Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits

Ashwin Rajadesingan, Paul Resnick, Ceren Budak
School of Information
{arajades, presnick, cbudak}@umich.edu
University of Michigan, Ann Arbor

Abstract

Online communities about similar topics may maintain very different norms of interaction. Past research identifies many processes that contribute to maintaining stable norms, including self-selection, pre-entry learning, post-entry learning, and retention. We analyzed political subreddits that had distinctive, stable levels of toxic comments on Reddit, in order to identify the relative contribution of these four processes. Surprisingly, we find that the largest source of norm stability is pre-entry learning. That is, newcomers' first comments in these distinctive subreddits differ from those same people's prior behavior in other subreddits. Through this adjustment, they nearly match the toxicity level of the subreddit they are joining. We also show that behavior adjustments are community-specific and not broadly transformative. That is, people continue to post toxic comments at their previous rates in other political subreddits. Thus, we conclude that in political subreddits, compatible newcomers are neither born nor made– they make local adjustments on their own.

1 Introduction

All online communities, because of user churn, need to attract new users to ensure their survival. While newcomers may bring in new ideas and perspectives that revitalize the community (Levine, Choi, and Moreland 2003), they are also more likely to violate community norms (Kiesler et al. 2012)

and increase workload for moderators (Kiene, Monroy-Hernández, and Hill 2016). Ideally, newcomers are already in sync with the norms of the community, but that is not always the case, as some communities have distinctive norms that are vastly different from the norms that exist outside. Thus, a major challenge for communities with distinctive norms is to attract and retain compatible newcomers and acculturate those whose outside behavior is not compatible.

There are four processes that can contribute to stability of norms despite an influx of newcomers. Newcomers can choose (not) to join upon inspecting posted rules and guidelines, and observing participation of existing members (Preece, Nonnecke, and Andrews 2004), a process known as *self-selection*. If they do join, those observations prior to making their first posts may also lead them to align their own commenting behavior to match what they perceive as the community's norms, a process we refer to as *pre-entry learning*. Once they start posting, newcomers may further adapt their behavior based on feedback from others (Lampe and Johnston 2005) as well as continued observation, a process we refer to as *post-entry learning*. Not all newcomers stay in a community; those whose initial behavior conforms to the community norms might be more likely to return, a process we refer to as *selective retention*.

All four of these processes have been identified in past research—albeit in isolation. While some study self-selection processes and their importance (Panciera, Halfaker, and Terveen 2009), others highlight the importance of learning (Tan and Lee 2015). Most research on norm conforming behavior focuses on post-entry socialization (Choi et al. 2010; Ren, Kraut, and Kiesler 2007; Ducheneaut 2005; Burke, Kraut, and Joyce 2010) highlighting that these processes, especially learning from the community's response to one's posts, are crucial to norm conformity. If attention in the research literature is any signal, it would seem that post-entry processes contribute to norm stability more than pre-entry processes. In this work, through data and modeling, we aim to determine if this indeed is the case for one particular norm, toxicity of comments. We move beyond examining these processes in isolation and examine the relative importance of pre-entry and post-entry processes on maintaining toxicity norms in political communities on Reddit.

While colloquially and in popular media, many online discussion forums are often railed against for being toxic[1][2], systematically classifying comments as toxic or not is a complex problem because of the inherent subjectivity of the task (Aroyo et al. 2019). In this work, we use the current state-of-the-art, Jigsaw Perspective API (Wulczyn, Thain, and Dixon 2017) to identify toxic comments. We do not expect different subreddits to hold the same notion of toxicity as that encoded by the Perspective API. In fact, it is precisely the differences in sensibilities that will lead to subreddits

[1] https://www.popsci.com/block-toxic-comments/

[2] https://www.cnbc.com/2019/05/20/microsoft-xbox-moderation-to-cut-back-toxic-content.html

having varying fractions of comments that the Perspective API labels as toxic.

In our analysis, we find that there are many political subreddits with toxicity levels (in terms of percentage of comments that are toxic) that are stable across many months, yet distinct from the toxicity levels in other subreddits. For example, the subreddit r/NeutralPolitics has few toxic comments while r/uncensorednews has many. Because the toxicity level is stable over time yet different from the level in other subreddits, it must somehow be reproduced through social processes. We investigate the sources of that stability by studying the aforementioned norm conforming processes. Furthermore, to the extent that individuals learn, adjusting their toxicity level to match the subreddit, we also investigate whether that learning is community-specific or transformative - that is, whether newcomers internalize these norms and apply them in other communities in which they participate in the future.

We summarize our contributions in this work as follows:

1. We find that pre-entry learning and to a lesser extent self-selection explain the stability of toxicity norms.
2. We find that individual learning of toxicity norms is community specific and not broadly transformative.

2 Related work

2.1 Social norms in online communities

Norms Social norms define acceptable behavior within a community. People learn to behave according to the norms of a community by observing typical behaviors exhibited by others (descriptive norms) and behaviors that are encouraged or sanctioned (injunctive norms) (Cialdini, Reno, and Kallgren 1990). Since Sherif's early autokinetic studies and Asch's conformity experiments, social psychologists have long identified the importance of social norms and other situational factors in constraining and influencing human behavior (Ross and Nisbett 2011). The power of social norms is observed in online communities as well. In fact, the Social Identity model of Deindividuation Effects (SIDE) suggests that visual anonymity, which is common in online settings, is associated with an increase in salience of group social identities and adherence to group normative behavior (Reicher, Spears, and Postmes 1995). Further, under conditions of anonymity, individuals are more prone to group influence than identifiable individuals.

Postmes, Spears, and Lea (2000) showed that norms are socially constructed over time, and are influenced by users' social identity in computer-mediated communication. Community managers may also influence the type of norms that become prominent in a community by constructing prosocial rules, showcasing good behavior (Park et al. 2016) and sanctioning disruptive behavior (see (Kiesler et al. 2012) for a detailed review). These norms may be formalized as community rules by moderators making them more visible, accessible and easier to follow for newcomers (Fiesler et al. 2018). Over time, different communities may develop different sets of norms that define acceptable behavior. Interestingly, specific to Reddit, Chandrasekharan et al. (2018) found that certain norms against personal attacks and racism are almost universally present in all subreddits (macro norms), while norms banning spam links are present in some subreddits (meso norms), still fewer subreddits have norms against Islamophobia and xenophobia (micro norms).

Norm conformity Researchers have studied different norm conforming processes that lead newcomers' behavior to resemble that of existing members in online communities. These processes can be grouped into two phases: pre-entry processes and post-entry processes, based on their period of influence relative to joining a community.

1. Pre-entry processes

Selection While most online communities don't explicitly choose their members, they attract like-minded people with similar beliefs who likely fit into their norms (Norris 2002). The importance of self-selection to norm conformity becomes evident during its absence, when online communities struggle to deal with a sudden influx of uninitiated newcomers whose behavior does not align with the norms of the community (Lin et al. 2017; Kiene, Monroy-Hernández, and Hill 2016).

Learning pre-entry Newcomers can start learning about community norms before they even interact with other community members. Social learning theory suggests that individuals learn by observing how others behave (Bandura 1978). Burke, Marlow, and Lento (2009) found that newcomers on Facebook who observe their friends sharing photos are likely to increase their own photo-sharing behavior. In addition to learning descriptive norms by observing others' behavior (descriptive norms), newcomers may also learn about injunctive norms by observing sanctions of bad behavior and by reading official policies (Matias 2019). Individuals may also ascribe culturally developed meanings to situations which in turn regulate their behavior. A variant of symbolic interactionism, affect control theory, suggests that people behave in ways in order to maintain the affective feelings evoked by an instance of a situation (Robinson, Smith-Lovin, and Wisecup 2006). Thus, individuals may appear to follow the norms of a community without ever having participated in that specific community before.

2. Post-entry processes

Learning post-entry Upon joining a community, the venues for learning expand. In addition to the pre-entry processes described above, newcomers can learn the norms through the feedback they receive from moderators and other community members. This phenomenon—learning by participation—has been well studied in online communities (Lampe and Johnston 2005; Danescu-Niculescu-Mizil et al. 2013). Members conform to linguistic norms as they become more experienced members (Danescu-Niculescu-Mizil et al. 2013; Nguyen and Rosé 2011). Newcomers learn to adhere to the norms of blogging communities by observing, engaging and mimicking existing bloggers (Dennen 2014). Why does participation lead to conformity? Kraut et al. (2010) and Choi et al. (2010) highlight the importance of socialization tactics used by moderators, while Tonteri et al. (2011)

underline the sense of belonging in the community that comes with participation.

Retention In some cases, users who are not a good fit to the community voluntarily leave (Schilling, Laumer, and Weitzel 2012), while in other extreme cases, non-conforming users may be expelled to sustain order in the community (Geiger and Ribes 2010). Retention of norm-conforming users is also essential for the growth and sustenance of online communities. Most online communities report low levels of newcomer retention (Panciera, Halfaker, and Terveen 2009). Thus, research on retention has primarily focused on limiting user churn and increasing contributions among newcomers (Burke, Marlow, and Lento 2009).

2.2 Toxicity and Incivility

While researchers have studied a myriad of toxic behaviors under different names, research on incivility is closest to toxicity. Here too, communication scholars have defined, quantified and operationalized incivility in different ways. Coe, Kenski, and Rains (2014) defined incivility to include name-calling, aspersion, lying, vulgarity, and pejorative remarks. Gervais (2015) identified insults, extreme language and emotionality as different markers of incivility. In synthesizing the varied definitions of incivility, Stryker, Conway, and Danielson (2016) note that a common strain is "a focus on rudeness in the political arena" (see (Papacharissi 2004) for an exception).

Wulczyn, Thain, and Dixon (2017), whose classifier we use, define *toxicity* to include elements of incivility but also a holistic assessment. They asked human labelers to judge whether a comment is, "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion". This naturally leads to a continuous measure based on people's subjective judgments: the toxicity of a comment is the percentage of human raters that would label it as toxic. The production version of their classifier, Perspective API, which we use in our study, has also been used in other Reddit research such as (Mittos et al. 2020) to measure toxicity.

2.3 Toxicity in Online Communities

Toxic behavior manifests in different forms in online communities and is studied in multiple contexts such as incivility (Borah 2014), trolling (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015), harassment (Blackwell et al. 2017) and cyberbullying (Kwak, Blackburn, and Han 2015). As online discussion communities are often riddled with toxic content, researchers have developed multiple machine learning techniques to detect and remove toxic comments at scale (Wulczyn, Thain, and Dixon 2017; Chandrasekharan et al. 2017b).

Given the widespread prevalence of toxicity, researchers have sought to understand the mechanics of such behavior online. Research on League of Legends, an online multiplayer game, showed that toxic players exhibit behavior similar to others at the start of a match but change their behavior during the match (Kwak, Blackburn, and Han 2015). In contrast, studying discussion communities, (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015) found that users who get banned write worse to begin with and degrade more than others over time. Qualitative research has shown that toxic behavior is not an isolated phenomenon but a consequence of more structural factors such as platform affordances (Leurs and Zimmer 2017), content moderation policies (Blackwell et al. 2017) and community culture (Massanari 2017). For example, Massanari (2017) noted the role of Reddit's geek culture and geek masculinity in supporting "toxic technocultures" that normalizes anti-feminist and misogynistic activism.

3 Intuition and Framework

Based on our literature review, we have identified four processes that enable norm stability despite turnover of members: selection, learning pre-entry, learning post-entry and retention. If toxicity levels are to remain stable, the newcomers must exhibit toxicity levels in the community that are similar to those of existing members. Figure 1 provides two illustrative examples for when a community has lower toxicity on average than the rest of the population. Configuration A shows a combination of selection and pre-entry learning. Configuration B shows a combination of retention and post-entry learning. In empirical analyses, we estimate the magnitude of each of these four processes for many political subreddits, during time periods when their toxicity levels are stable and distinctive from the overall average toxicity level.

More formally, we quantify these different processes as follows: For a community C, we identify *joiners* as individuals who have never previously posted in C and have posted in C for the first time on month t, and *non-joiners* as individuals who do not participate in C until (and including) month t.

3.1 Selection Effect

To quantify the average selection effect (SE) for a community C, we compare the toxicity exhibited by joiners ($\beta_{elsewhere\_last\_month}$) and non-joiners ($\beta_{control\_last\_month}$) in the month prior to the joiners making their first comments.

$$SE = \beta_{elsewhere\_last\_month} - \beta_{control\_last\_month}$$

3.2 Learning Pre-Entry

To quantify Learning Pre-Entry ($LPre$) for a community C, we compare the toxicity exhibited by joiners in their first comment in community C (β_{first}), to the toxicity expressed by them in the previous month in other political subreddits ($\beta_{elsewhere\_last\_month}$).

$$LPre = \beta_{first} - \beta_{elsewhere\_last\_month}$$

3.3 Retention Effect

To quantify retention effect (RE) for a community C, we compare the toxicity exhibited by joiners who leave after posting only one comment in their first month ($\beta_{first\|exiter}$), to the toxicity exhibited in the first comment of joiners who stay to post more comments in their first month ($\beta_{first\|returner}$).

$$RE = \beta_{first\|returner} - \beta_{first\|exiter}$$

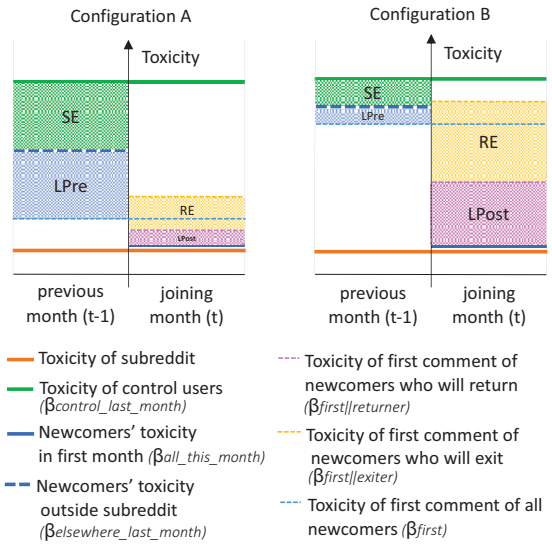

Figure 1: Two ways for newcomers' toxicity level (—) to match the community's toxicity level (—) which is lower than the toxicity elsewhere (—). In Configuration A, the community attracts new users whose natural toxicity level (--) is lower than the population at large leading to a large selection effect (SE,). In addition, the newcomers' first posts in the community (---) have lower toxicity than their posts the previous month in other communities, indicating a large effect of learning pre-entry ($LPre$,). By contrast, in Configuration B, those effects are small. Instead, the newcomers who return to post after their first comment (---) are much less toxic than those who leave (---), showing a large effect of retention (RE,). These newcomers who stay also adjust to the community norms during their joining month, thus showing a large effect of learning post-entry ($LPost$,).

3.4 Learning Post-Entry

To quantify Learning Post-Entry ($LPost$), for all joiners who return to post more than one comment in their first month in community C, we compare the toxicity exhibited by these joiners in their first comment ($\beta_{first||returner}$) to the toxicity expressed by them through all their comments in their first month ($\beta_{all\_this\_month}$).

$$LPost = \beta_{all\_this\_month} - \beta_{first||returner}$$

3.5 Transformative learning

To identify if newcomer adjustment results in transformative learning, we compare the toxicity exhibited in other political communities by joiners in the month prior to them joining community C ($\beta_{elsewhere\_last\_month}$), to the toxicity exhibited by the same joiners in other political communities in the month after they join C ($\beta_{elsewhere\_next\_month}$). Transformative learning is not shown in Figure 1 to avoid visual overload.

$$TL = \beta_{elsewhere\_next\_month} - \beta_{elsewhere\_last\_month}$$

In order to measure the strength of the norm conforming processes and possible transformative learning effects, we need to estimate the toxicity of comment groups such as *elsewhere_last_month* and *control_last_month* which indicate the context in which the comment is posted. The exact definitions of each comment group and details on how its toxicity is estimated is given in Section 5.

4 Data

4.1 Reddit data

Reddit can be viewed as a community of communities (called subreddits). These subreddits are organized based on topic and can have millions of subscribers. They are relatively autonomous with their own moderators, separate rules and website design through CSS. These factors allow communities to develop their own distinct norms, providing an ideal setting to examine how newcomers conform to diverse toxicity norms in different communities. In this work, we analyze political subreddits using public Reddit comments by accessing PushShift's BigQuery Reddit dump.[3]

4.2 Data quality

The BigQuery dataset does not contain comments removed by moderators. If we were to analyze only the comments that were not removed, it would produce a biased estimate of the strength of the norm conforming processes as removed comments may be disproportionately norm violating. Therefore, we also retrieve, whenever available, details of the removed comments by querying the PushShift Search API. For comments posted from April 2017, the API returns comments that were collected within a few seconds of posting, that is, *before* they were likely removed by the moderators, using an approach similar to (Chandrasekharan et al. 2018).

We restrict our analysis to subreddits from April 2017 to February 2018, the period for which we have access to the removed comments. In total, there are 1,410,203 removed comments for the political subreddits identified in Section 4.4. We could not retrieve information for 57% of the removed comments of which more than 90% were removed in under 30 seconds after posting (median = 1 second), likely by moderator bots such as AutoModerator. The AutoModerator Bot[4] is a customizable bot that helps moderate subreddits based on simple regular expression rules set by community moderators. Any obvious rule-breaking content is quickly acted upon by the AutoModerator and the content may be removed even before other community members see them. Our dataset does not include information on such auto-moderated comments, a limitation that we will return to later in Section 9.

4.3 Ethical considerations

In performing this research, we took into account work on conducting ethical research using social media data (Zimmer and Kinder-Kurlanda 2017; Fiesler and Proferes 2018). Of particular concern are the moderator-removed comments that were likely norm violating. Though these removed comments were publicly posted, users would not reasonably expect those comments to be made public and attributable to

[3] https://pushshift.io/using-bigquery-with-reddit-data/

[4] https://reddit.com/wiki/automoderator/full-documentation

them. Thus, to protect user privacy, we choose not to release the raw dataset containing the removed content. Instead, we will release a fully anonymized dataset containing only the aggregate number of comments posted and fraction of those comments that was classified as toxic for each user, where users will be anonymized through hashing[5]. We will not be releasing the actual text content posted by any user and will also not indicate if a comment was removed or not. We also clarify that we do not retrieve or use comments deleted by the users themselves. We believe that the benefits of this research, which provides valuable insights into how newcomers conform to toxicity norms and therefore informs relevant design implications to improve content moderation in online communities, outweighs the minimal risk to privacy which we mitigate by releasing only aggregated, anonymized statistics.

4.4 Identifying political subreddits

Most politics-related research on Reddit predominantly rely on either hand selection (Soliman, Hafer, and Lemmerich 2019; Guimaraes et al. 2019; An et al. 2019) or crowdsourced lists (Nithyanand, Schaffner, and Gill 2017) to identify political subreddits. The former approach usually includes only popular political subreddits, while the latter – crowdsourced lists – are usually incomplete and not regularly updated. For example, the most comprehensive crowdsourced list built by moderators of r/politics [6] is no longer public or maintained. Therefore, in this work, to identify political subreddits, we first train a classifier to detect if a comment is politics-related and use this classifier to identify subreddits that host political discussions. One advantage of this classifier-driven approach over using crowdsourced lists is that this approach is replicable and the list of political subreddits identified can be regularly updated. Further, as we are classifying subreddits as political based on their comments rather than their name or description, we are able to find subreddits that are not specific to politics but nevertheless host mostly political discussion.

We built an L1-regularized logistic regression model trained on unigram, bigram and trigram word features, using as training data, a random sample of comments from r/politics for the "politics" class and about 2000 comments from each of the other default subreddits[7] as the negative class. In total, we used 189,916 comments, evenly split across the two classes for training the classifier[8]. Assessed through 10-fold cross validation, we obtain an accuracy of 81% (84% precision and 78% recall for politics class). To identify political subreddits, we sample 2000 comments each from all subreddits on Reddit and classify them as political (or not) according to the clas-

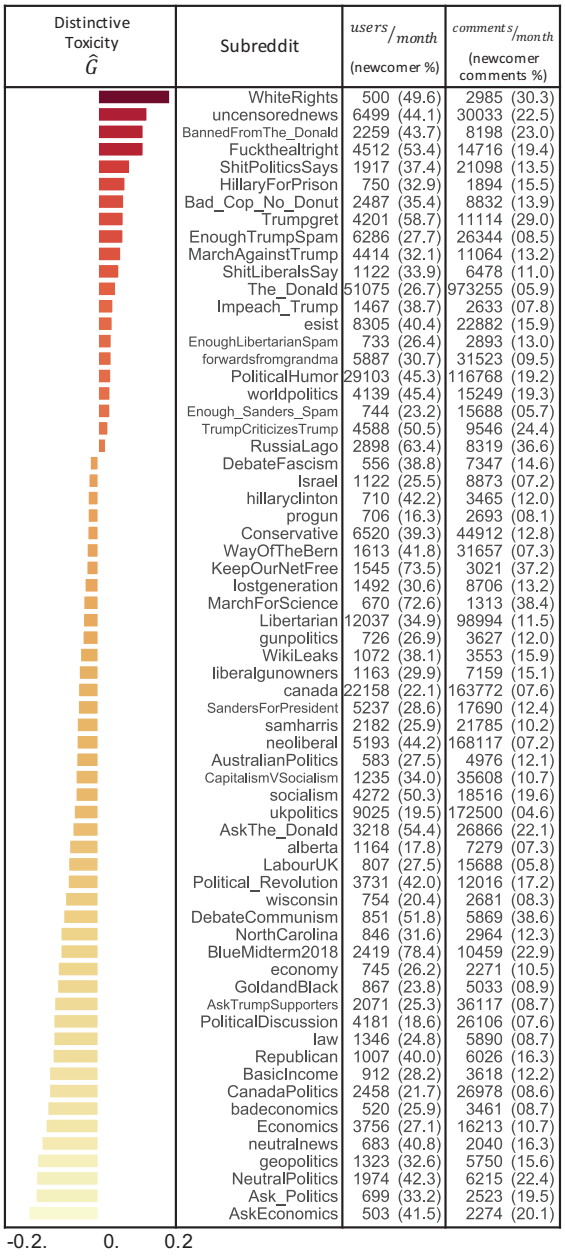

Figure 2: All 65 political subreddits with stable and distinctive toxicity norms listed in descending order of their distinctive toxicity. For each subreddit, the center-right column contains the average number of users per month and the percentage of newcomers per month during their stable period. The right-most column lists the average comments per month and the percentage of comments produced by newcomers per month during their stable period.

[5]Contact the first author for the aggregated data, classifier and the full list of identified subreddits.

[6]https://web.archive.org/web/20190502124604/https://www.reddit.com/r/politics/wiki/relatedsubs

[7]Until 2017, new users would automatically be subscribed to 50 subreddits to showcase to them a representative variety of content on Reddit, one of which was r/politics.

[8]There may be some politics-related comments for the negative class in the training data as the other default subreddits may also contain political content, albeit to a very small degree. However, we don't expect it overly to affect our identification of political subreddits as we make that determination based on a large sample of 2000 comments from each subreddit.

sifier predictions. Then, we select subreddits which have a majority of their comments ($> 50\%$) classified as political for further analysis. In order to focus our analysis on reasonably established communities, we excluded subreddits having fewer than 5000 commenters. Unlike the aforementioned crowdsourced list, this approach yields not only subreddits typically construed as political such as r/democrats and r/Conservative but also location-based subreddits such as r/wisconsin and r/NorthCarolina, humor subreddits such as r/libertarianmeme and r/PoliticalHumor and issue-based subreddits such as r/PanamaPapers which host large amounts of political content.

4.5 Measuring toxicity norms

One way to think about differences in toxicity norms between communities is to think of each community having its own definition of toxicity – content that would make people want to leave *that community*. For example, many comments in r/WhiteRights, a (now banned) subreddit, would have been troubling to members of r/AskEconomics but not to members of r/WhiteRights.

We take a different approach. We apply a single universal definition of toxicity, but recognize that content that a general population would identify as toxic is not perceived as problematic in all communities. We do not directly observe what communities view as problematic or not (except for comments removed by moderators). With a universal definition of toxicity, however, we expect a lower prevalence of toxicity in communities that find such content problematic.

Thus, we try to measure the community's norm for toxicity, the amount of it that members see. Specifically, we define a community's toxicity norm as the observed prevalence of content identified as toxic according to a general population's assessment of toxicity. Concretely, we use the Perspective API classifier, which was trained using crowdsourced labeling, to identify comments as toxic.[9]

Note that our measure is intended to capture aspects of both descriptive and injunctive norms. We measure the toxicity norm for each subreddit in each month based on all the unremoved comments posted that month.

We exclude removed comments because these clearly violate the injunctive norm of the community and because the removal makes them less visible to community members, affecting the descriptive norm.[10]

Alternate measurement of toxicity norms Community members also signal injunctive norms through voting on content, which affects visibility of content (descriptive norms). Thus, an alternative measure of the toxicity norm could be developed based on whether less toxic comments are more likely to get upvotes. However, users may vote on content for a variety of reasons unrelated to toxicity. Indeed, examining the unremoved comments, we find little correlation ($r = 0.028$) between the voting points that the comments accrued and its toxicity (measurement described below) . Thus, we do not include voting points in our operationalization of toxicity norms.

Evaluating the Perspective API classifier for Reddit comments Perspective API's toxicity classifier has been used to identify toxic comments in multiple online communities such as Wikipedia, New York Times[11] and even Reddit (Mittos et al. 2020). However, to be suitable for use for our task – identifying toxic comments in political subreddits – there are two robustness concerns.

The first concern is bias against subgroups especially with respect to misclassifying (i) comments mentioning marginalized subgroups and (ii) comments posted in non White-aligned dialects such as African American English (AAE) (Sap et al. 2019). To address the former, the Perspective team recently published documentation detailing performance characteristics across comments mentioning fifty different minority, marginalized and intersectional groups [12]. In recent versions, the AUC scores for the classifier across these groups are almost always above 0.85, allaying our concerns about bias against comments mentioning subgroup identities. While concerns about misclassifying AAE still remain unresolved, for this to affect the measurement of the stability of community toxicity norms in our work, the level of such misclassifications would have to vary considerably between months, which is unlikely.

The second robustness concern is generalizability. Since the classifier has been trained on Wikipedia data, we need to ensure that it can detect toxic Reddit comments with reasonable accuracy. To evaluate Perspective's performance on Reddit comments, we compared the classifier assigned toxicity score to a ground truth computed as the majority of eleven human labelers' evaluations collected using Amazon Mechanical Turk (MTurk) [13]. As a baseline for comparison, we checked how well a single MTurk evaluation can predict the majority of eleven human labelers' evaluations. We sampled 100 comments each from r/WhiteRights (most toxic community according to Perspective), r/AskEconomics (least toxic) and also 100 comments from the other subreddits. Following Perspective's approach to establishing ground truth, Turkers labeled each comment as very toxic, toxic, neither, healthy contribution or very healthy contribution. Again following Perspective, we binarized the labels, treating 'very toxic' or 'toxic' labels as toxic and the rest as not-toxic.

Table 1 shows the F1-score, precision and recall between the Perspective classifier and the majority of Mturk labelers, and similarly, metrics for the baseline single labeler

[9] We use the 5th iteration of their classifier (TOXICITY@5) and binarize the toxicity score, classifying comments with score >0.5 as toxic. Varying this threshold to 0.7 and 0.9 did not qualitatively change the results.

[10] We do include removed comments in calculating individual contributors' toxicity levels, as we want to measure the individual's willingness to post toxic material, not the community's willingness to accept it.

[11] https://blog.google/technology/ai/new-york-times-using-ai-host-better-conversations/

[12] https://github.com/conversationai/perspectiveapi/blob/master/2-api/model-cards/English/toxicity.md

[13] We selected only raters based in the US who had previously completed atleast 1000 accepted tasks on MTurk. We paid 0.10$ per comment, averaging to 12$ an hour.

Classifier	Overall			AskEconomics			WhiteRights			Other Subreddits		
	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
Perspective	0.769	0.817	0.726	0.667	1.0	0.5	0.774	0.837	0.72	0.764	0.763	0.765
Baseline	0.731	0.696	0.773	0.411	0.284	0.792	0.799	0.81	0.791	0.632	0.563	0.723

Table 1: Evaluation of Perspective classifier performance against MTurk labelers. We find that overall, the Perspective classifier gives us a better approximation of the ground truth than we would get using a single human labeler, except on the most toxic subreddit, r/WhiteRights where a single labeler was slightly better.

compared to the the majority of Mturk labelers.[14] Inferring from Table 1, overall, the Perspective classifier gives us a better approximation of the ground truth than we would get using a single human labeler, except on the most toxic r/WhiteRights subreddit, where a single labeler was slightly better, making it a reasonable choice for identifying toxic content on Reddit.

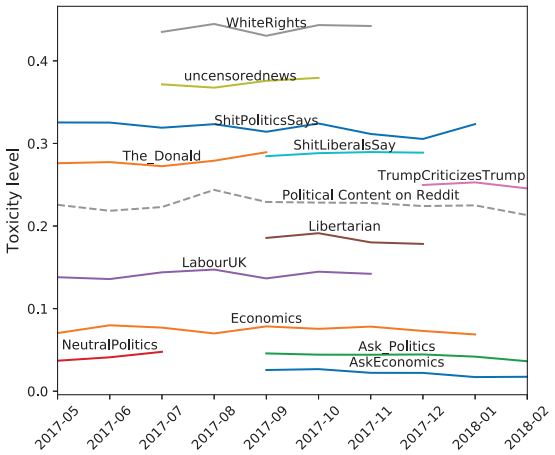

Figure 3: A sample of stable periods of subreddits with distinctive toxicity norms: The y-axis is toxicity level (mean toxicity per month). The grey dotted line represents the overall average toxicity per month in all political subreddits. The colored lines indicate the stable period of a subreddit (name of the subreddit, above the line). This graph includes only a subset of subreddits for visual interpretability.

4.6 Identifying communities with stable and distinctive norms

Among the political subreddits we identify in Section 4.4, we chose communities that exhibit stable toxicity levels. A community whose toxicity measure shows a major fluctuation from month to month may be in transition or may not have clear norms, so it will be harder to notice whether and how newcomers are adapting to those norms. Also, we specifically focus on communities that have distinctive norms – toxicity levels that are different from the overall average toxicity in Reddit's political communities. These communities will require more newcomer adaptation because the normative behavior is uncommon in the larger Reddit political discussion environment. Periods of stable, uncommon norms provide the best opportunity to understand how newcomers adapt. Note that we are not trying to predict whether newcomers conform to community norms. Instead, we are aiming to understand the strength of different norm conforming processes when people do conform. Thus, we select communities to study that have distinctive and stable toxicity norms as follows:

1. We consider a community to have stable toxicity norms if the frequency of toxic content posted in the community per month is fairly constant. Specifically, we define that a community has a stable toxicity norm over time period t if the maximum difference in toxicity levels recorded per month during t does not exceed 0.02 ($\Delta tox_{max} \leq 0.02$).
2. For each community C, we construct rolling time windows of t months ($t = 3$ to $t = 10$ to account for the whole length of the dataset, April 2017 - February 2018) and select windows that have $\Delta tox_{max} \leq 0.02$.
3. Among subreddits with stable windows, we identify windows with distinctive toxicity levels as those with mean toxicity levels more than 0.02 points ($> 2\sigma$) away from the mean overall toxicity level in political communities during the same time period.
4. If there are multiple stable windows for a community, we choose the longest window. If there are multiple such windows, we choose the window with the most number of participants. We call these selected windows stable periods.
5. In some cases, a community's norm may be neither distinctive nor stable during any time window. We omit those communities from our analysis.

We identify 65 political subreddits that have a period of stable and distinctive toxicity levels, a subset of which is shown in Figure 3 for visual interpretability; the complete list of 65 subreddits with their toxicity is shown in Figure 2 along with details about user activity and comments posted per month in each subreddit.

From Figure 3, we make two key observations: (1) communities have stable norms at different, sometimes overlapping time periods (ii) the overall toxicity of comments in political subreddits (grey dotted line), though reasonably stable, is not entirely constant over months. Therefore, to make accurate comparisons between subreddits that are stable during different time periods, we measure the *distinctive toxicity*. A subreddit's distinctive toxicity ($\hat{G}$) is the difference between its average toxicity and overall average for all political subreddits during the same time period:

[14] We performed cross-validation, selecting eleven labels for ground truth and one for the human baseline each time. All results are averaged over the twelve folds.

$$\hat{G} = \mu_C - \mu_{pol}$$

where,

1. μ_C is the toxicity level of community C at its stable time period.
2. μ_{pol} is the mean of the overall toxicity exhibited in political subreddits during stable period of community C.

5 Modeling norm conforming processes

To quantify the effect of norm conforming processes on newcomers for each community identified, we estimate the probability of posting toxic comments in different posting contexts characterized in Section 3.

For any given subreddit, for each month t in that subreddit's stable period, we select comments from *joiners*, people who have never posted in C previously and posted for the first time in C during month t, and *non-joiners*, people who did not participate in C up through month t. For each joiner, we include comments they posted in C during month t but also comments they posted in other subreddits in the previous and following month. For each non-joiner, we include comments during the previous month $t-1$. As we have many *non-joiners*, we sample a set of 1000 such users for each month and community. For each comment, we have the user u who posted it, the month t that is was posted, whether the comment was civil or not, and which of the following comment groups it belongs to.

1. $elsewhere\_last\_month$. All the comments from joiners in month $t-1$ in other political subreddits.
2. $control\_last\_month$. All the comments from non-joiners in month $t-1$ in other political subreddits.
3. $first$. The first comment in C during month t for each joiner who posted at least one comment in another political subreddit in month $t-1$.
4. $first \| returner$. The first comment in C during month t for each joiner who posted in C more than once during month t.
5. $first \| exiter$. The first (only) comment in C during month t for each joiner who posted in C exactly once during month t.
6. all_this_month. All the comments in C during month t from joiners.
7. $elsewhere\_next\_month$. All the posts by joiners in other political subreddits in month $t+1$. This is restricted to only joiners who also commented in another political subreddit in month $t-1$.

To calculate the effects of norm conforming processes, we estimate the toxicity levels for each of aforementioned groups. To do this, for each subreddit, we conduct a mixed effects logistic regression using the $lme4$ package (Bates et al. 2014) modeling toxicity of users. The count of toxic posts is modeled as the number of *successes* and the total posts as the number of Bernoulli *trials* in a binomial distribution.

Concretely, for each subreddit, we estimate the following model:

$$T_{u,t,g} = Binomial(P(toxicity), N_{u,t,g})$$
$$P(toxicity) = logit(\alpha_u + \Sigma\beta_t month_t + \Sigma\beta_g G_g)$$

We model the count of toxic posts $T_{u,t,g}$ made by user u in each comment group g in month t. $N_{u,t,g}$ is the total comments made by user u in each comment group g in month t. The independent variables of interest are the dummy variables G_g for the comment groups described above. In addition, we include a fixed effect, $month_t$, for each month t to account for any time-specific phenomena that affect all messages in that month. Finally, we include a random effect for user, α_u, to account for individual variability and to model the correlation among a user's own posts.

We can interpret the coefficient for a message group (β_g) as the toxicity level of an average user's comments in that message group.[15] Since some users post more often than others, this estimate is not the same as the mean toxicity of the comments in that subgroup. This approach prevents some extremely prolific users from having outsized influence, as they would if we estimated the mean toxicity of comments.

The coefficients β_g ($\beta_{elsewhere\_last\_month}$, $\beta_{control\_last\_month}$, β_{first}, $\beta_{first\|returner}$, $\beta_{first\|exiter}$, $\beta_{all\_this\_month}$, $\beta_{elsewhere\_next\_month}$) are the key estimates that are used to calculate the size of the different norm-conforming processes as described in Section 3. The significance threshold for each of the β_t parameters is adjusted via Bonferroni correction during model estimation at $p = 0.05$, so that the effective p-value is $p = 0.05/7 = .007$.

To illustrate, the subreddit r/NeutralPolitics (marked in magenta (•) in Figures 4 and 5) had distinctive toxicity $\hat{G} = -0.180$ from May-July 2017, reflecting that 4.2% of its comments were toxic as compared to the overall mean 22.2%. We estimated a selection effect $SE = -0.029$, indicating that people who joined made 2.9% fewer toxic comments than people who did not join. The estimated pre-entry learning effect was $LPre = -0.132$, meaning that 13.2% less of newcomers' comments in r/NeutralPolitics were toxic than their comments in other subreddits. The estimated retention effect was $RE = -0.008$ and the estimated post-entry learning effect was $LPost = +0.007$. The estimated transformational learning was $TL = -0.001$, meaning that the newcomers' posts in other subreddits had essentially the same toxicity before and after joining r/NeutralPolitics. Indeed, the standard errors of the estimates reveal that RE, $LPost$, and TL were not statistically significant ($p > 0.05$).

As another illustration with a more toxic subreddit, r/uncensorednews marked in red (•) in Figures 4 and 5 had

[15] The same comment may appear in multiple comment groups. For example, the first message from a joiner who also posts other messages will be duplicated in groups $first$ and $first \| returner$. To ensure that the means of the comment groups are estimated accurately, we recode the user ids to mark messages in different message groups to have different user ids. This strategy ensures that there is still partial pooling when we estimate random effects for each user, but we lose some constraints by allowing multiple random intercept values for the different aliases of the same user.

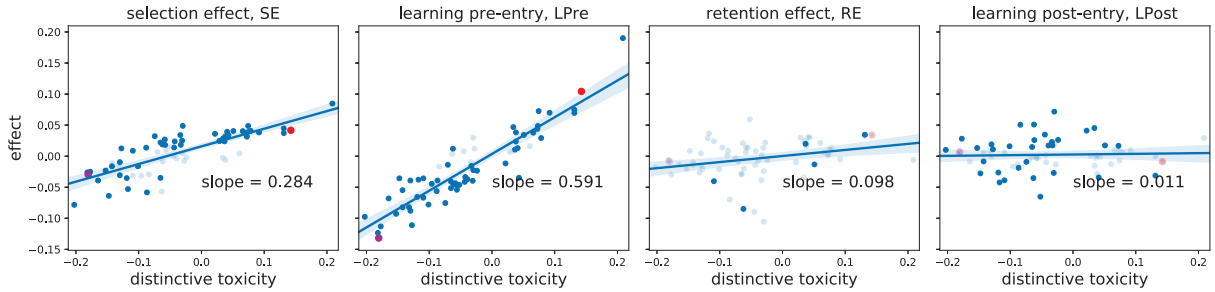

Figure 4: Comparing the effect of norm conforming processes ($SE, LPre, RE, LPost$) against the distinctive toxicity of existing users. The points correspond to the 65 subreddits with distinctive toxicity norms. The points are lighter if the effect is not statistically significant at $p = 0.05$. r/NeutralPolitics is marked in magenta (•) and r/uncensorednews in red (•).

$\hat{G} = +0.143$ from July-October 2017. The estimated coefficients for it were $SE = +0.042$, $LPre = +0.104$, $RE = +.0034$, $LPost = -0.009$, and $TL = +0.003$. Similar to r/NeutralPolitics, the standard errors of the estimates reveal that RE, $LPost$, and TL were not statistically significant ($p > 0.05$).

6 Results

6.1 Learning Pre-Entry contributes most to norm conformity

Using the estimates of the different processes for each subreddit, we construct Figure 4 to measure the strength of each norm conforming process across subreddits[16]. Figure 4 shows one graph for each of the processes, with the size of that process' effect shown on the y-axis. Each point represents one subreddit. On each graph, points where the effect of the process is not statistically significant are lightly shaded. To provide context, we plot these effect sizes against the distinctive toxicity norm ($\hat{G}_{sub}$) on the x-axis. As we include only subreddits with distinctive norms, there are no subreddits with distinctive toxicity scores in the range $[-.02, +.02]$[17]. The effect sizes on the y-axis are comparable across sub-figures and can be interpreted as being in the same units as the amount of distinctiveness which is shown on the x-axis. The reason is that all of the beta coefficients that are used to define these effect sizes are conveying the mean toxicity of the users for some group of messages.

The left-most sub-figure shows the size of the selection effect, the difference between the behavior in the previous month of joiners and non-joiners. For many communities, this effect is statistically significant, but it is rarely larger than five percentage points, even for communities with very distinctive norms (fifteen or twenty percentage points away from the mean of all political subreddits.) The best fit line shows that each extra percentage point of distinctiveness leads to attracting people who were .284 points closer to the community's norm.

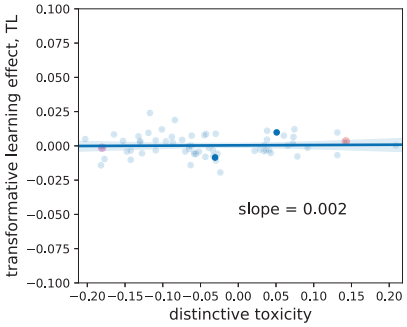

Figure 5: Comparing the transformative learning effect to the distinctive toxicity norm for each community. The points represent each subreddit, lighter shades indicate that the transformative learning effect for that subreddit is not statistical significant (at $p = 0.05$). The flat line suggests that users return to their past behavior after interacting with the community. r/NeutralPolitics is marked in magenta (•) and r/uncensorednews in red (•).

The second sub-figure shows the size of pre-entry learning, the difference between joiners' behavior in other communities and their first post in this community. For almost all communities, this effect is statistically significant, and it is also noticeably larger than the selection effect. The best fit line shows that each extra percentage point of distinctiveness leads to the average newcomer adjusting their behavior by .591 points closer to the community's norm.

The third sub-figure shows the size of retention, the difference between the behavior of joiners who leave and joiners who stay in the community. For almost all communities, this effect is not statistically significant, and it is also noticeably smaller than the pre-entry learning effects. The best fit line shows that each extra percentage point of distinctiveness leads to the average newcomer adjusting their behavior by .098 points closer to the community's norm.

[16]To ensure that our findings are robust to behavior of a few prolific users in these subreddits, we removed the top 5% of the most active users in each subreddit (95th percentile or higher) and performed the same analysis. Removing the most prolific users did not materially change the estimated effects and the resulting slopes in these graphs.

[17]When including subreddits with non-distinctive toxicity scores, all effect sizes are small and do not materially change the slope in these graphs.

The final sub-figure shows the size of post-entry learning, the difference between joiners' behavior in their first comment and their behavior in the first month. For almost all communities, this effect is not statistically significant, and with the effect size near 0. The best fit line shows that each extra percentage point of distinctiveness leads to practically no adjustment towards the community's norm.

6.2 No Transformative learning

Figure 5 compares the transformative learning effect TL against the distinctive toxicity $\hat{G}$ of the corresponding subreddit. A positive slope would indicate the presence of transformative learning, as in that case, newcomers' toxicity in the future month would be closer to the norm of the community the user joined in the previous month. However, we observe no such learning effect. Users are not "transformed" and continue their usual previous behavior in other communities in the month after joining.

7 Discussion

We find that pre-entry processes, in general, are much more influential than post-entry processes for newcomer conformity. Even among pre-entry processes, given the extensive evidence of homophily (McPherson, Smith-Lovin, and Cook 2001), one would reasonably expect self-selection to play a crucial role in norm conformity. While self-selection does play an important role, we find that other pre-entry learning processes contribute most to matching newcomers toxicity level to that of the community. For communities with distinctive toxicity norms, newcomers on average seem to know enough about a community's toxicity norms *before* they engage with the community they are joining, and adjust the behavior they exhibit in other political subreddits.

One possible explanation for high pre-entry learning is that toxicity norms are relatively easy to grasp by observation. Users can probably tell how much a community tolerates toxicity through a glance of the community's rules as well as the past comments made by existing users. Unlike more complicated norms such as the norm of "suspended disbelief" in r/NoSleep subreddit which "requires all commenters to act as if stories are factual" (Kiene, Monroy-Hernández, and Hill 2016), toxicity norms likely take less time and effort to absorb, leading to high pre-entry learning.

Another reasonable explanation is the role of lurking in learning norms of the community (Preece, Nonnecke, and Andrews 2004). In this work, we define *joining a community* as the first time the user posted a public comment. In reality, users may have "joined" a community much earlier and remained a passive learner, observing the norms of the community. Thus, the phenomenon of almost instantly adjusting behavior to match the community norms may actually be a more drawn-out process invisible to the researcher without user log data. Regardless of how much time it might take for a passive lurker to turn into contributor, we observe that when users do switch to posting, on average they nearly match the norms of the community quickly.

We find that, on average, there appears to be little or no real retention effect and learning post-entry. This is counter-intuitive given that most research on norm conforming focuses on post-entry socialization (Choi et al. 2010; Ren, Kraut, and Kiesler 2007; Ducheneaut 2005; Burke, Kraut, and Joyce 2010). We speculate that socialization tactics (e.g., verbal sanctions against norm violators) impact not only newcomers after joining, but also prospective newcomers who observe and possibly gain insights into community norms before posting. Further, given that post-entry learning, by definition, occurs after the observed strong pre-entry learning and since we find that first posts largely match the community norms, there is little need or opportunity for post-entry learning.

We find that there is little evidence of transformative learning, highlighting the situational nature of norms. It appears that users return to their previous behavior in the other communities they participate in. The relative ease and speed with which individuals adjust to community toxicity norms can help explain the lack of spillover effects. This finding is also consistent with past work on Reddit communities which found that users' language style does not carry over and is different in different communities (Tan and Lee 2015). Further, this observation that users, on average, don't carry forward their behavior in one community to another community supports Chandrasekharan et al. (2017a)'s finding that after Reddit banned hate subreddits in 2015, members of those subreddits did not engage in similar hate speech in other subreddits that they subsequently participated in.

8 Implications

Our results provide an existence proof that most communities, even ones that might seem unruly in their discussions, are able to reproduce toxicity norms largely by getting new members to adhere to them even in their first posts. Therefore, we recommend that communities invest in making their norms more visible to prospective newcomers by posting explicit guidelines and highlighting exemplars. Further, we recommend that, whenever possible, communities make certain moderator actions visible (such as leaving a visible trace of deletions or explanations for deletions (Jhaver, Bruckman, and Gilbert 2019)) so that these actions have ripple effects on prospective newcomers, allowing them to learn about the community norms through observation before joining. Finally, since there is no transformative learning and users are quick to adjust to community norms, we infer that, on average, a user's behavior in one subreddit is not indicative of behavior outside that subreddit. Thus, barring extreme situations, we recommend that moderators primarily judge newcomers by what they do in the community only and not based on their past behavior elsewhere.

9 Limitations and Future Work

Pre-entry learning aggregates three different ways that people may learn: observing others' messages, reading posted guidelines, and having messages removed by moderators. Post-entry learning includes these as well as feedback from the community (e.g. upvotes/downvotes). Lacking access to user logs, we cannot reliably separate out the individual drivers of learning. Instead we consider them in aggregate

and focus on the phases of learning. Also, we note that there is a possibility that we are missing some deleted comments. In particular, comments removed by the AutoModerator bot are invisible to us, as those comments are removed almost as soon as they are published (comments removed by human moderators were visible to us through the PushShift API). While the use of moderator bots are not necessarily driven by toxicity norming goals, past work has shown that use of auto-moderators can curtail abusive language (Young 2018). It is possible that the Automoderator bot deletes a large number of toxic first-time comments by new users. Hence, the first comment we identify for a subset of users might not be the first comment they composed. However, we observe that even newcomers in *distinctively toxic* communities on average match the *higher* levels of toxicity when posting for the first time, a phenomenon which can't be explained by AutoModerator removals of toxic comments.

In this work, we focus our analysis on toxicity norms. It would be interesting to analyze other norms of discussion, such as reciprocity, in a future study to validate if the *pre-entry learning* is equally strong for other norms. Another extension to this study would be to understand how these effects relate to platform logics and algorithmic curation. Though this work provides insights into different norm conforming processes, we have not identified why the strength of some of these effects are different in different communities, a possible avenue for future research.

10 Conclusion

In this work, we have examined the processes that affect norm conformity among newcomers in political communities on Reddit. Interestingly, we find that pre-entry norm conforming processes contribute much more than post-entry processes to norm conformity. Specifically, we find that most of the norm conforming occurs through *pre-entry learning* – newcomers adjust to community toxicity norms while making their first comment. We also find that the adjustments made to conform to community norms are not permanently transformative. Users conform when participating in that community but continue to behave differently in other communities. Since selection effects are small, and there is little personal transformation, we conclude that compatible newcomers are neither born nor made; they merely adjust.

11 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1717688. We thank Jeff Lockhart for data collection, Summer Nyugen for help in vizualization, and the ARC-TS team at Michigan for Cavium-Thunderx Hadoop cluster support. We also thank participants of the Computational Social Science Seminar for feedback on earlier drafts.

References

An, J.; Kwak, H.; Posegga, O.; and Jungherr, A. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 68–79.

Aroyo, L.; Dixon, L.; Thain, N.; Redfield, O.; and Rosen, R. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, 1100–1105. ACM.

Bandura, A. 1978. Social learning theory of aggression. *Journal of communication* 28(3):12–29.

Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Blackwell, L.; Dimond, J.; Schoenebeck, S.; and Lampe, C. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW):24:1–24:19.

Borah, P. 2014. Does It Matter Where You Read the News Story? Interaction of Incivility and News Frames in the Political Blogosphere. *Communication Research* 41(6):809–827.

Burke, M.; Kraut, R.; and Joyce, E. 2010. Membership claims and requests: Conversation-level newcomer socialization strategies in online groups. *Small group research* 41(1):4–40.

Burke, M.; Marlow, C.; and Lento, T. 2009. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI conference on human factors in computing systems*.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017a. You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW):31:1–31:22.

Chandrasekharan, E.; Samory, M.; Srinivasan, A.; and Gilbert, E. 2017b. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 3175–3187. New York, NY, USA: ACM.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):32.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial Behavior in Online Discussion Communities. In *ICWSM*, 61–70.

Choi, B.; Alexander, K.; Kraut, R. E.; and Levine, J. M. 2010. Socialization tactics in wikipedia and their effects. In *Proceedings of the conference on Computer supported cooperative work*. ACM.

Cialdini, R. B.; Reno, R. R.; and Kallgren, C. A. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology* 58(6):1015.

Coe, K.; Kenski, K.; and Rains, S. A. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64(4):658–679.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW.

Dennen, V. P. 2014. Becoming a blogger: Trajectories, norms, and activities in a community of practice. *Computers in Human Behavior* 36:350–358.

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

Fiesler, C., and Proferes, N. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*.

Fiesler, C.; McCann, J.; Frye, K.; Brubaker, J. R.; et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

Geiger, R. S., and Ribes, D. 2010. The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*.

Gervais, B. T. 2015. Incivility Online: Affective and Behavioral Reactions to Uncivil Political Posts in a Web-based Experiment. *Journal of Information Technology & Politics* 12(2):167–185.

Guimaraes, A.; Balalau, O.; Terolli, E.; and Weikum, G. 2019. Analyzing the traits and anomalies of political discussions on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 205–213.

Jhaver, S.; Bruckman, A.; and Gilbert, E. 2019. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3(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. ACM.

Kiene, C.; Monroy-Hernández, A.; and Hill, B. M. 2016. Surviving an "Eternal September" – How an Online Community Managed a Surge of Newcomers.

Kiesler, S.; Kraut, R.; Resnick, P.; and Kittur, A. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press*.

Kraut, R.; Burke, M.; Riedl, J.; and Resnick, P. 2010. Dealing with newcomers. *Evidencebased Social Design Mining the Social Sciences to Build Online Communities* 1:42.

Kwak, H.; Blackburn, J.; and Han, S. 2015. Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 3739–3748. ACM.

Lampe, C., and Johnston, E. 2005. Follow the (Slash) Dot: Effects of Feedback on New Members in an Online Community. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '05, 11–20. ACM.

Leurs, K., and Zimmer, M. 2017. *Platform values: an introduction to the# AoIR16 special issue*. Taylor & Francis.

Levine, J. M.; Choi, H.-S.; and Moreland, R. L. 2003. Newcomer innovation in work teams. *Group creativity: Innovation through collaboration* 202–224.

Lin, Z.; Salehi, N.; Yao, B.; Chen, Y.; and Bernstein, M. S. 2017. Better when it was smaller? community content and behavior after massive growth. In *ICWSM*, 132 – 141.

Massanari, A. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19(3):329–346.

Matias, J. N. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116.

McPherson, M.; Smith-Lovin, L.; and Cook, J. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*.

Mittos, A.; Zannettou, S.; Blackburn, J.; Cristofaro, D.; and Emiliano. 2020. "and we will fight for our race!" a measurement study of genetic testing conversations on reddit and 4chan. In *ICWSM*.

Nguyen, D., and Rosé, C. P. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media*, 76–85. Association for Computational Linguistics.

Nithyanand, R.; Schaffner, B.; and Gill, P. 2017. Online political discourse in the trump era. *arXiv preprint arXiv:1711.05303*.

Norris, P. 2002. The bridging and bonding role of online communities.

Panciera, K.; Halfaker, A.; and Terveen, L. 2009. Wikipedians are born, not made: a study of power editors on wikipedia. In *Proceedings of international conference on Supporting group work*. ACM.

Papacharissi, Z. 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society* 6(2):259–283.

Park, D.; Sachar, S.; Diakopoulos, N.; and Elmqvist, N. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 1114–1125.

Postmes, T.; Spears, R.; and Lea, M. 2000. The formation of group norms in computer-mediated communication. *Human communication research* 26(3):341–371.

Preece, J.; Nonnecke, B.; and Andrews, D. 2004. The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior* 20(2):201–223.

Reicher, S. D.; Spears, R.; and Postmes, T. 1995. A social identity model of deindividuation phenomena. *European review of social psychology* 6(1):161–198.

Ren, Y.; Kraut, R.; and Kiesler, S. 2007. Applying common identity and bond theory to design of online communities. *Organization studies* 28(3):377–408.

Robinson, D. T.; Smith-Lovin, L.; and Wisecup, A. K. 2006. Affect control theory. In *Handbook of the sociology of emotions*. Springer.

Ross, L., and Nisbett, R. E. 2011. *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678.

Schilling, A.; Laumer, S.; and Weitzel, T. 2012. Who will remain? an evaluation of actual person-job and person-team fit to predict developer retention in floss projects. In *2012 45th Hawaii International Conference on System Sciences*, 3446–3455. IEEE.

Soliman, A.; Hafer, J.; and Lemmerich, F. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 259–263. ACM.

Stryker, R.; Conway, B. A.; and Danielson, J. T. 2016. What is political incivility? *Communication Monographs* 83(4):535–556.

Tan, C., and Lee, L. 2015. All Who Wander: On the Prevalence and Characteristics of Multi-community Engagement. In *Proceedings of the International Conference on World Wide Web*, WWW '15.

Tonteri, L.; Kosonen, M.; Ellonen, H.-K.; and Tarkiainen, A. 2011. Antecedents of an experienced sense of virtual community. *Computers in Human Behavior* 27(6):2215–2223.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1391–1399.

Young, L.-Y. 2018. The effect of moderator bots on abusive language use. In *Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence*, 133–137. ACM.

Zimmer, M., and Kinder-Kurlanda, K. 2017. *Internet research ethics for the social age: New challenges, cases, and contexts*. Peter Lang International Academic Publishers.