Understanding Community Resilience: Quantifying the Effects of Sudden Popularity via Algorithmic Curation

Jackie Chan[1], **Charlotte Lambert**[1], **Fred Choi**[1], **Stevie Chancellor**[2], **Eshwar Chandrasekharan**[1]

[1] University of Illinois Urbana-Champaign
[2] University of Minnesota Twin Cities
{jackiec3, cjl8, fc20, eshwar}@illinois.edu, steviec@umn.edu

Abstract

The sudden popularity communities gain via algorithmically-curated "trending" or "hot" social media feeds can be beneficial or disruptive. On one hand, increased attention often brings new users and promotes community growth. On the other hand, the unexpected influx of newcomers can burden already overworked moderation teams. To examine the impact of sudden popularity, we studied 6,306 posts that reached Reddit's front page—a feed called *r/popular* that millions of users browse daily—and the effects of sudden popularity within 1,320 subreddits. We find that on average, r/popular posts have 45 times the comments, 42 times the removed comments, and 70 times the number of newcomers compared to posts from the same community that did not reach r/popular. Additionally, r/popular posts led to a peak 85% median increase in the subreddit's comment rate, and these effects lingered for about 12 hours. Our regression analysis shows that stricter moderation and previous r/popular appearances were associated with shorter and less intense effects on the community. By quantifying the differential effects of sudden popularity, we provide recommendations for moderators to promote stability and community resilience in the face of unexpected disruptions.

Introduction

Online communities are vital in both social and professional life. However, online communities often undergo disruptive periods induced by events ranging from media attention (Chandrasekharan et al. 2017) to malicious, organized trolling efforts (Kumar et al. 2018). These disruptions frequently increase activity from non-regular users (Kiene, Monroy-Hernández, and Hill 2016) and challenge community management and disrupt regular users' participation.

Disruptions are often caused by algorithmically-curated "hot" or "trending" social media feeds, bringing drastic attention to highlighted content (Chan et al. 2022). Algorithms that cause popularity spikes are often opaque about how they operate (Eslami et al. 2015), and thus create unexpected moments of popularity or virality. Prior work on virality focuses on the content that became viral (Berger and Milkman 2012) or the users who create content (Gurjar et al. 2022).

However, algorithmic curation can disrupt communities, precipitating negative effects from unexpected surges in attention (Cheng et al. 2014). These volatile periods are critical because they can make or break a community, where they can either thrive and gain new membership or falter and destabilize the community (DeVito 2022; Zhang et al. 2021). Moreover, sudden popularity can substantially increase moderator workload and hurt existing members' ability to engage regularly with their communities (Li, Hecht, and Chancellor 2022). Previous studies typically focus on a single community (Kiene, Monroy-Hernández, and Hill 2016), one-time events (Matias 2016; Horta Ribeiro et al. 2021), or niche communities that require expertise, such as Wikipedia (Zhang et al. 2019) or GitHub (Maldeniya et al. 2020). However, we have not studied how popularity spikes affect communities generally or whether communities are impacted differently when exposed to unexpected surges in attention. By studying sudden popularity spikes, we can understand what factors influence how distinct communities emerge after their popularity subsides.

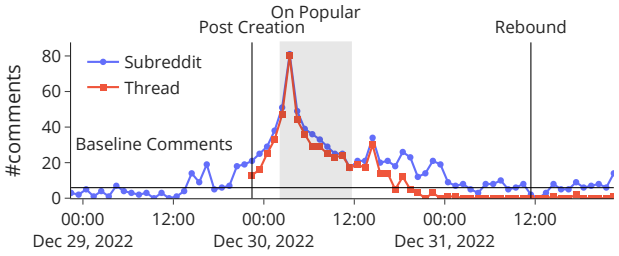

Figure 1: An example of an r/popular post's impact on a subreddit's activity (in UTC). *Baseline comments* refers to the average number of comments per hour on the subreddit during the week before the r/popular post was created.

Our Contribution

We present a large-scale study to quantify the differential effects of sudden popularity introduced by algorithmically-curated feeds on Reddit posts and communities. We do this by studying posts becoming "popular" through Reddit's r/popular feed.[1] The r/popular feed is an algorithmically-curated feed of active, highly-voted content that populates

[1] https://www.reddit.com/r/announcements/comments/5u9pl5/introducing_rpopular/

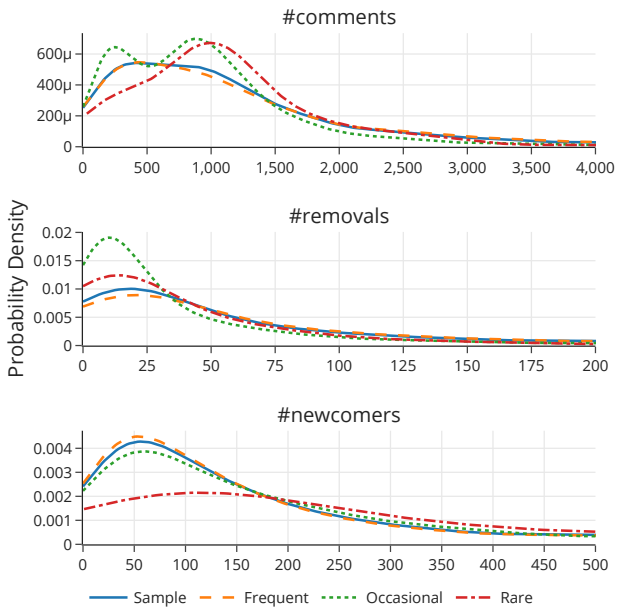

Figure 2: KDE plots illustrating the distributions for each dataset (seen in Table 1) and outcome variable.

Reddit's home page, reaching millions of daily users who browse Reddit. Despite Reddit not revealing how it curates the r/popular feed, from our analysis of thread-level outcomes, the feed consistently contains posts that are 45 times more active than their non-r/popular counterparts, showing r/popular's exceptional disruptive capabilities (see Table 2). Auditing r/popular posts allows us to examine the differential impacts of algorithmic hot/trending feeds on "discussion threads" (i.e., r/popular posts and their associated comments) and the respective communities from which these r/popular posts originate. Given this interest in differential effects, we analyzed the number of comments, moderator interventions (in the form of removed comments), and newcomers—which prior literature has noted as being disruptive to communities (Kiene, Monroy-Hernández, and Hill 2016; Lin et al. 2017)—at both a thread and subreddit level. To attend to subreddit-level effects, we examine the longevity and intensity of effects and quantify a community's ability to withstand and rebound from disruptions.

In this paper, we explore the effects of sudden popularity on Reddit posts through the following research questions:

RQ1: How are threads that appear on r/popular affected by sudden spikes in popularity? (*Thread-Level Effects*)

RQ1a: What is the effect of popularity on the number of comments, removals, and newcomers within the thread?

RQ1b: What factors influence such thread-level effects?

RQ2: How does a post appearing on r/popular affect the community it originated from? (*Community-Level Effects*)

RQ2a: How long do community-level effects of popularity last on the number of comments, removals, and newcomers participating in the community (i.e., *longevity*)?

RQ2b: How intense are the community-level effects?

RQ2c: What factors affect *longevity* and *intensity*?

To answer these questions, we examined 6,306 r/popular posts collected over 11 months from 1,320 distinct subreddits. From these posts, we find that the type of content affects its thread activity, e.g., link and video content receive more comments, but text posts have more newcomer activity. Additionally, u/AutoModerator presence, an automated moderation tool on Reddit, dampens the subreddit activity spikes at the cost of more comment removals and fewer newcomers within the thread. Lastly, we quantify how r/popular affects communities that infrequently appear on r/popular.

Empirically studying tumultuous periods caused by algorithmic feeds like r/popular can improve our understanding of how sudden popularity impacts communities, helping us to construct mitigating strategies against the undesired outcomes of sudden popularity. Our findings help moderators promote stability and community resilience in the face of unexpected disruptions caused by popularity feeds.

Related Work

Popular Content & Algorithmic Curation

Prior work has studied how content becomes popular, both in offline (Kollock and Smith 1999) and online contexts. This work comes from the early 2010s when researchers focused on predicting cascades (Cheng et al. 2014) and information diffusion (Garg, Smith, and Telang 2011) as social sharing practices mediated virality and popularity. Empirical studies have also predicted what content goes viral on platforms like X, formerly Twitter, and YouTube (Figueiredo, Benevenuto, and Almeida 2011; Weng, Menczer, and Ahn 2013). Recent research has also examined the consequences of popularity shocks on online users (Gurjar et al. 2022).

It is important to understand the mechanisms behind these feeds because they are often the default page on many social platforms, determining what many users perceive as "high quality" information (Ciampaglia et al. 2018). Such feeds are often "black boxes" that do not explain their decisions, so people struggle to reason why content may be featured (Eslami et al. 2015). Moreover, appearing unexpectedly on these feeds is connected to negative impacts on the well-being of individual creators (DeVito 2022) and communities (Lin et al. 2017). In our prior work (Chan et al. 2022), we studied r/popular and the subreddits that appear by correlating r/popular appearances to the daily number of comments, authors, removals, and newcomers. However, we did not examine potential factors that may affect these measures which we do in the current paper along with studying the longevity and intensity of these effects.

Community Success & Responses to Crises

Considering communities directly, what indicates that a community may be doing well? As Cunha et al. (2019) note, the concept of success is overloaded and has been quantified in many ways (Kraut and Resnick 2012). One way is through the size of the community (Cunha et al. 2019). Another strategy is to look at the community's longevity, i.e.,

	# Popular Posts	# Subreddits
Dataset ($\mathcal{D}$)	112,372	1,320
Sample ($\mathcal{D}^{Samp}$)	6,306	1,320
Frequent ($\mathcal{D}^{Freq}$)	5,223	310
Occasional ($\mathcal{D}^{Occa}$)	7,393	543
Rare ($\mathcal{D}^{Rare}$)	611	467

Table 1: The number of r/popular posts in our complete dataset ($\mathcal{D}$), representative 5% sample ($\mathcal{D}^{Samp}$), and subsamples: a 5% sample of subreddits on r/popular more than 50 times during the study period ($\mathcal{D}^{Freq}$), all r/popular posts from subreddits with 3 to 50 posts ($\mathcal{D}^{Occa}$), and all r/popular posts from subreddits with 1 or 2 appearances ($\mathcal{D}^{Rare}$).

how long it persists (Kairam, Wang, and Leskovec 2012). Other papers have proposed alternative metrics, such as how many users drop out or return to the community (Danescu-Niculescu-Mizil et al. 2013; Yang, Kraut, and Levine 2017). Yet another way is by collecting what stakeholders in that community think. For instance, community members may think their small, tight-knit communities are better (Lin et al. 2017) and seek to keep it that way, countering some of the previous quantification narratives. Other factors also influence community success, such as moderator commitment to community management (Li, Hecht, and Chancellor 2022).

Online communities are also prone to crises that can affect their sustainability—and many crises come from sudden changes in a community's popularity or outsider attention. Prior work by Kiene, Monroy-Hernández, and Hill (2016) serves as an excellent qualitative analysis of r/NoSleep, a creative writing subreddit dedicated to horror stories, which experienced a rapid rise in membership. Community leadership, moderator coordination, and automated moderating tools were important in determining the outcome of these turbulent moments (Kiene, Monroy-Hernández, and Hill 2016). Additional examples of crises include repositories that appear on GitHub's trending page (Maldeniya et al. 2020), rapid increases in Wikipedia page editorship (Zhang et al. 2019), and breaking news events on Wikipedia (Keegan, Gergle, and Contractor 2013).

Data Collection & Sampling Approach

Our dataset contains Reddit posts found on the r/popular feed: the default, algorithmically-curated feed that presents the most active content on the platform. We gather additional posts and comments using Pushshift's historical dataset (Baumgartner et al. 2020).

Collecting r/popular Posts from Reddit

To systematically collect r/popular posts in real-time, we used the Reddit API[2] to request a snapshot of the top 100 posts every two minutes. We collected r/popular posts from March 24, 2022, to February 8, 2023, with a brief two-day down period on July 16 and 17, 2022, resulting in 134,661 unique r/popular posts originating from 1,432 subreddits (see Table 1).

[2]https://praw.readthedocs.io/en/stable/

Constructing Our Sample

Before constructing a sample, $\mathcal{D}^{Samp}$, we performed the following filtering steps (n represents the number of drops made during each step sequentially): (1) remove r/popular posts from 11 banned subreddits ($n = 240$), (2) remove the posts that existed on r/popular for less than an hour ($n = 21,963$), and (3) remove prediction tournaments posts as they remained on r/popular for an artificially long time (weeks) due to users *automatically* upvoting them when participating in the embedded poll (Perez 2021) ($n = 86$). These steps resulted in 112,372 r/popular posts from 1,320 subreddits seen in Table 1—we refer to this dataset as $\mathcal{D}$.

After filtering, we shrunk our dataset to a tractable size from $\mathcal{D}$ by sampling 5% of r/popular posts from each subreddit in $\mathcal{D}$. If a subreddit has fewer than 20 r/popular posts (877 subreddits), we select a random r/popular post from that subreddit. This sampling approach became $\mathcal{D}^{Samp}$: our representative sample.

Grouping by Frequency of Popularity

We produced three subsamples to highlight r/popular's differential effects on communities that appeared on r/popular at different rates. Prior work suggests that the impacts of hitting r/popular may differ based on how often they experience unexpected popularity (Kiene, Monroy-Hernández, and Hill 2016; Lin et al. 2017; Chan et al. 2022). However, due to their size, some subreddits appear quite often on r/popular (e.g. r/politics). This means that only analyzing $\mathcal{D}^{Samp}$ may not show the differential effects of popularity on communities that do not hit r/popular often.

Thus, we construct three subsamples, based on the number of appearances on r/popular. First, we construct $\mathcal{D}^{Freq}$, containing 5% of all r/popular posts from subreddits that appeared more than 50 times (approximately once a week during the study period) on r/popular. Similarly, we construct $\mathcal{D}^{Occa}$, with subreddits that appeared occasionally on r/popular ($3 \leq n \leq 50$ times), and $\mathcal{D}^{Rare}$, with subreddits that rarely appeared on r/popular ($n \leq 2$ times). Since $\mathcal{D}^{Occa}$ and $\mathcal{D}^{Rare}$ contained fewer posts compared to $\mathcal{D}^{Freq}$, we did not downsample to preserve our stratified analyses.

Thread-Level Outcomes of Popularity

Next, we detail our methods and findings for r/popular's thread-level effects. To start with, we examine three thread-level outcomes:

1. *Comments:* number of comments the thread received,
2. *Removals:* number of comments on the thread removed by moderators,
3. *Newcomers:* number of comments by users who did not participate in the subreddit since January 2020.

Comments measure activity level in the thread, whereas *removals* quantify the stress put on moderators. This is because removed comments signify that a thread is experiencing norm violations (Chandrasekharan et al. 2018). Flagging *newcomers* estimates activity from users who do not normally participate within the r/popular thread's subreddit.

We examine newcomers because they can be disruptive to both moderators (Zhang et al. 2019) and to regular users who must engage with people missing norms, knowledge, and historical context of a subreddit.

Comparing Against Non-r/popular Threads

Instead of analyzing the raw number of each outcome listed above, we normalize by using the previous week's activity on the subreddit as a baseline to see how much an r/popular thread deviates from the average non-r/popular thread. Thus, for each r/popular thread, we divide the number of comments, removals, and newcomers by the numbers seen on non-r/popular threads in the previous week. For example, if an r/popular thread receives 1,000 comments and non-r/popular threads on the respective subreddit in the previous week receive 50 comments on average, then we get 20 which represents the *multiplicative increase* from a normal, non-r/popular thread. We compute this for removals and newcomers as well. We conduct a second evaluation by filtering for the top 5th percentile of non-r/popular threads in each outcome measure and calculating the same multiplicative increase. Comparing thread-level outcomes for r/popular posts against posts that attained "organic popularity" (i.e., top 5th percentile) within the same community allows us to estimate the effects of sudden popularity introduced via Reddit's r/popular feed versus organic popularity. The comparisons for both baselines can be seen in Table 2a and 2b.

Thread Results (RQ1a)

Table 2 shows the *median* multiplicative increase from non-r/popular threads to r/popular threads from the same subreddit. To see if the distributions for multiplicative increases are different from one another, we performed a Kruskal-Wallis test on each outcome (H-statistics in Table 2). All Kruskal-Wallis tests returned a $p < 0.01$, meaning at least one distribution is significantly different. Thus, a post hoc test, specifically the Conover-Iman test, is required to identify the pairwise differences.

For Table 2a, the post hoc tests revealed that $\mathcal{D}^{Rare}$ is statistically different than all other datasets, including $\mathcal{D}^{Samp}$, for all three outcomes (everything stated is $p < 0.01$, Bonferroni corrected). Similarly, $\mathcal{D}^{Occa}$ is statistically different from $\mathcal{D}^{Freq}$ for all outcomes except #comments. Lastly, $\mathcal{D}^{Samp}$ and $\mathcal{D}^{Occa}$ difference with #newcomers was statistically significant. Relationships not mentioned were not statistically significant.

For Table 2b, $\mathcal{D}^{Rare}$ is again statistically different than all other datasets ($p < 0.01$, Bonferroni adjusted). $\mathcal{D}^{Occa}$ has statistically significant differences when compared to $\mathcal{D}^{Samp}$ and $\mathcal{D}^{Freq}$ for #comments and #newcomers, but not #removals. Comparing $\mathcal{D}^{Samp}$ and $\mathcal{D}^{Freq}$, we only find that the difference for #newcomers is significant.

In short, r/popular's raw effects differ based on the number of prior appearances of a subreddit on r/popular. The multiplicative impact is largest with subreddits that only appear once or twice a year, i.e., $\mathcal{D}^{Rare}$.

	#comments	#removals	#newcomers
$\mathcal{D}^{Samp}$	44.59	42.42	70.12
$\mathcal{D}^{Freq}$	40.73	40.58	65.96
$\mathcal{D}^{Occa}$	40.49	44.71	79.74
$\mathcal{D}^{Rare}$	52.24	61.57	124.74
H-stat.	41.50	49.03	160.0

(a) All Threads

	#comments	#removals	#newcomers
$\mathcal{D}^{Samp}$	3.56	2.74	4.84
$\mathcal{D}^{Freq}$	3.37	2.69	4.41
$\mathcal{D}^{Occa}$	3.89	2.67	5.76
$\mathcal{D}^{Rare}$	5.76	3.43	10.46
H-stat.	152.12	25.80	380.88

(b) Top 5th Percentile

Table 2: Median multiplicative increases, representing the magnitude difference between non-r/popular and r/popular thread activity. The first table considers all threads in the previous week whereas the second considers only highly active non-r/popular threads in the previous week. The last row contains the H-statistics ($df = 3, p < 0.01$) for the Kruskal-Wallis tests.

			Rebound Time		
	To Pop.	On Pop.	Com.	Rem.	New.
$\mathcal{D}^{Samp}$	4:39	6:00	2:19	2:23	2:10
$\mathcal{D}^{Freq}$	4:24	6:12	1:47	2:14	1:11
$\mathcal{D}^{Occa}$	6:12	5:11	4:15	3:02	5:03
$\mathcal{D}^{Rare}$	6:39	5:02	8:57	3:14	10:23
H-stat.	1,093.24	81.10	476.47	73.01	1,001.05

Table 3: Median lifespan, defined in our community-level outcomes section, presented in hour : minute format. The last row contains the H-statistics ($df = 3, p < 0.01$) for the Kruskal-Wallis tests performed.

Community-Level Outcomes of Popularity

We quantify the impact of an r/popular post on its respective subreddit for RQ2 by defining two features: (1) the r/popular post's *lifespan* which measures the effect's longevity and (2) *intensity* which assesses the effect's disruptiveness.

Computing Longevity of Subreddit Effects

To capture how long an r/popular post impacts its home subreddit, we define a post's *lifespan* as three intervals (see Figure 1): (1) the time the post took to reach r/popular from post creation, (2) the duration the post is on r/popular top 100, and (3) *the rebound time*, which is the time the subreddit took to return to baseline activity levels. We define "baseline" as 1.1 times the average hourly rate in the previous week. Rebound times are outcome-variable specific, so each r/popular post has three lifespans. We will use the lifespan as a "treatment period" on posts. Table 3 contains details on how long each interval is for each dataset.

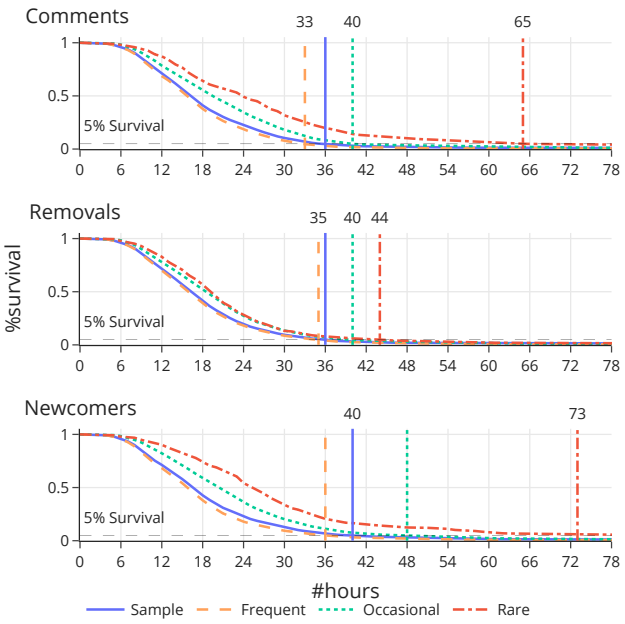

Figure 3: The survival probability that a subreddit will still have above baseline activity at a certain hour, outcome, and dataset. Vertical lines denote when 95% of subreddits return to baseline activity for their respective color.

Longevity Results (RQ2a)

To identify statistically significant differences, we performed Kruskal-Wallis tests on each interval, which all were significant ($p < 0.01$). H-statistics are in the last row in Table 3. Regarding the time to reach r/popular, only the difference between $\mathcal{D}^{Rare}$ and $\mathcal{D}^{Occa}$ was insignificant ($p > 0.01$) according to the pairwise Conover-Iman post hoc tests. All other pairs fell below our significance threshold. The significant relationships for on r/popular times were sparse. We find that $\mathcal{D}^{Rare}$ and $\mathcal{D}^{Occa}$ are significantly different compared to $\mathcal{D}^{Freq}$ ($p < 0.01$), all other pairs do not pass our significance threshold. Lastly, all pairwise tests for comment and newcomer rebound times were statistically significant ($p < 0.01$). However, for removal rebound times, the differences between $\mathcal{D}^{Samp}$ and $\mathcal{D}^{Freq}$ and between $\mathcal{D}^{Occa}$ and $\mathcal{D}^{Rare}$ were both not significant ($p > 0.01$). All other pairs for removal rebound times were statistically significant according to our post hoc tests.

The Kaplan-Meier estimator (KME) plots (Goel, Khanna, and Kishore 2010) seen in Figure 3 provide a higher resolution visual of the lifespans for each outcome metric. Figure 3 shows how survival rates change throughout time with vertical lines denoting when 95% of r/popular posts rebound (i.e., go back to baseline activity levels) for each outcome metric. We observe that $\mathcal{D}^{Rare}$ takes drastically longer to rebound compared to other datasets when looking at comment and newcomer rates. However, the 95% survival points for removals are comparably close together. This could hint that moderators stop removing comments within a thread even though comments and newcomers continue to appear.

Computing Intensity Using Peak Activity

Each r/popular post has an *intensity* metric for each outcome variable. We calculate intensity by dividing the subreddit's maximum hourly rate *during the thread's lifespan* (defined earlier in this section) with the subreddit's average hourly rate in the previous week as a baseline. For example, if the subreddit peaked at 100 comments per hour during an r/popular thread's comment lifespan and the home subreddit's previous week's average was 25 comments per hour, then the intensity for #comments will be 4. As a robustness check, we exclude the activity within the r/popular thread and redo the intensity calculation to strictly quantify spillover effects. See Table 4 for the medians and Kruskal-Wallis H-statistics.

Comparing Against High-Performing Control Posts: To estimate how much of the intensity is causally attributable to r/popular, we pair each r/popular post with a group of high-performing non-r/popular control posts from the same subreddit. The controls must meet the following criteria: (1) made in the previous week of the corresponding r/popular post's subreddit, (2) not made 24 hours before the r/popular post, (3) placed in the top 5th percentile threads for the number of comments, and (4) and did not reach top 100 of r/popular. For tractability, if more than 5 posts satisfy these conditions, we randomly select 5. Only 19 r/popular posts had no candidates and were dropped for this analysis.

Using the pairs we created, we build a Bayesian model (McElreath 2020) to estimate the standardized effect sizes (i.e., analogous to Cohen's d) between treatment and control group intensity distributions. Bayesian modeling allows us to customize our model because no out-of-the-box model estimates Cohen's d where each treatment observation is matched to a group of control observations. Koshy et al. (2023) also use Bayesian modeling in a social computing context and list benefits like the ability to specify priors based on outside expertise and partial pooling which helps the model share information across observations—we utilize this for our control groups. Kay, Nelson, and Hekler (2016) also provide an overview of the method's fit within HCI.

To help describe our standardized effect sizes model, we will start with the formula for Cohen's d:

$$d = \frac{\mu_t - \mu_c}{\sigma}$$

The above can be rearranged to better mimic the model we will describe.

$$\mu_c = \mu_t - d \cdot \sigma$$

Because we have matched data, with multiple control posts paired to each treatment post, we cannot apply this formula as written. Instead, we adapted the formula to fit our situation. Let $t_o[g]$ be the log value for outcome o for the gth r/popular post and $c_o[g, i]$ be the log value of the ith control post for outcome o that was matched with r/popular post g. Lastly, let d_o be the effect size for outcome o. We log scale these values because they are log-normal distributions. We will then substitute the terms we introduced into the formula above.

$$\mu_o[g] = t_o[g] - d_o \cdot \sigma_o[g]$$

	#comments	#removals	#newcomers
$\mathcal{D}^{Samp}$	1.85	2.59	2.22
$\mathcal{D}^{Freq}$	1.72	2.29	1.98
$\mathcal{D}^{Occa}$	2.74	5.41	4.10
$\mathcal{D}^{Rare}$	5.91	12.09	10.81
H-stat.	3,499.59	4,241.78	4,261.02

(a) With r/popular

	#comments	#removals	#newcomers
$\mathcal{D}^{Samp}$	1.47	2.05	1.58
$\mathcal{D}^{Freq}$	1.42	1.89	1.49
$\mathcal{D}^{Occa}$	1.62	3.21	1.84
$\mathcal{D}^{Rare}$	2.14	5.90	2.73
H-stat.	494.47	1,831.13	687.73

(b) Without r/popular

Table 4: Median intensities, defined in our community-level outcomes section, for each dataset and outcome variable. The last row contains H-statistics ($df = 3$, $p < 0.01$) for the Kruskal-Wallis tests performed.

Where $\mu_o[g]$, $\sigma_o[g]$, and d_o are unobserved, however, are linked to observations through the following model where control values are assumed to be drawn from a normal distribution. Here we used $\tau_o[g] = \frac{1}{\sigma_o[g]^2}$ for convenience.

$$c_o[g,i] \sim \mathcal{N}(\mu = \mu_o[g], \sigma = \frac{1}{\sqrt{\tau_o[g]}})$$
$$= \mathcal{N}(\mu = t_o[g] - d_o \cdot \sigma_o[g], \sigma = \frac{1}{\sqrt{\tau_o[g]}})$$

We used a Bayesian approach in estimating the unobserved parameters where we pooled the $\tau_o[g]$ across control groups using a gamma prior with weakly informative hyperpriors on α_o and β_o.

$$\tau_o[g] \sim \Gamma(\alpha_o, \beta_o)$$
$$\alpha_o, \beta_o \sim Exp(\lambda = 1)$$

The effect size for outcome o, d_o, is assigned to an informed prior tightly centered around zero, which corresponds to a strong assumption that the effect size is small if it exists at all.

$$d_o \sim \mathcal{N}(\mu = 0, \sigma = 0.5)$$

By running the model above, we learn the effect size between our treatment and control groups. Table 5a contains the percentage difference between treatment and control which we produced by removing the normalizing term, $\sigma_o[g]$, from $c_o[g,i]$'s μ term in the normal distribution.

Intensity Results (RQ2b)

Table 4 shows the median intensities for each outcome and dataset, capturing how much above baseline the subreddit's activity reaches after being placed on the r/popular feed. For example, in $\mathcal{D}^{Samp}$, we find that the peak hour after appearing on the r/popular feed was 85% above baseline—which we defined as the average activity in the previous week.

	#comments	#removals	#newcomers
$\mathcal{D}^{Samp}$	7.36%	7.68%	14.57%
$\mathcal{D}^{Freq}$	3.87%	4.50%	7.79%
$\mathcal{D}^{Occa}$	30.34%	32.05%	58.72%
$\mathcal{D}^{Rare}$	76.47%	71.94%	138.93%

(a) Mean effect on intensity from appearing on r/popular

	#comments	#removals	#newcomers
$\mathcal{D}^{Samp}$	0.193	0.134	0.235
$\mathcal{D}^{Freq}$	0.092	0.072	0.121
$\mathcal{D}^{Occa}$	0.580	0.347	0.734
$\mathcal{D}^{Rare}$	1.065	0.600	1.310

(b) Effect size (Cohen's d)

Table 5: Results from estimating the difference between r/popular posts intensity and non-r/popular posts intensity for our three outcome measures.

To compare between datasets, we ran Kruskal-Wallis tests for each column seen in Table 4 and all returned statistically significant results ($p < 0.01$) leading us to run post hoc Conover-Iman tests like we did for previous sections. For this section, all pairwise post hoc tests came back significant ($p < 0.01$), allowing us to freely compare.

One consistent trend is the rise in intensity as you move from communities in $\mathcal{D}^{Freq}$ to $\mathcal{D}^{Rare}$. The higher intensities measured for $\mathcal{D}^{Rare}$ could be due to lower baseline activity levels, but that still means that there are strong deviations from typical activity levels—even when we remove the r/popular thread from the subreddit activity seen in Table 4b. Additionally, removal and newcomer spikes consistently surpass the ones seen in comment activity which match our thread-level findings. Lastly, comparing between subtables, we can see that intensities drop substantially when we remove r/popular activity from the subreddit activity, however, there still is a measurable jump, especially for removals for $\mathcal{D}^{Occa}$ and $\mathcal{D}^{Rare}$.

Results From Comparing Against Control Posts: To determine how much of the intensities measured are attributable to r/popular, we rely on our effect size model using treatment-control pairs—summarized in Table 5. From Table 5a, we can see the estimated mean effect on intensities caused by the r/popular feed. Looking at $\mathcal{D}^{Samp}$ and $\mathcal{D}^{Freq}$, the r/popular feed's effect on intensity is minimal, however, existent. Where we see the r/popular feed's effects more clearly is with $\mathcal{D}^{Occa}$ and $\mathcal{D}^{Rare}$. The results in Table 5a match the estimated effect sizes (i.e., Cohen's d) in Table 5b. When looking across outcomes, the r/popular feed's effect on newcomers surpasses the other two indicating that the popular feed serves as an avenue for new users to contribute within a community.

Identifying Factors that Influence Observed Effects of Popularity

Next, we describe the features for the regression analysis on $\mathcal{D}^{Samp}$ to identify factors that influence the effects of

r/popular posts on threads and communities.

Feature Development

To understand what influences the thread and subreddit outcomes, we identify three categories of features: *author*, *thread*, and *subreddit*.

Author Features: We first identify potential factors about the Reddit user who posted the r/popular post. Post authors can significantly influence the post's activity through their topic selection, titling, and early participation in the thread. We use the following features to capture their background:

1. *Age:* time between the author creating their account to when they posted their r/popular post.
2. *Post/Comment Karma:* net karma (i.e., reputation score) from the author's most recent 1,000 posts/comments before the r/popular post.
3. *Is Post/Comment Newcomer:* whether the author has posted/commented on the subreddit between January 2020 to the r/popular post's creation time.

Thread Features: We consider characteristics of the r/popular thread because the content type, community response, and features describing a post's appearance on r/popular are all likely to impact outcomes. We separate metadata-related features that are set at creation time and do not change.

1. *# Awards:* number of awards the post received.
2. *# Crossposts:* number of times the post was posted to a different subreddit.
3. *Time To Popular:* time between the post's creation and its first appearance on r/popular.
4. *Time On Popular:* time between the post's first and last appearance on r/popular.
5. *Got Locked:* whether moderators locked the thread.
6. *Got Stickied:* whether moderators stickied the thread.

Next, the metadata features:

1. *Is Image/Text/Video/Link:* whether the post contains an image/text/video/link.
2. *Title Length:* title character count.
3. *Body Length:* post body character count.

Subreddit Features: Our dataset contains 1,320 distinct subreddits. To understand the differential effects of popularity on subreddits, we use features like community size, moderation team size, and baseline activity levels.

1. *# Moderators:* number of moderators—no historical data exists, queried on August 2023.
2. *Contains AutoMod:* whether the moderation team contains u/AutoModerator—a common moderation tool.
3. *# Previous Popular:* number of times this subreddit has reached r/popular in the previous week.
4. *# Subscribers:* number of subscribers to the subreddit.
5. *Comment Rate:* average hourly number of comments in this subreddit over the previous week.
6. *% Removed:* percentage of comments removed of all comments made in the previous week—not used while regressing on #removals because it is too correlated.

Multi-Collinearity Check: We calculated the variance inflation factor (VIF) for each independent variable to detect multi-collinearity. Using a VIF threshold of 5 (James et al. 2013), we pruned variables representing the proportion of each content type (e.g., image, video, etc.) posted on the subreddit. Previous iterations of the model pruned the age of the subreddit, whether the subreddit description contained rules, and the percentage of deleted comments as variables.

Regression Analysis for Thread Outcomes

To quantify how each feature impacts an r/popular thread, we created three multilevel Bayesian regression models, one for each outcome variable, that estimate the log number of outcomes (e.g., log number of comments) to get the percentage change to the outcome. The results are summarized in Table 6. We previously experimented with negative binomial and Poisson models, but Bayesian approaches[3] provided greater flexibility and handling of uncertainty.

To help define our model, let $y_o[i]$ be the ith r/popular post where o is the outcome we are measuring. Let $y_o[i]$ be drawn from the following distribution:

$$\ln(y_o[i]) \sim N(\mu = \vec{p}_o \cdot \vec{x}[i] + h_{o,c[i]}, \sigma)$$

Notice that we take the natural log of the outcome to see the multiplicative changes applied by each feature. For our μ term, let $p_{o,j}$ be the coefficient for the jth feature for outcome o which gets multiplied by the feature vector $\vec{x}[i]$, denoting the features for the ith r/popular post. Next, we have $h_{o,c[i]}$ which is the pooled intercept term for outcome o where $c[i]$ is the content type of the ith r/popular post.

With those terms defined, we specify the priors below:

$$\begin{aligned}
\sigma &\sim HalfCauchy(\beta = 10) \\
p_{o,j} &\sim \mathcal{N}(\mu = 0, \sigma = 20) \\
h_{o,c} &\sim \mathcal{N}(\mu_{pool}, \sigma_{pool}) \\
\mu_{pool} &\sim \mathcal{N}(\mu = 0, \sigma = 20) \\
\sigma_{pool} &\sim InverseGamma(\alpha = 1, \beta = 1)
\end{aligned}$$

Essentially, this Bayesian model runs a log-normal regression where the intercept term changes for each content type, but is pooled using a shared hyperprior to learn across content types. As with a typical regression, the coefficients from $\vec{p}_o$ will learn the relationship between our features and logged outcome measures.

Regression Analysis for Subreddit Outcomes

In the previous section, we defined two properties of an r/popular post to assess its impact on its respective subreddit: (1) the lifespan capturing the longevity of r/popular and (2) the intensity representing the disruptiveness of said event.

We analyze the former using a Cox proportional-hazard model (Lin and Wei 1989),[4] a survival regression that produces hazard ratios describing how each feature influences

[3] Using the PyMC package.

[4] https://lifelines.readthedocs.io/en/latest/

	Com.	Rem.	New.
AUTHOR FEATURES			
Age (1 year)			-1.3%
Is Comment Newcomer			14.1%
Is Post Newcomer			23.7%
Comment Karma (1k)			-0.1%
THREAD FEATURES			
# Awards	0.1%	0.3%	0.3%
# Crossposts		0.2%	0.2%
Got Locked	-36.9%	98.5%	-45.1%
Got Stickied	87.2%		52.8%
Time to Popular (1 hour)	5.1%	2.0%	4.7%
Time on Popular (1 hour)	10.8%	10.6%	13.2%
THREAD METADATA			
Title Length (10 char.)	1.4%		
Body Length (100 char.)			-0.5%
SUBREDDIT FEATURES			
Contains AutoMod	9.6%	23.4%	-27.9%
# Previous Popular	1.1%	2.5%	
% Removed	-0.4%	NA	-2.2%
Comment Rate (100 / hour)	3.7%	2.6%	1.3%
# Subscribers (100k)			0.1%
# Moderators		0.2%	
INTERCEPTS			
Is Link	284.8	9.5	40.3
Is Text	225.8	8.6	49.6
Is Video	269.8	13.4	28.6
Is Image	186.5	5.6	35.4

Table 6: Results from multilevel Bayesian models estimating the log number of outcomes within an r/popular thread. The abbreviated headers correspond to the number of comments, removed comments, and newcomers, respectively. Percentages indicate change based on one unit increase—if not specified, the unit is 1. The content types serve as unique intercepts for the model. Low-significance relationships were omitted for conciseness.

the baseline hazard ratio seen in Figure 3. Note that the time to r/popular and time on r/popular are removed as independent variables as they are direct components in the post's lifespan. The results are shown in Table 8.

For our intensity metric, we use a similar multilevel model defined above to estimate each feature's relation to each outcome intensity. See Table 7 for those results.

Results

Factors Affecting Thread-Level Outcomes (RQ1b): The percentages in Table 6 represent the corresponding change when the feature is increased by one unit. For Boolean features (e.g., is comment newcomer), the percentages correspond to when the feature is true.

We find that being on r/popular an hour longer corresponds to a 10.8% increase in comments, 10.6% increase in removed comments, and 13.2% increase in newcomers within the thread. Intuitively, being on such a prominent feed would lead to more traffic, and our model confirms it and measures the extent of that relationship. We see a similar, but less drastic, relationship with the time a post takes to reach r/popular—because there is more time to accrue activity.

	Com.	Rem.	New.
THREAD FEATURES			
# Awards	0.2%	0.2%	0.3%
Got Locked	11.9%	51.4%	
Got Stickied	211.4%	168.7%	273.8%
Time to Popular (1 hour)		1.1%	
Time on Popular (1 hour)	1.7%	2.1%	2.4%
SUBREDDIT FEATURES			
Contains AutoMod	-9.0%	-8.6%	-9.5%
# Previous Popular	-1.3%	-2.1%	-1.8%
% Removed		-1.9%	
Comment Rate (100 / hour)	-0.6%	0.9%	-0.7%
INTERCEPTS			
Is Link	2.58	3.90	3.47
Is Text	2.37	4.01	3.03
Is Video	2.76	4.49	3.42
Is Image	2.45	4.27	3.37

Table 7: Results from multilevel Bayesian models estimating the log outcome intensity. The abbreviated headers correspond to the number of comments, removed comments, and newcomers, respectively. Percentages indicate the amount of change based on one unit increase—if not specified, the unit is 1. Content types serve as unique intercepts for the model. Low-significance relationships were omitted.

Our model finds the largest relationships with moderator actions like locking and stickying. If a moderator locks a post, we find substantial decreases in the number of comments (-36.9%) and newcomers (-45.1%) and a strikingly large increase in removals (98.5%). Moderators may lock the thread when they sense a significant amount of rule-violating behavior to stop further from happening. However, determining the directionality of stickying is more difficult because it could be that moderators sticky highly active posts or stickying resulted in more attention. Regardless, stickying does not have any significant relationship with removals within a thread.

To identify differences between content types, which are the intercepts within our model, we will be using credible intervals (CI). They are similar to confidence intervals in that they describe a range of possible values but from a posterior distribution (McElreath 2020). Regardless, we see fewer comments on image threads (94% CI [172, 201]) compared to link (94% CI [257, 314]), video (94% CI [242, 298]), and text threads (94% CI [206, 246]). Video posts, however, see more removals (94% CI [11, 16]) than image (94% CI [5, 6]), link (94% CI [8, 11]), and text posts (94% CI [7, 10]) which could indicate that videos attract more norm-violating behavior or are more difficult to moderate, but the differences are fairly small. Lastly, text posts attracted more newcomers (94% CI [44, 55]) compared to image (94% CI [32, 39]) and video posts (94% CI [25, 32]), but only slightly above link posts (94% CI [35, 46]). This could hint at the accessibility of text posts attracting newcomers.

Surprisingly, we find that the author being a comment

newcomer and post newcomer correspond to a 14.1% and 23.7% increase in newcomers within a popular thread, respectively. This relationship may indicate that authors who are newcomers post more accessible content for other newcomers. However, given that there may be unknown confounders producing this relationship, a more detailed investigation would be needed to prove that theory.

The last trend pertains to moderation. We find that AutoMod presence has a substantial inverse relationship with removals and newcomers which could suggest a balancing act between enforcing norms and newcomer participation. Similarly, a higher percentage of removals within a subreddit showed decreases in the number of newcomers (-2.2%). Moderation team size also positively correlated with the number of removals within a thread (0.2%).

Factors Affecting Longevity & Intensity (RQ2c): Table 8 shows the few statistically significant relationships identified by our Cox proportional-hazards model. The sparseness in Table 8 matches the low concordance indices that measure model performance. Despite the poor performance, the models did highlight how repeated appearances on r/popular correlate with a faster return to baseline activity levels as indicated by the positive percentages. Additionally, we find that removal rates on link and video posts fall slower, -13.65% and -12.01% respectively, compared to image posts which the model uses as a reference category. All other relationships identified played fairly minimal roles in the hazard ratio. Our model's poor performance may be attributable to our use of "static" features—i.e., features that do not change since the post's creation time. Static features have limited predictive power, thus making it hard to forecast activity levels when a thread "dies"—typically hours after post creation.

Our intensity regressions in Table 7, however, find more significant relationships than our longevity analysis. The most notable are the strong positive relationships for locking and sticking posts. For example, locked posts corresponded with an 11.9% higher peak in comments, surpassed by the 51.4% increase in removal intensity. Similarly, stickied posts correlated with peaks more than twice as high for all outcomes. However, because we flagged posts if they were stickied or locked at any time during their tenure on r/popular, it is difficult to tell the directionality of this effect.

Other than stickying and locking, we find that AutoMod's presence corresponded with consistent reductions in spikes for all three outcomes. This finding can have practical significance for moderators as they decide whether to use AutoMod to assist in their tasks. Regarding repeated appearances on r/popular, our models find that it reduces lifespan and intensity by about 2%. When looking at the intercepts, the credible intervals are pretty intertwined between different content types, so the model cannot find any substantial differences between them. One final point, however, is how a percentage increase in removal rate (% removed) corresponded with a disproportionate 1.9% decrease in peak removals which could indicate that stricter subreddits are more resilient towards removal spikes.

	Com.	Rem.	New.
AUTHOR FEATURES			
Comment Karma (1k)	-0.12%		
THREAD FEATURES			
# Awards	-0.28%	-0.28%	-0.31%
THREAD METADATA			
Is Link		-13.65%	
Is Video		-12.01%	
Title Length (10 char.)	-0.86%	1.11%	-1.12%
SUBREDDIT FEATURES			
# Previous Popular	2.02%	1.67%	2.16%
Concordance Index	0.600	0.5944	0.6183

Table 8: Results from our Cox proportional-hazards models on r/popular post lifespans. The abbreviated headers correspond to the number of comments, removed comments, and newcomers, respectively. Positive percentages indicate a faster decay rate to baseline activity levels given a unit increase. Image posts are our reference category, i.e., link post removal rates are -13.65% less hazardous compared to image posts. Only significant relationships are shown ($\alpha = 0.05$ Bonferroni corrected). A concordance of 0.500 is expected from random predictions, so our models do not explain much of the hazard rate.

Discussion

We demonstrated how appearing on Reddit's r/popular feed impacts threads and subreddits, capturing and modeling dynamics on r/popular. We empirically showed that communities that rarely hit r/popular ($\mathcal{D}^{Rare}$) experience more pronounced spillover effects than those that appear more often on r/popular. Next, we discuss our findings in light of our research questions and how communities may benefit from this knowledge.

Disruptive & Differential Effects of Algorithmic Curation

When answering our RQs, we observed significant spikes in thread and community activity due to appearing on the algorithmically-curated feed r/popular. The relationships between our independent variables and thread- and subreddit-level outcomes indicate that previous appearances on r/popular, AutoMod presence, and locking threads correspond to substantial changes in the metrics we measured.

Our results also confirm qualitative findings that an appearance in trending feeds causes events beyond just the post itself (DeVito 2022). We empirically demonstrate the presence of spillover effects. Additionally, in the same community, newcomer participation on threads not on r/popular increases as do comment and removal levels, suggesting the potential for disruptions across the community.

Finally, our results show differential effects on communities when they hit the r/popular feed. For RQ1, we demonstrate that $\mathcal{D}^{Rare}$, subreddits that rarely reach the r/popular feed, experience more intense relative effects because of popularity, especially newcomer participation. Our survival analyses for RQ2 show that these effects lasted longer

in subreddits less accustomed to exposure. In most cases, the increased volume of comments came with significantly higher comment removal rates by moderators. Taken together, this shows that unexpected popularity substantially increases moderator workload for smaller communities.

Supporting Community Resilience

Understanding how communities respond to disruptions builds on prior work about "resilience" in the face of disruptive events. Resilience is the ability to effectively respond to disruptions and has been theorized in online community research (Butler et al. 2014; Chan et al. 2022). For example, Butler et al. (2014) define community resilience as the ability of a community to maintain users despite topic changes. Garcia, Mavrodiev, and Schweitzer (2013) define resilience to mean a user can leave without triggering other users to leave the site. The prior work shares the same intuition that resilient communities can withstand change. In our case, the change (reaching the r/popular feed) is a sudden and drastic one that mimics the literature in offline communities and disaster preparedness (Cohen et al. 2013) where resilience is the ability to operate and withstand crises.

How can communities be more resilient to disruptions due to sudden popularity via appearances on r/popular? Moderators of subreddits operate like community leaders (Seering et al. 2022; Seering, Kaufman, and Chancellor 2022), so our recommendations focus on these key users. Important to moderator success is platform support to help communities be more resilient. First, we imagine temporary support for communities through short-term, increased moderator presence. This was shown to be effective in Discord at limiting hate and harassment during Pride Month (Seering et al. 2022). More nuanced automated moderation can also provide moderators more bandwidth for responding to disruptive events like the r/popular feed and limit the negative effects of u/AutoModerator, also proposed in prior work (Chandrasekharan et al. 2019). Other strategies include more distributed and proactive (Zhang et al. 2018) approaches to facilitate prosocial (Bao et al. 2021) and resilient (Lambert, Rajagopal, and Chandrasekharan 2022) conversations within online communities.

Implications for Online Moderation

Our results allow us to formulate pragmatic recommendations that moderation teams and developers can consider when dealing with disruptions:

Post Content Type Matters: We find that different types of content behave differently when they reach virality through the r/popular feed. For example, video posts contained more removed comments than other types and affected the subreddit's removal rates for longer compared to image posts. These findings may change for individual subreddits, so we encourage moderators to pay attention to how different posts behave and monitor accordingly.

Deploying Moderator Actions More Effectively: We also find strong correlations between the amount of exposure a thread receives and the activity contained within the thread. We recommend that moderators closely monitor posts if they have been stickied or highlighted on the r/popular feed for extended periods. These forms of prolonged exposure may encourage elevated commenting rates and substantial newcomer participation.

Employing Automated Moderation: Our results uncovered the potential impact that moderation can have on thread participation in times of increased exposure. We find that employing u/AutoModerator resulted in more removed comments and fewer newcomers within the thread. Additionally, AutoMod presence reduces the peak activity levels for all three outcomes. Because r/popular affects communities that infrequently appear on the feed more, we advise them to invest time in their moderation efforts and consider employing automated moderation tools.

Limitations & Future Work

Examining Causal Factors: There are limitations with correlational studies on public, observational data. We cannot infer causal relationships between our independent variables and the participatory outcomes we study. However, it is difficult to artificially induce and control sudden spikes in popularity, making controlled experiments infeasible (and potentially unethical). Future work should explore quasi-causal methods to establish the causal link between our features and different outcomes of popularity.

Including Information about Removed Comments: In our analysis, we could identify occurrences of removed comments; however, we were unable to identify *who* got removed and *what* got removed. Obtaining information about removed content is a significant technical challenge as this data is not publicly available through APIs. This could help determine whether the majority of removals in r/popular posts and the originating communities are authored by newcomers, and future work could identify the specific types of violations resulting from sudden popularity.

Auditing Other Black-Box Curation Algorithms: Our goal was to study the impacts of sudden popularity caused by Reddit's black-box curation algorithm and understand the disruptive effects on communities. Future work should empirically explore the impact of similar black-box algorithms (e.g., TikTok) and audit the effects of algorithmic "trending" and "hot" feeds on communities and their moderators.

Broader Perspectives & Ethics

Our data is taken from r/popular through public APIs, and thus direct consent was not obtained from users. The Institutional Review Boards at our institutions do not require us to obtain consent from users whose posts/comments appear in public datasets, which is common in computational Reddit research and ICWSM (Proferes et al. 2021). However, there are still risks in gathering and processing this data. We heed the advice from social media ethicists (Proferes et al. 2021) and model papers in ICWSM and complementary venues (Chandrasekharan et al. 2017; Li, Hecht, and Chancellor 2022) and adopt their recommendations.

We handle moderator-removed and user-deleted Reddit comments; however, our analysis does not use text or user IDs of these removals/deletions (Chancellor, Lin, and De Choudhury 2016). We store data on secured, firewalled servers at the primary author's institution. We anonymize

data as appropriate for the analysis—we use usernames to gather information about poster histories but do not analyze user data for any other purpose. Finally, we do not provide thread- or comment-level details to prevent re-identification.

Our work has one notable consequence: by helping to "decode" or audit the black box of the r/popular feed, this provides more information to the public about how it works and what influences it. We discussed how more transparency can help users and communities prepare for and reason about how the r/popular feed impacts them. On the other hand, increased transparency about trending and popular feeds may lead to system gaming by actors with questionable intentions, like spammers and bad-faith influencers. There also could be more nefarious use of our findings through targeted attacks on "rare" communities. We believe that the benefits of more empirical insight in this space outweigh the downsides—and that more information can help communities better respond to all unwanted negative attention, no matter their origin.

Conclusion

Reasoning about the "why" of algorithmic black boxes like the r/popular feed on Reddit is tricky, in part because these systems are not transparent. Although there are reasonable concerns about intellectual property, the lack of transparency of such algorithmic feeds prevents people from reasoning about or responding to how content appears on them. Many current practices are "defensive"—they assume the affected party has no control over their presence on algorithmically curated feeds. We suggest several "offensive" post/thread-level options that can help the community better cope with the increase in attention. First, we suggest an option to delist the post from trending or hot feeds, like the r/popular feed, once it happens. This could be done at the user level (for platforms like X or TikTok) or the community level (for Reddit). Second, we imagine that a community could choose to temporarily limit newcomers' ability to post on the rest of the community for a brief period (similar to *locking* a post), limiting the disruptiveness they cause to the remainder of the community and managing moderator workload.

References

Bao, J.; Wu, J.; Zhang, Y.; Chandrasekharan, E.; and Jurgens, D. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset.

Berger, J.; and Milkman, K. L. 2012. What makes online content viral? *Journal of marketing research*.

Butler, B.; Bateman, P.; Gray, P.; and Diamant, I. 2014. An Attraction-Selection-Attrition Theory of Online Community Size and Resilience. *MIS Quarterly*, 38.

Chan, J.; Atreyasa, A.; Chancellor, S.; and Chandrasekharan, E. 2022. Community Resilience: Quantifying the Disruptive Effects of Sudden Spikes in Activity within Online Communities. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*.

Chancellor, S.; Lin, Z.; and De Choudhury, M. 2016. " This Post Will Just Get Taken Down" Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1157–1162.

Chandrasekharan, E.; Gandhi, C.; Mustelier, M. W.; and Gilbert, E. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW).

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).

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *Proceedings of the International Conference on World Wide Web*.

Ciampaglia, G. L.; Nematzadeh, A.; Menczer, F.; and Flammini, A. 2018. How algorithmic popularity bias hinders or promotes quality. *Scientific reports*, 8(1).

Cohen, O.; Leykin, D.; Lahad, M.; Goldberg, A.; and Aharonson-Daniel, L. 2013. The conjoint community resiliency assessment measure as a baseline for profiling and predicting community resilience for emergencies. *Technological Forecasting and Social Change*, 80(9).

Cunha, T.; Jurgens, D.; Tan, C.; and Romero, D. 2019. Are All Successful Communities Alike? Characterizing and Predicting the Success of Online Communities. In *The World Wide Web Conference*, WWW '19. New York, NY, USA: Association for Computing Machinery.

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 WWW*.

DeVito, M. A. 2022. How transfeminine TikTok creators navigate the algorithmic trap of visibility via folk theorization. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2).

Eslami, M.; Rickman, A.; Vaccaro, K.; Aleyasen, A.; Vuong, A.; Karahalios, K.; Hamilton, K.; and Sandvig, C. 2015. "I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 2015 CHI conference on human factors in computing systems*.

Figueiredo, F.; Benevenuto, F.; and Almeida, J. M. 2011. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining*.

Garcia, D.; Mavrodiev, P.; and Schweitzer, F. 2013. Social resilience in online communities: the autopsy of friendster.

In *Proceedings of the first ACM conference on Online social networks*, COSN '13. New York, NY, USA.

Garg, R.; Smith, M. D.; and Telang, R. 2011. Measuring information diffusion in an online community. *Journal of management information systems*, 28(2).

Goel, M. K.; Khanna, P.; and Kishore, J. 2010. Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research*, 1(4): 274–278.

Gurjar, O.; Bansal, T.; Jangra, H.; Lamba, H.; and Kumaraguru, P. 2022. Effect of Popularity Shocks on User Behaviour. *Proceedings of the International AAAI Conference on Web and Social Media*, 16.

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2).

James, G.; Witten, D.; Hastie, T.; and Tibshirani, R. 2013. *An Introduction to Statistical Learning*. Springer.

Kairam, S. R.; Wang, D. J.; and Leskovec, J. 2012. The life and death of online groups: predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12. New York, NY, USA: Association for Computing Machinery.

Kay, M.; Nelson, G. L.; and Hekler, E. B. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 4521–4532. New York, NY, USA: Association for Computing Machinery.

Keegan, B.; Gergle, D.; and Contractor, N. 2013. Hot off the wiki: Structures and dynamics of Wikipedia's coverage of breaking news events. *American behavioral scientist*.

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 CHI Conference on Human Factors in Computing Systems*.

Kollock, P.; and Smith, M. A. 1999. *Communities in cyberspace*. Routledge London.

Koshy, V.; Bajpai, T.; Chandrasekharan, E.; Sundaram, H.; and Karahalios, K. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 286:1–286:36.

Kraut, R. E.; and Resnick, P. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press.

Kumar, S.; Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

Lambert, C.; Rajagopal, A.; and Chandrasekharan, E. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16.

Li, H.; Hecht, B.; and Chancellor, S. 2022. All That's Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Lin, D. Y.; and Wei, L. J. 1989. The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*, 84(408): 1074–1078.

Lin, Z.; Salehi, N.; Yao, B.; Chen, Y.; and Bernstein, M. 2017. Better When It Was Smaller? Community Content and Behavior After Massive Growth. *Proceedings of the International AAAI Conference on Web and Social Media*.

Maldeniya, D.; Budak, C.; Robert Jr., L. P.; and Romero, D. M. 2020. Herding a Deluge of Good Samaritans: How GitHub Projects Respond to Increased Attention. In *Proceedings of The Web Conference 2020*, WWW '20. New York, NY, USA: Association for Computing Machinery.

Matias, J. N. 2016. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16. New York, NY, USA: Association for Computing Machinery.

McElreath, R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. New York: Chapman and Hall/CRC, 2 edition.

Perez, S. 2021. Reddit adds a new way to post with launch of 'Predictions' feature.

Proferes, N.; Jones, N.; Gilbert, S.; Fiesler, C.; and Zimmer, M. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2): 20563051211019004.

Seering, J.; Dym, B.; Kaufman, G.; and Bernstein, M. 2022. Pride and Professionalization in Volunteer Moderation: Lessons for Effective Platform-User Collaboration. *Journal of Online Trust and Safety*, 1(2).

Seering, J.; Kaufman, G.; and Chancellor, S. 2022. Metaphors in moderation. *New Media & Society*, 24(3).

Weng, L.; Menczer, F.; and Ahn, Y.-Y. 2013. Virality prediction and community structure in social networks. *Scientific reports*, 3(1).

Yang, D.; Kraut, R.; and Levine, J. M. 2017. Commitment of newcomers and old-timers to online health support communities. In *Proceedings of the 2017 CHI conference on human factors in computing systems*.

Zhang, A. F.; Wang, R.; Blohm, E.; Budak, C.; Jr, L. P. R.; and Romero, D. M. 2019. Participation of New Editors after Times of Shock on Wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. Association for Computational Linguistics.

Zhang, J. S.; Keegan, B.; Lv, Q.; and Tan, C. 2021. Understanding the Diverging User Trajectories in Highly-related Online Communities during the COVID-19 Pandemic. *Proceedings of the International AAAI Conference on Web and Social Media*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see our Broader Perspectives & Ethics section for a discussion on privacy and anonymity in our data.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see our Discussion section for contributions on which the abstract and introduction were based.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the sections where we describe outcome measures and models.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No, because there are no artifacts in the data relevant to our analysis.
   (e) Did you describe the limitations of your work? Yes, see our Limitations & Future Work section.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see our Broader Perspectives & Ethics section.
   (g) Did you discuss any potential misuse of your work? Yes, see our Broader Perspectives & Ethics section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see our Broader Perspectives & Ethics section.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? See our Broader Perspectives & Ethics section.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, we footnoted the packages we used.
   (b) Did you mention the license of the assets? No, the footnoted links have license information.
   (c) Did you include any new assets in the supplemental material or as a URL? No, because the data is not anonymized.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see our Broader Perspectives & Ethics section.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see our Broader Perspectives & Ethics section.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? No, because we are not making the data public.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, because we are not making the data public.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA