–10. Association for Computational Linguistics.

Seering, J. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. In *Proceedings of the ACM on Human-Computer Interaction*, CSCW2, 1–28.

Seering, J.; Kraut, R.; and Dabbish, L. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on Computer Supported Cooperative Work and Social Computing*, 111–125.

Seering, J.; Wang, T.; Yoon, J.; and Kaufman, G. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7): 1417–1443.

Sridhar, D.; and Getoor, L. 2019. Estimating Causal Effects of Tone in Online Debates. In *International Joint Conference on Artificial Intelligence*.

Wilcox, R. R. 2011. *Introduction to robust estimation and hypothesis testing*. Academic Press.

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