

Trajectories of Blocked Community Members: Redemption, Recidivism and Departure

Jonathan P. Chang
Cornell University
jpc362@cornell.edu

Cristian Danescu-Niculescu-Mizil
Cornell University
cristian@cs.cornell.edu

ABSTRACT

Community norm violations can impair constructive communication and collaboration online. As a defense mechanism, community moderators often address such transgressions by temporarily blocking the perpetrator. Such actions, however, come with the cost of potentially alienating community members. Given this tradeoff, it is essential to understand to what extent, and in which situations, this common moderation practice is effective in reinforcing community rules.

In this work, we introduce a computational framework for studying the future behavior of blocked users on Wikipedia. After their block expires, they can take several distinct paths: they can reform and adhere to the rules, but they can also recidivate, or straight-out abandon the community. We reveal that these trajectories are tied to factors rooted both in the characteristics of the blocked individual and in whether they perceived the block to be fair and justified. Based on these insights, we formulate a series of prediction tasks aiming to determine which of these paths a user is likely to take after being blocked for their first offense, and demonstrate the feasibility of these new tasks. Overall, this work builds towards a more nuanced approach to moderation by highlighting the tradeoffs that are in play.

ACM Reference Format:

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*, May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3308558.3313638>

1 INTRODUCTION

The health of online communities is often threatened by individuals that violate their norms and rules [24]. Communities have limited defense mechanisms against such offenses, ranging from the deletion of problematic comments [4], to

temporarily blocking the perpetrators, to outright excluding them [22]. Such actions, however, come at the cost of potentially alienating valuable community members and raise issues of freedom of speech, fairness and bias [22, 25, 45]. Considering their potential negative impacts, it is important to understand to what extent, and in which situations, these moderation mechanisms are effective at enforcing community norms.

In this work we take a first step towards answering these questions by focusing on *temporary blocks* in the Wikipedia community of editors. Unlike other moderation actions, these have the explicitly stated purpose of “detering any future possible repetitions of inappropriate conduct” by the perpetrators, while reintegrating them in the community (thus the temporary nature of the block). However, following their initial block, nearly half of first-time offenders either recidivate—18% of established members violate the rules again within 6 months of the block—or straight-out abandon the community—30% of established members depart within 6 months (a third of which depart during the block). This discrepancy between the stated goal and the empirical outcomes of temporary blocks motivates an investigation into the factors that determine the likely trajectory a community member will take after their block expires: redemption, recidivism or departure.

We ground our investigation in offline theories of deterrence and defiance [46], which point to at least three broad classes of factors determining future compliance: those pertaining to the characteristics of the perpetrator [48], those pertaining to the actual severity of the punishment [16, 27] and those pertaining to whether the perpetrator perceives their treatment as fair [40]. While relying on these theories to guide our selection of factors and potential confounds to consider in our analysis, we refrain from drawing any direct parallels between offline and online rule-enforcement strategies given their distinct nature and goals.

Individual characteristics. Behavior of community members and their adherence to community norms has been shown to vary according to their level of involvement with the community [6, 10, 18, 42]. We find that individuals that exhibit a high level of community involvement (e.g., have interacted with more users over a longer period of time) are less likely to abandon the community during a temporary

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6674-8/19/05.

<https://doi.org/10.1145/3308558.3313638>

block. On the other hand, these individuals are also more likely to recidivate soon after the block. This suggests that users with high community involvement are more immune to the effects of blocks, both undesired (departure) and desired (reforming).

Block duration. Existing literature on offline rule compliance varies in their accounts of how the nature of the punishment determines the likelihood of recidivism [49]. The traditional (institutional) perspective is that the severity of the punishment is the main factor determining compliance [16, 27]. A contrasting (normative) perspective, however, argues that perhaps a more important factor is whether the perpetrator regards the punishment as being fair and justified [32, 48]. In the context of the Wikipedia community, we find that even though longer—and thus more severe—blocks come at the cost of higher departure rates, they do not appear to have any effect on recidivism.

Perceived fairness. To capture whether an individual perceives their initial block as fair, we leverage a mechanism specific to online moderation: block appeals. Blocked individuals can contest a block by opening a discussion with Wikipedia moderators during the block. In turn, moderators have the option to lift the block early—i.e., to *unblock* the individual—as a result of these discussions. This process provides us with a glimpse into a blocked individual’s perception of fairness: we can analyze their language as well as the outcome of the appeal.

Using this methodology, we find that users who perceive the block to be unfair have an increased rate of recidivism, while those that acknowledge their wrongdoing and apologize have a decreased rate. Furthermore, even though the block duration itself does not appear to affect recidivism, the post-hoc reduction of this duration by a moderator drastically diminishes recidivism rates, perhaps by improving the user’s perception of the moderation system as being fair and reasonable. Combined, these results suggest that the effectiveness of temporary blocks is not so much dependent on their actual severity, but rather on the blocked user’s perception of how appropriate the block is.

To summarize, in this work we:

- investigate the effectiveness of temporary blocks by considering possible trajectories an individual can take after their block expires;
- reveal factors that are indicative of these future trajectories, reflecting both the engagement patterns of the blocked individual, as well their perception of the fairness of the block;
- analyze the relative importance of each of these factors by using them in a series of forecasting tasks aiming to determine which path a user is likely to take after their first block.

More broadly, we propose a computational framework for analyzing the outcomes of a common moderation practice in online communities. In doing so, we expose the underlying tradeoffs and provide the means to further explore factors that mediate them.

2 BLOCKS ON WIKIPEDIA

While Wikipedia is primarily known as an online encyclopedia, it also plays host to a vibrant community of editors who continually write new articles and improve existing ones. To support this community, Wikipedia has a feature known as *talk pages*: special pages on which editors can discuss a particular article or Wikipedia policy, or simply unwind with casual conversation. Every Wikipedia article has an associated talk page, on which editors can discuss proposed edits to the article. Similarly, every registered user has an associated *user talk page*, where other editors can engage directly with the respective user on issues pertaining to their general behavior. This is where the social functions of the community surface: editors discuss inappropriate behavior and norm violations, ask for support, offer advice or encourage each other. In this work we use the complete conversational history between English Wikipedia editors on both article and user talk pages. With over 90 million conversations between 4 million users on 24 million talk pages, this is one of the largest collections of public conversations [20].

Like many online communities, Wikipedia has a moderation system aimed at imposing community rules. What makes Wikipedia particularly suitable for our proposed investigation is that its rules apply uniformly to the entire community and moderation is entirely transparent to the public [13]. This is unlike other large communities, such as Reddit, where rules can vary drastically between sub-communities and where aspects of the moderation remain private [4].

Blocking mechanism. Wikipedia moderators (formally referred to as administrators, though we will continue to use the term “moderator” as we are focused on their role in maintaining community norms) are elected from among community members through public elections [1, 30]. Among other privileges, moderators have the ability to block users from making any edits or comments. They can either block a user indefinitely, in cases of accounts that are clearly used for bad-faith purposes, or temporarily if they see the user as a potentially valuable member of the community that has momentarily gone awry. In fact, the stated purpose of temporary blocks is “to deter any future possible repetitions of inappropriate conduct”, thus keeping the door open for reintegration after the expirations of the blocks.¹ Our focus in this work is on temporary blocks (henceforth simply called blocks), and their effectiveness in reforming blocked users.

¹https://en.wikipedia.org/wiki/Wikipedia:Blocking_policy

Total number of blocks	104,245
Number of blocked users	72,332
Number of blocking moderators	1,706
Users with block reason in disruption subset	21,043
Users with first block in disruption subset	18,909
Users remaining after minimum activity filters	6,026

Table 1: Statistics of the blocked users dataset.

When temporarily blocking an individual, moderators are required to notify the user by posting a message on their user talk page. This notification includes the reason for the block and its duration. The duration of the block is at the discretion of the moderator, and can vary greatly—the standard deviation of block durations is 194 days. However, this high variance is largely the result of a small set of extreme outliers, and in practice most blocks are very brief—the median duration is 1 day.

For the duration of the block, the targeted user may not contribute to Wikipedia articles or make any comments on talk pages, with the sole exception of their own user talk page, where they are allowed to discuss the block.

Block reason. When moderators block a user, they must cite a reason that reflects one of the community rules. Wikipedia groups these reasons under two broad categories: *protection* and *disruption*. The former concerns behaviors that may put Wikipedia editors or Wikipedia itself at risk, such as making legal threats, releasing personal information or violating copyright law. The latter concerns the health of the community—or, as Wikipedia phrases it, maintaining a “civil, collegial atmosphere”. We identify a subset of four such disruption block types relating to breaches of community norms (rather than to legal concerns specific to Wikipedia): personal attacks and incivility, harassment, edit warring, and disruptive editing. These reflect norms that can be broken even by established and valuable members of the community, unlike spam or vandalism which most often come from illegitimate, disruption-only accounts.

Blocked users dataset. We extract information about the nature of the blocks from the *block log*, a centralized and public record of all blocks that have ever happened on Wikipedia containing over 136,000 actions (including both blocks and manual block modifications) at time of collection. We automatically extract from this log the block’s duration,² its type, its date,³ and the involved users (blocking moderator and blocked individual). We seek to focus our analysis on actual

members of the community as opposed to disruption-only accounts, because such accounts are made in bad faith and hence have no reason to reform once caught [15]. To this end, we discard accounts whose first block was not one of the four disruption block types previously described, and impose a minimum activity before the first block: at least one month and eight comments. Finally, we join the resulting block meta-data with the conversational record. Details of the resulting dataset are summarized in Table 1. We make this processed dataset publicly available⁴ to encourage the study of moderation in a community where all activity is by design fully transparent and public.⁵

3 TRAJECTORIES AFTER THE BLOCK

After their block expires, a user can take several distinct trajectories: they can reform and adhere to the rules, but they can also recidivate, or straight-out abandon the community. Here we formally define these (un)desired outcomes and reveal their relative prevalence.

Throughout, we focus on the very *first* block a user receives, thus discarding any potential lingering effects of previous blocks and putting all users in our dataset on equal ground at the time of investigation. We consider a maximum future horizon of 6 months, and thus are constrained to discard users whose initial block takes place less than 6 months before the end of our dataset (June 2018) since for these we cannot know what the future holds. This is a common solution for limited horizon in longitudinal studies [10].

Departure. Wikipedia and other collaborative communities have traditionally struggled to retain their members [17]. Thus, the risk of alienating users needs to be taken into consideration when evaluating the efficacy of moderation. This is particularly true for temporary blocks, which by their very nature aim at re-integrating the targeted user into the community rather than discouraging them to contribute again (for which a moderator would use a permanent block).

We consider a user to *depart* from the community at the time of their last comment on community talk pages; we use this instead of their last edit on an article as the indicator of quitting because we are specifically interested in their engagement with the community. We impose the same 6 month maximum future horizon as we did when considering recidivism. Out of all 6,026 blocked users selected for this study (Section 2), 30% depart in the 6 months following their first block, and 10% depart *during* the block itself.

To illustrate the connection between departure rates and the block itself, Figure 1 compares the probability of a user

²We merge consecutive blocks and account for post-hoc reduction in block duration.

³In our analysis we discard blocks happening in the first five years of the community, allowing for Wikipedia community norms to stabilize [13, 17].

⁴As part of ConvoKit (convokit.cornell.edu)

⁵All editors are expressly informed, during all edits, that “Work submitted to Wikipedia can be edited, used, and redistributed—by anyone”, where “work” is defined as content on any Wikipedia page, talk pages included. See https://en.wikipedia.org/wiki/Wikipedia:Ownership_of_content.

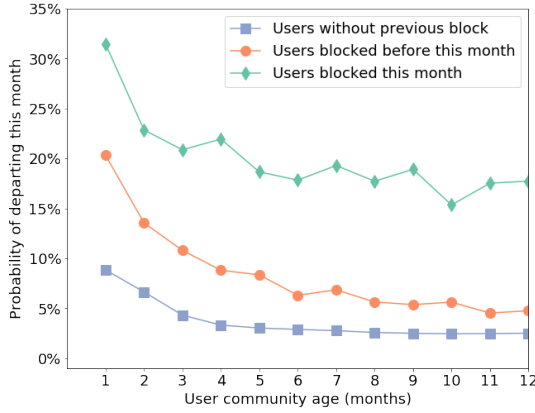


Figure 1: Probability of a user quitting at a given community age (in months), given that they were never blocked before (blue squares), blocked anytime prior to that month (orange circles) or blocked in that same month (green diamonds). The trends continue beyond the 12 months illustrated here.

abandoning the community at different stages in their community life (i.e., time since their first activity), conditioning on whether (and when) they were blocked before. We first note that departure is not independent from blocking events (compare blue squares with orange circles): users are much more likely to depart if they were blocked before, especially in the first months of their life in the community. But even controlling for the fact that a user was previously blocked (compare orange circles with green diamonds), they are much more likely to depart during the month of their block. This strong recency effect shows that blocking and departure are events that are also temporally connected.

Recidivism. A previously-blocked user *recidivates* if they are blocked again for any other breach of community rules, and *reforms* if they never get blocked again. There is a wide variance in how long it takes recidivist users to re-offend. While some get blocked again within a week, others remain “clean” for years before re-offending. Only 47.7% of blocked users in our dataset have no (blocked) offenses after their first block. Among the users who do have a second known offense, the median time to reoffense is 32 days.

To study recidivism at different temporal scales, we consider two categories of users: *long-term recidivists* are those who re-offend within the next 6 months after their initial block, and *short-term recidivists* are the subset who re-offend within the first week after their first offense. Out of all blocked users selected for this study, 38.5% breach the norms again in the 6 months following their initial block; and 15.4% recidivate in just one week after the initial block.

To put recidivism rates into perspective and differentiate the phenomenon from general rule violations, Figure 2 compares the probability of a user getting (re)blocked at different

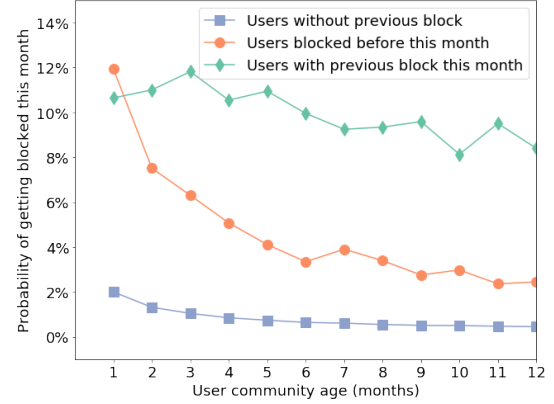


Figure 2: Probability of a user getting blocked at a given community age (in months), given that they were never blocked before (blue squares), previously blocked anytime prior to that month (orange circles) or previously blocked in that same month (green diamonds).

stages in their community life, conditioned on whether (and when) they were already blocked or not. Perhaps unsurprisingly, individual block events are not independent: the probability of being blocked is much higher for individuals who were already blocked before than for those that were not, even after controlling for community age (compare blue squares with orange circles). We also observe a dramatic decrease in probability of recidivism during the first year of a user’s community age. Importantly, this effect is specific to recidivism, rather than just reflecting a tendency to get blocked in general—the probability of getting blocked for the first time (blue squares) stays relatively flat over time. Finally, as in the case of departure, we also observe a recency effect (compare green diamonds with orange circles), showing that consecutive blocks are temporally related events.

Having established a relation between blocks and departure or recidivism, we now move towards a more thorough analysis of the factors that are in play. We will explore two broad classes of factors inspired by theories of deterrence and defiance [46]. The first class of factors concerns the characteristics of the perpetrator (Section 4): some types of individuals might simply be more likely to (re)offend or depart. The second class of factors concerns the actual context of the block: its actual severity and how the perpetrator perceives it (Section 5).

4 USER CHARACTERISTICS

We have already seen that community age is an important factor mediating the likelihood of departure and (re)offense: overall, newcomers are more likely than experienced users to follow one of these undesired paths after being temporarily blocked. Here we explore how these outcomes relate to

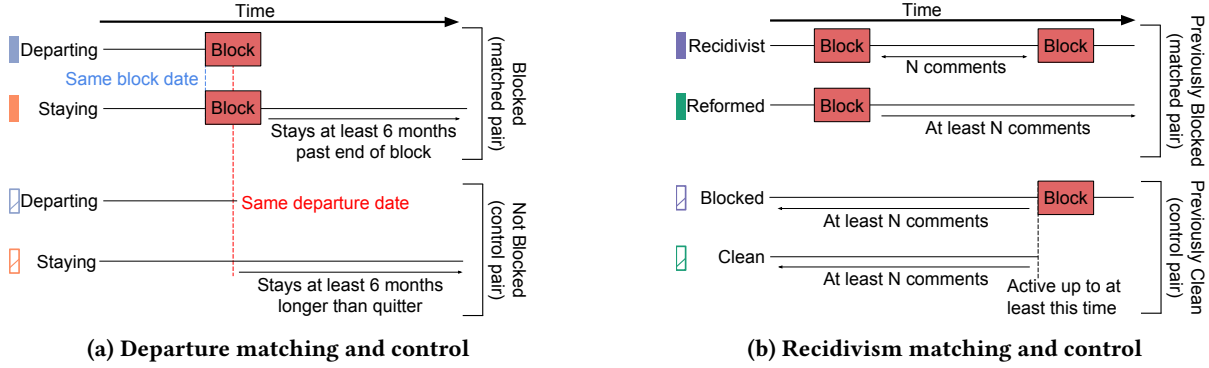


Figure 3: We matched comparable blocked users with opposite outcomes: (a) departs or not during the first block; (b) offends or not after the first block. For each matched pair, we also select a control pair of users (with the same opposite outcomes) who were not blocked.

other user characteristics, focusing on factors indicating a member’s level and type of engagement in the community.

We consider two broad types of features: how active the user is and how broad their social connections are in the community. To foreshadow our findings, we find that individuals that are more active in the community are overall more immune to blocks: they are both less likely to depart during a block, but also less likely to reform. However, the way an individual spreads their activity is also informative: if their activity is directed at a broader set of users they are more likely to either depart or reform.

Activity level. We consider both the number of comments a user contributes to other users’ talkpages (contributed activity), as well as the number of comments they receive from other users on their own talkpage (received activity) before the block. To disentangle a user’s activity from their community age, we normalize these measures by the number of days between the user’s first recorded activity in the community and the time of the block.

Activity spread. An active individual can either engage a broad audience, or instead concentrate on a narrow set of individuals. We capture the *received activity spread* of an individual by counting the number of unique users who wrote comments on their talk page before the block, then normalizing it by the total number of such comments. A higher value of this ratio indicates brief interactions with a relatively large number of peers, while conversely, a smaller ratio indicates extended interactions with a relatively small group of peers. Analogously, we define *contributed activity spread* as the ratio between the number of other users’ talk pages to which an individual contributed comments and the number of such comments they contributed.

Departure. To make meaningful comparisons between users who depart *during* their first block and those who don’t,

hence eliminating trivial confounds, we use a matching technique inspired by causal inference methods [43]. As illustrated in Figure 3a (top), we pair each departing user with a user whose first block took place on the same date (within 1% tolerance) but who stayed active for at least 6 more months (our maximum future horizon). This procedure results in 1,156 paired departing and staying users.

As a secondary question, we also want to understand whether any differences we observe are specific to blocked individuals, or generally characterize community departure. To this end, for each matched pair of blocked departing-staying individuals, we construct a control pair of departing-staying individuals that were not blocked. This control pair consists of a non-blocked departing individual with the same departure date (within tolerance) as the blocked departing individual, and a staying individual who remained active at least 6 months past that date. This process is also illustrated in Figure 3a (bottom).

Figure 4 summarizes how our engagement factors relate to departure during the block (solid bars). Departed users receive and contribute less daily activity in their life before the block than non-departed users. The activity they do receive comes from a broader set of users, potentially signaling they are not integrated into a tight social group.⁶ This is however not true for the spread of the activity they contribute, which is similar between departing and staying users.

We note that these relations between activity and departure are specific to blocked individuals (compare hashed bars in Figure 4). Departing users in our matched pairs do not have significantly different activity levels or activity spreads than their staying counterparts. This finding reinforces the

⁶These trends continue to hold, and remain significant, if we consider only the top half of active users in our comparison, thus separating them from the previously described trends in raw activity counts.

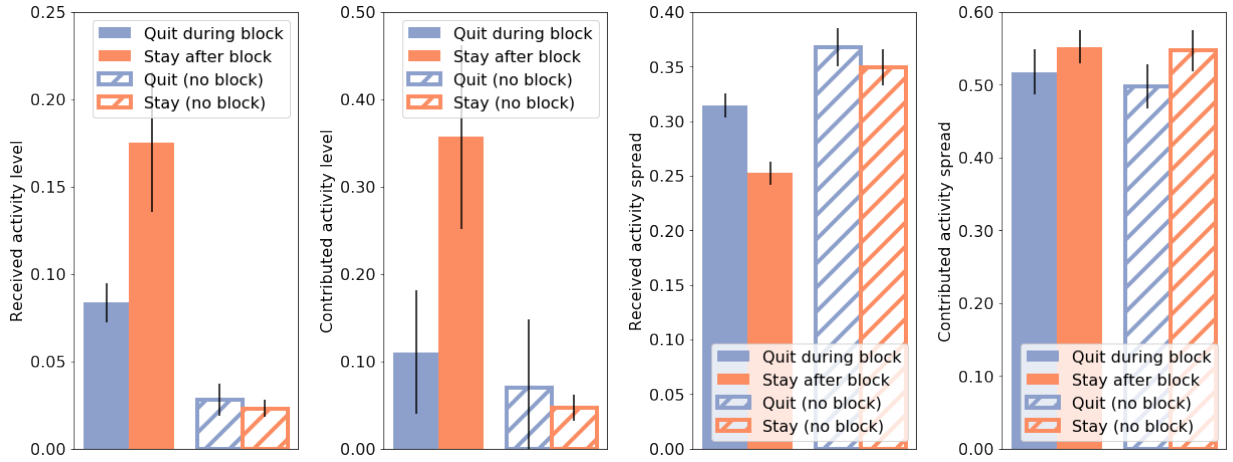


Figure 4: Difference between the community engagement patterns of departing and staying users. From left to right: receiving activity level, contributed activity level, receiving activity spread, contributed activity spread. Solid bars compare blocked departing and staying users; hashed bars compare non-blocked departing and staying users. Throughout errorbars indicate 95% binomial proportion confidence intervals.

notion that departure during a block is a distinct phenomenon from departure in general.

Recidivism. Having shown that activity level and activity spread are indicative of departure, we now explore whether they might also signal recidivism. As before, we use matching to eliminate trivial confounds. In this setting, the primary concern is that higher activity might trivially result in higher likelihood of recidivism merely because users with more activity tend to write more comments, which means they have more things they could be blocked for, probabilistically speaking. We therefore match each recidivist user by counting the number of actions (new comments, comment edits, or comment deletions) they performed between consecutive blocks, and choosing a reformed user who made at least that many actions in their lifetime after their first (and only) block; this is visualized in Figure 3b (top). This ensures that both users had at least as many opportunities to be blocked, so any differences that remain must result from the quality, not quantity, of their actions. This pairing results in 1,332 paired short-term recidivists and reformed users and 3,278 paired long-term recidivists and reformed users.

As in the departure analysis, we also find a set of control pairs to help us understand whether any differences we observe are specific to recidivism, or apply generally to norm violations. Each control pair contains a first-time offender and a user who is completely clean up to the time of the first-time offender’s block. Selection is done on number of actions, just like the pairing of the experimental group. This procedure is visualized in Figure 3b (bottom).

Figure 5 summarizes how our engagement factors relate to recidivism and to blocking in general. For space considerations, we only present results for long-term recidivism; short term recidivism shows the same trends and statistical significance levels. Recidivist users receive and contribute more daily comments, but the received activity comes from a narrower set of users (solid bars). Taken together, these indicate that recidivist users’ interaction patterns have more depth than breadth; they engage with a smaller set of users but each engagement is more prolonged. This may signal the presence of sustained back-and-forth with a restricted circle of editors. These engagement patterns echo general blocking patterns (hashed bars): unsurprisingly, the type of individuals that are likely to get blocked in general are also more likely to recidivate (although this was not the case for departure). In the following sections we will complement this general observation by exploring factors that relate to the context of the first block, and thus are specific to recidivism.

Predicting future trajectory. To estimate the relative predictive power of these user characteristics factors and compare them with natural baselines we formulate three new *future* prediction tasks: one for departure, one for long-term recidivism, and one for short-term recidivism. Each task is run on the respective paired sets of users previously described, resulting in a balanced prediction task.

For transparency, we train a simple SVM classifier with our activity features as input, employing both the normalized and unnormalized versions of the activity level features. In each prediction task, hyperparameters for the SVM are selected through iteration over different combinations on a held-out

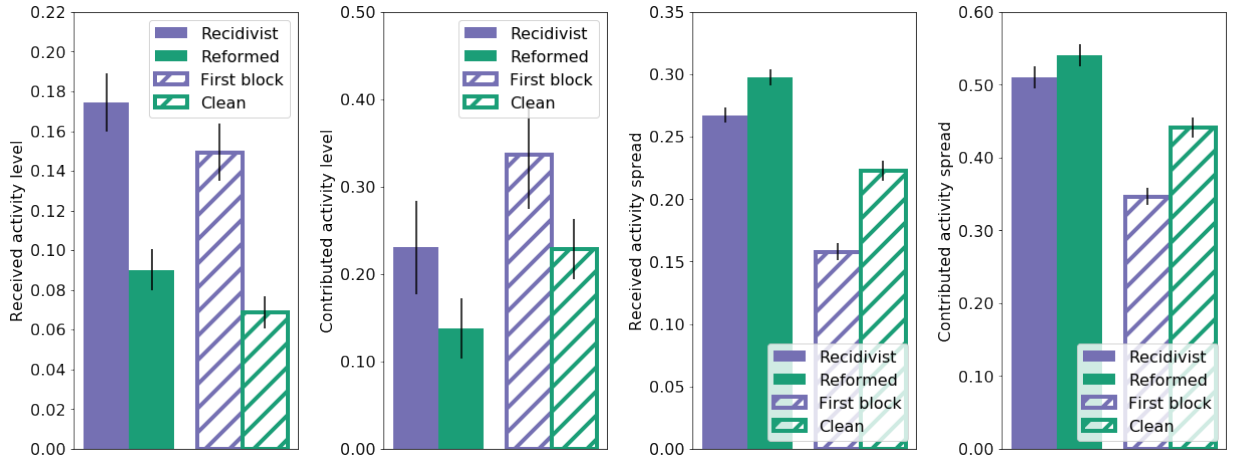


Figure 5: Difference between the community engagement patterns of recidivist and reformed users (solid bars). From left to right: receiving activity level, contributed activity level, receiving activity spread, contributed activity spread. For reference, hashed bars compare users that get blocked for the first time with those that never get blocked.

Feature set	Departure	Long-term recidivism	Short-term recidivism
Baseline: reason	59.0%	52.9%	51.9%
Baseline: duration	56.7%	50.0%	43.8%
Community age	58.6%	58.7%*	56.3%*
Engagement patterns	61.4%*	59.0%*	59.1%*
Engagement + Age	66.2%*	60.6%*	58.8%*

Table 2: Accuracies on three post-block trajectory prediction tasks (random baseline is 50%). *s indicate results that are significantly outperforming the best performing baseline for the respective task ($p < 0.05$).

development set, and accuracy is reported using leave-one-out cross-validation on the rest of the data.

We consider as features the user characteristics introduced earlier, calculated based only on information available *before* the block: community age and engagement patterns (activity level and spread). We compare their predictive power to two natural baselines that indicate the type and duration of the initial block. These account for the fact that certain kinds of offenses might be inherently more or less likely to lead to departure or recidivism.

Table 2 summarizes the results of the prediction experiments. We find that both community age and engagement patterns outperform baselines in all three prediction settings, showing they encode additional information about the future trajectory of the individual. Furthermore, they also combine to outperform each individual feature, showing that they capture different aspects of the phenomenon.

Overall, these results demonstrate that forecasting the future post-block trajectory of a user is at least a feasible and well-formed task. While our main goal in formulating these prediction tasks is to better understand the relative importance of different factors, we advise against using the resulting classifiers for any applications of algorithmically-assisted moderation. Given the sensitive nature of such applications, a public-facing model would need to be thoroughly tuned to ensure fair outcomes and exhaustively studied to understand if any biases in the data are being encoded [7, 12, 54]. Understanding the sources of such potential biases in the data, and devising mechanisms to eliminate them, constitutes an important avenue for future work.

5 BLOCK CONTEXT

Having established the relation between general user characteristics and post-block trajectories, we now seek to understand how the nature and context of the block itself mediates recidivism. Deterrence and defiance theories contrast two possible factors that relate to future compliance [46]. The traditional perspective is that the more severe the punishment is, the more likely it is to deter the perpetrator from misbehaving in the future [16, 27]. More recent work, however, suggests that a factor that is at least as important is whether the perpetrator perceives the punishment to be fair and justified [40, 52]: intuitively, individuals are unlikely to follow rules that they perceive to be imposed by an unfair and antagonistic society. We explore both these perspectives in the context of online recidivism, and bring supporting evidence for the perceived fairness perspective.

Block severity

The severity of temporary blocks corresponds naturally to their duration. For the purposes of this analysis, we only consider blocks that never had their duration changed. Since block durations are highly concentrated on the shorter end of the scale, with extreme outliers being sparsely spread in a long tail, we binarize durations as short (1 day or less) and long (more than 1 day), where the cutoff of 1 day was chosen because it is the median duration. Perhaps surprisingly, we find no statistically significant difference (chi-squared test) between these two classes in terms of recidivism rates, neither for long-term blocks, nor for short-term blocks.

While block duration does not appear to impact future compliance, it does correlate with departure rates: 30% of the users receiving long blocks quit within one month after the start of the block, compared with 24% for short blocks (significant at $p < 0.001$).⁷ This result holds even when controlling for the age and activity of the blocked individual.

Perceived fairness

Considering that in our online setting the actual severity of a block does not appear to impact its effectiveness, we can investigate how the way a user *perceives* their block relates to their future compliance with community rules.⁸

The main challenge in addressing questions related to user perception is capturing such a subjective notion at a sufficiently large scale. Unlike offline studies that rely on interviews with punished individuals to assess their perception of the fairness of the punishment [40], our observational data does not provide direct access to such variables. We can, however, leverage a specific Wikipedia mechanism in order to estimate them for a subset of users: block appeals.

A block appeal is a formal avenue through which a blocked user can contest the block, admit their wrongdoing, or ask for clemency. In response to an appeal, any moderator—not necessarily the one responsible for the original block—may choose to reduce the duration of the block. This happens rather rarely, with only 15% of blocked users having their block duration reduced. As described on Wikipedia’s official policy page on block appeals,⁹ there are two reasons a moderator may choose to grant an appeal. The most common reason is clemency: if the blocked user takes responsibility for their offense, acknowledges that they were in the wrong, and convincingly argues that they will not offend again, the reviewing moderator may grant a shortened block. The other reason is if the moderator judges the original block to be

in error, e.g., due to a case of mistaken identity. However, according to the cited Wikipedia policy, the latter scenario is “extremely rare”, due in part to the public nature of the moderation in Wikipedia and to the strict procedure which blocking moderators need to undergo in order to block somebody, which involves identifying a specific reason.

Linguistic signals. Block appeals provide two separate opportunities to (admittedly imperfectly) estimate how users perceive the block. The first one is through the language the blocked individual uses in their appeals. They can signal whether they perceived the block as being justifiable and acknowledge the wrongdoing or, conversely, whether they disagree with the justification for the block and contest it.

To capture such signals we consider three types of linguistic cues (Table 3) that are intuitively related to the user’s perception of the fairness of the block. Users that apologize signal that they acknowledge their wrongdoing; we expect such users to perceive the block as being fair and thus to be less likely to recidivate. Conversely, users that openly and directly challenge their block, either by aggressively questioning the administrators or by explicitly claiming unjust treatment, signal that they perceive the block as unfair and thus we would expect that they are more likely to recidivate.

To capture apologizing and direct questioning we use off the shelf tools¹⁰ that perform regular expression matching on dependency trees [9]. To capture explicit claims of unfairness we experimented with hand-building a small lexicon by examining a sample of examples from the data:¹¹ “unjust”, “illegitimate”, “illegal”, “unfair”, “not fair”, “accuse” (and different forms), “wrongly”, “falsely”, “injustice”, “unfounded”, “allege” (and different forms), and “unwarranted”.

We extract these linguistic cues from the messages recidivist and reformed users direct at moderators during their block, capturing their use in both formal and informal appeals. Figure 6 (a-c) compares the relative ratios of recidivist and reformed users among those that used a given type of fairness cue or not. Compared to a baseline ratio of 45% recidivism and 55% reform that consistently holds in all three cases, apologizing corresponds to a reduction in the rate of recidivism (chi-squared test $p < 0.05$), while direct questioning and explicitly calling for unfairness correspond to a moderate increase in the rate of recidivism ($p < 0.05$ and $p = 0.053$, respectively).

Granted appeals. A second opportunity to estimate the user’s perception of a block is by analyzing *granted* appeals. When an appeal is granted, the duration of the block is shortened. This positive response from the moderators arguably signals to the user that the system is fair and that their views

⁷Rather than analyzing departure during of the block as in the previous analysis, we are constrained here to consider a fixed time window in order to account for the time a user has to depart.

⁸While for brevity we only discuss long-term recidivism, the same effects and statistically significant levels hold for short-term recidivism.

⁹https://en.wikipedia.org/wiki/Wikipedia:Appealing_a_block

¹⁰From ConvoKit: convokit.cornell.edu

¹¹We also performed a Fightin’ Words analysis [34] to identify salient cues not present in the sample data we examined.

Feature	Examples
Apologizing	It is true that the tag-war got out of hand. I apologize if it caused any disruption. I am deeply sorry for not understanding the whole situation, and ask for your forgiveness. Given the recent history on the article, forgive me if claims of lawyers are met with skepticism.
Direct questioning	Why was I blocked, when I had not the right, but rather the duty to remove BLP inaccuracies? So what policy, precisely have I violated? How is it this person won't leave me alone and I repeatedly ask to be left alone and I get blocked?
Unfairness lexicon	this block is unjustified . none of the changes were in violation of the 3rr rule. i have alerted another administrator about your blatant and unwarranted abuse of power. iv been wrongly accused of making up information

Table 3: Example blocked user messages exhibiting linguistic indicators of (un)fairness perception. Bolded words indicate matched patterns.

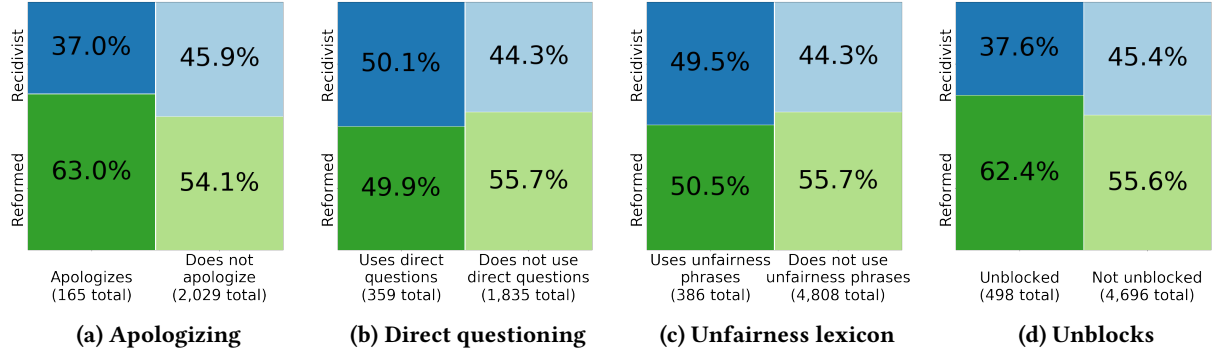


Figure 6: Normalized contingency tables showing the relative ratios of recidivist and reformed users conditioned on the presence or absence of fairness perception indicators. Table cell heights are scaled to the frequencies they represent in order to facilitate comparison between the two conditions (compare left to right in each figure).

are taken seriously. Paternoster et al. ([40], p. 169) shows that such perceptions about the enforcing authorities decrease the rate of recidivism offline, thus we hypothesize that this could hold true in online settings as well. Figure 6d shows that this principle is indeed echoed on Wikipedia: users who are unblocked are significantly more likely to reform ($p < 0.001$). This effect is even more striking when considering that, as discussed above, the duration of the block itself has no noticeable effect on recidivism and that granted appeals often shorten the duration of the block by less than one day (the median reduction is 22 hours).

Alternative explanations. One explanation that does not relate to perceived fairness could be that these appeal-based effects are simply differentiating between individuals who did not deserve to be blocked in the first place (and thus did not commit an offense and are unlikely to offend at all) and those that did commit an offense. While our observational approach precludes ruling out this alternative explanation completely, we argue that it is at least unlikely. First, as noted earlier, it is extremely rare for users to be blocked in error on Wikipedia. Second, some of the linguistic effects go in

the opposite direction of what such an explanation would predict: for example, apologizing generally comes with an admittance of guilt, and under the alternative explanation we would not expect such guilty individuals to have a reduced rate of recidivism. In fact, we find that individuals who apologize are significantly ($p < 0.001$, by chi-squared test) more likely to get unblocked than ones who do not, a pattern that is consistent with what we would expect if unblocks tend to be granted because of clemency (acknowledgment of wrongdoing being one of the factors that moderators look for when deciding whether to grant an appeal).

In an attempt to further discard explanations based on user characteristics we also repeat these experiments after controlling for age and activity (i.e., pairing recidivist and reformed users based on these variables before conducting the analysis) and found similar trends, with exception of the unfairness lexicon (which had a moderate effect to start with and was likely sensitive to the sample size reduction).

We argue that these observations, while not of causal nature, at least motivate further investigations into the role of perceived fairness in online moderation.

6 FURTHER RELATED WORK

Antisocial behavior. Online moderation largely addresses the problem of antisocial behavior, which occurs in the form of harassment [50], cyberbullying [47], and general aggression [23]. Approaches to moderating such content include decentralized, community-driven methods [14], as well as top-down methods relying on designated community managers or moderators [33]. Prior research in this area ranges from understanding the actors involved in antisocial behavior [6, 28, 36, 42, 51] to analyzing its effects [39] to tools for identifying such behavior [38, 53], and even forecasting future instances [31, 55]. While inspired by this line of work, our present study extends it by focusing on what happens *after* a user is blocked for violating community rules.

Effects of moderation. Prior work has examined the effects of different kinds of moderation on online platforms. Community-driven moderation can affect overall user participation [8, 35] and comment quality [5] in conversations. Centralized moderation can also have effects on conversation, mainly through the moderator’s role as an authority figure [44]. These existing studies of moderation effects have focused on the short term, largely operating at the level of individual conversations. By contrast, in this work we intend to study the long term, *user* level effects of moderation: what happens to a user in the days, weeks, and months after a moderator takes action against them?

Norms and engagement. A major factor governing engagement in online communities is a sense of belonging [37], which in many communities engenders the emergence of community-specific norms, such as specific patterns of language [3, 10, 19, 56]. Wikipedia is no exception: it relies on group dynamics to promote editing productivity [26] and quality [41], as well as participation in governance [1, 30].

Fairness in moderation. Moderation is fundamentally a difficult process, sparking discussions about design and best practices [2, 24], and moral questions regarding bias [45] and free speech [21]. One interesting avenue of research in this broad space deals with understanding the *perception* of fairness. Prior work has looked at this from the perspective of moderators, showing that moderators on the technology news forum Slashdot frequently judge each others’ actions as unfair, but that this rarely leads to reversal of moderation decisions [29]. We instead explore the perspective of users, exploring ways to determine when a user regards a moderation decision as unfair.

7 DISCUSSION

We introduced a computational framework for characterizing the possible trajectory a user can take after a temporary block: departure, recidivism, or redemption. Taking inspiration from theories of offline punishment, our investigation

focused on two types of factors that are tied to these outcomes: those related to a user’s involvement in the community, and those related to their perception of the moderation action affecting them.¹² Overall, our work builds towards a more nuanced approach to moderation that more broadly accounts for the tradeoffs between possible outcomes and that considers how the affected individuals might perceive the moderators and their actions.

Our study has several limitations that naturally translate into opportunities for future work. First, while we have focused on the blocked users, it would be valuable to also understand the normative effect of blocks on other users that are witnessing them. Additionally, our study is limited to revealing correlational effects, and different methodologies (involving surveys or experimentation) would be necessary to study the underlying causal mechanism.

In terms of methodology, there are several ways in which the definitions of the post-block trajectories themselves could be refined. While we measure departure as a binary value, in reality community participation occurs on a variable scale and it is possible that blocked users do not depart, but drastically reduce their participation (or conversely, increase it). Similarly, defining recidivism in terms of a second block is an imperfect approach because it is possible that some users reoffend without getting blocked. This is especially true in the case of personal attacks, where each individual’s threshold of acceptable behavior may vary. One possible way of addressing this would be to combine our approach with prior work in automated detection of toxic language [11, 42, 53].

We also note possible ways in which the factors we explore could be expanded upon. The activity features we introduced in Section 4 effectively serve as crude approximations of network effects. Future work could apply a more formal network analysis to uncover more powerful representations of activity. Similarly, our linguistic measures of fairness are currently limited. One immediate way of expanding them would be to build a more comprehensive lexicon of unfairness, perhaps using a semi-automated approach based on word co-occurrences with the currently existing lexicon items. A more sophisticated approach might involve engineering dependency-based features similar those used to extract apologies and direct questioning. Understanding whether a user feels they are treated unfairly constitutes an interesting research question on its own.

Acknowledgements. The authors thank Lucas Dixon, Liye Fu, Jon Kleinberg, Lillian Lee, Dario Taraborelli, Justine Zhang and Andrew Wang and the anonymous reviewers for their helpful comments. This work is supported in part by a Google Faculty Research Award, NSF CAREER award IIS-1750615 and NSF Grant SES-1741441.

¹²We refrain from drawing any explicit connections between online and offline punishment, acknowledging their different nature and goals.

REFERENCES

- [1] Moira Burke and Robert Kraut. 2008. Mopping up: Modeling Wikipedia Promotion Decisions. In *Proceedings of CSCW*.
- [2] Carlos Castillo, Francesco Fabbri, and Diego Saez-Trumper. 2017. Current Practices of Online Community Managers. (2017).
- [3] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities. In *Proceedings of CHI*.
- [4] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Red-dit Norm Violations at Micro, Meso, and Macro Scales. In *Proceedings of CSCW*, Vol. 2.
- [5] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How Community Feedback Shapes User Behavior. In *Proceedings of ICWSM*.
- [6] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Antisocial Behavior in Online Discussion Communities. In *Proceedings of CSCW*.
- [7] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of KDD*.
- [8] Tiago Oliveira Cunha, Ingmar Weber, Hamed Haddadi, and Gisele L. Pappa. 2016. The Effect of Social Feedback in a Reddit Weight Loss Community. In *Proceedings of DH*.
- [9] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of ACL*.
- [10] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of WWW*.
- [11] Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.
- [12] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of KDD*.
- [13] A. Forte and A. Bruckman. 2008. Scaling Consensus: Increasing Decentralization in Wikipedia Governance. In *Proceedings of HICSS*.
- [14] R. Stuart Geiger. 2016. Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space. *Information, Communication & Society* 19, 6 (June 2016).
- [15] R. Stuart Geiger and David Ribes. 2010. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proceedings of CSCW*.
- [16] Harold G. Grasmick and George J. Bryjak. 1980. The Deterrent Effect of Perceived Severity of Punishment. *Social Forces* 59, 2 (Dec. 1980).
- [17] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System. *American Behavioral Scientist* 57, 5 (May 2013).
- [18] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work. In *Proceedings of WikiSym*.
- [19] William L. Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in Online Communities. In *Proceedings of ICWSM*.
- [20] Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community. In *Proceedings of EMNLP*.
- [21] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The View from the Other Side: The Border between Controversial Speech and Harassment on Kotaku in Action. *First Monday* 23, 2 (Feb. 2018).
- [22] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Block-lists. In *Proceedings of CSCW*.
- [23] Joseph M. Kayany. 1998. Contexts of Uninhibited Online Behavior: Flaming in Social Newsgroups on Usenet. *Journal of the American Society for Information Science* 49, 12 (Jan. 1998).
- [24] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating Behavior in Online Communities. In *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press.
- [25] Gary King, Jennifer Pan, and Margaret E. Roberts. 2013. How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review* 107, 02 (May 2013).
- [26] Aniket Kittur, Bryan Pendleton, and Robert E. Kraut. 2009. Herding the Cats: The Influence of Groups in Coordinating Peer Production. In *Proceedings of WikiSym*.
- [27] Steven Klepper and Daniel Nagin. 1989. Tax Compliance and Perceptions of the Risks of Detection and Criminal Prosecution. *Law & Society Review* 23, 2 (1989).
- [28] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web.
- [29] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of CHI*.
- [30] J Leskovec, D P Huttenlocher, and Kleinberg. 2010. Governance in Social Media: A Case Study of the Wikipedia Promotion Process. In *Proceedings of ICWSM*.
- [31] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Proceedings of ICWSM*.
- [32] Toni Makkai and John Braithwaite. 1994. Reintegrative Shaming and Compliance with Regulatory Standards. *Criminology* 32, 3 (Aug. 1994).
- [33] Aiden R McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Ann Crane-field. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. In *Proceedings of ICIS*.
- [34] Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis* 16, 04 (2008).
- [35] Lev Muchnik, Sinan Aral, and Sean Taylor. 2013. Social Influence Bias: A Randomized Experiment. *Science* 341 (2013).
- [36] Kevin Munger. 2017. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior* 39, 3 (Sept. 2017).
- [37] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Proceedings of ICWSM*.
- [38] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of WWW*.
- [39] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The Effect of Extremist Violence on Hateful Speech Online. In *Proceedings of ICWSM*.
- [40] Raymond Paternoster, Robert Brame, Ronet Bachman, and Lawrence W. Sherman. 1997. Do Fair Procedures Matter? The Effect of Procedural Justice on Spouse Assault. *Law & Society Review* 31, 1 (1997).
- [41] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. 2018. Mind Your POV: Convergence of Articles and Editors Towards Wikipedia's Neutrality Norm. In *Proceedings of CSCW*.
- [42] Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgilio A. F. Almeida, and Wagner Meira Jr. 2018. Characterizing and Detecting Hateful Users on Twitter. In *Proceedings of ICWSM*.

- [43] Paul R Rosenbaum. 2010. *Design of Observational Studies*. Springer, New York.
- [44] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of CSCW*.
- [45] Qinlan Shen, Michael Miller Yoder, Yohan Jo, and Carolyn P. Rose. 2018. Perceptions of Censorship and Moderation Bias in Political Debate Forums. In *Proceedings of ICWSM*.
- [46] Lawrence W. Sherman. 1993. Defiance, Deterrence, and Irrelevance: A Theory of the Criminal Sanction. *Journal of Research in Crime and Delinquency* 30, 4 (Nov. 1993).
- [47] Vivek K. Singh, Marie L. Radford, Qianjia Huang, and Susan Furrer. 2017. "They Basically like Destroyed the School One Day": On Newer App Features and Cyberbullying in Schools. In *Proceedings of CSCW*.
- [48] Michael Tonry. 2008. Learning from the Limitations of Deterrence Research. *Crime and Justice* 37, 1 (Jan. 2008).
- [49] Tom R. Tyler. 1990. *Why People Obey the Law*. Yale University Press, New Haven.
- [50] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of CSCW*.
- [51] Svitlana Volkova and Eric Bell. 2017. Identifying Effective Signals to Predict Deleted and Suspended Accounts on Twitter across Languages. In *Proceedings of ICWSM*.
- [52] Kirk R. Williams. 2005. Arrest and Intimate Partner Violence: Toward a More Complete Application of Deterrence Theory. *Aggression and Violent Behavior* 10, 6 (Sept. 2005).
- [53] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of WWW*.
- [54] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of WWW*.
- [55] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*.
- [56] Justine Zhang, William L. Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community Identity and User Engagement in a Multi-Community Landscape. In *Proceedings of ICWSM*.