

Combating the Harassment of Online Activists in Authoritarian Settings *

Romain Ferrali † Horacio Larreguy ‡

December 17, 2024

Abstract

Online harassment is widespread on social media, with severe consequences in authoritarian regimes where opposition figures are targeted. While counterspeech can mitigate harassment in democracies, its impact in authoritarian contexts is less understood. An online experiment conducted by an NGO in an authoritarian regime shows that public counterspeech reduced victim activity and follower engagement, likely due to fear of retaliation, while private counterspeech increased both. Harassers were unaffected by counterspeech. These findings highlight the limitations of counterspeech in authoritarian settings and emphasize the need for context-sensitive approaches, particularly addressing the fear of repression that shapes victim behavior in these environments.

Keywords: Harassment, Social Media, Content Moderation, Authoritarian Context

JEL Codes: D74, L86, C93.

1 Introduction

In authoritarian settings, political activists are often silenced on social media through harassment. Such online harassment frequently involves hate speech and may be orchestrated by government agents (Guriev and Treisman, 2019). While there is increasing evidence of social media’s contribution to hate speech and its harmful impact both offline and online on targeted communities (Müller and Schwarz, 2020; Bursztyn et al., 2024; Beknazar-Yuzbashev et al., 2023; Müller and Schwarz, 2023), our understanding of the policies that effectively curb hate speech on social media, particularly in authoritarian contexts, remains limited.

Online harassment through hate speech may be more severe in authoritarian than in democratic settings. In democratic settings, hate speech often centers around xenophobic attacks. In authoritarian settings, these issues are compounded by politically-motivated, often personally-targeted attacks. Prominent opposition figures are routinely targeted by posts that question their values and loyalty to the country. Less popular accounts that share similar views may also be targeted on similar grounds, either when they post content that diverges from the regime’s official stance on key political issues, or through what appears to be coordinated campaigns following major political events.

*This study was pre-registered in the Open Science Foundation (OSF) Registry (<https://doi.org/10.17605/OSF.IO/P36EZ>) and obtained IRB approval from Instituto Tecnológico Autónomo de México. Larreguy gratefully acknowledges funding from the French Agence Nationale de la Recherche under the Investissement d’Avenir program ANR-17-EURE-0010. Ferrali acknowledges that the project leading to this publication has received funding from the French government under the “France 2030” investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University - A*MIDEX. We thank Daniela Pinto-Veizaga for her excellent research assistance.

†Aix-Marseille University, CNRS, AMSE, Marseille, France, romain.ferrali@univ-amu.fr.

‡Instituto Tecnológico Autónomo de México, horacio.larreguy@itam.mx.

Moreover, third-party interactions with victims and harassers might be more meaningful in authoritarian settings than in democratic settings, as they may reveal the third-party stance towards the regime. Engaging with victim content by liking or sharing it may be viewed as supporting the opposition, potentially making users targets of harassment themselves. Conversely, interacting with harasser content may be viewed as endorsing the regime, inviting criticism from opposition supporters.

While previous work in democracies has shown that hate-speech content moderation may curb hate speech (Jiménez Durán, 2023; Beknazar-Yuzbashev et al., 2023; Jiménez Durán et al., 2023), platforms such as X (formerly Twitter) have limited the extent to which they engage in content moderation. Meanwhile, counterspeech that confronts hate speech has gained strong policy appeal. Policies that counter hate speech, often implemented by international and nongovernmental organizations (I/NGOs), are visible and may shape the incentives of harassers, while providing victims with a sense of protection. Counterspeech that promotes empathy (Hangartner et al., 2021), highlights likely sanctioning by close ties or other in-group users who can observe the user harassing on social media (Munger, 2017), or primes common identity (Siegel and Badaan, 2020) has successfully reduced retrospective or prospective hate speech.

However, the strategies that have worked in democratic contexts are unlikely to work in authoritarian contexts. Precisely because harassment of opposition users is justified based on dubious values and treason to the country, it is hard to induce empathy, highlight the social cost of harassment, or prime a common identity. Organizations that counter hate speech in these contexts are only left with the threat of reporting such behavior to platforms (Yildirim et al., 2023; Jiménez Durán, 2023), while signaling support to the victims.

We analyze a randomized controlled experiment conducted on Twitter by NGO Y in authoritarian regime Z.¹ The NGO first identified a network of harassers and their victims on the platform within regime Z. This network largely resembles a collection of stars, where a few highly active harassers target many victims, and many harassers target a few highly popular victims. Harasser-victim pairs were randomly assigned to one of three conditions: public messaging, private messaging, or control. This randomization was stratified based on the size of the star structure and whether the victim allowed direct messages (DMs). The NGO reported the harassment posts of treated harassers to Twitter and posted either public or private messages. These messages highlighted the reporting, condemned the harasser’s behavior, and expressed support for the victims, with private messages sent as DMs to the victims.

We deal with spillovers, which are natural in network settings like the one we study, in two ways. First, we assisted the NGO in decomposing the network into a well-defined collection of stars, where a few harassers target many victims, and many harassers target a few victims. Specifically, we identified central nodes connected to many peripheral nodes and had few connections to adjacent stars. The resulting decomposition features stars of various sizes, ranging from isolated harasser-victim pairs to large stars with dozens of peripheral nodes, with few links across stars. Focusing on (relatively) isolated stars alleviates the problem of spillovers across stars, as discussed below. Second, we test for the relevance of within-star spillovers and show robustness to accounting for across-star spillovers. The former exploits that, by design, isolated pairs cannot exhibit spillovers, and thus, we can compare the behavior of treated peripheral nodes in isolated pairs to that of treated peripheral nodes in stars. All the analyses use Ordinary Least Squares (OLS) regressions, supplemented by randomization-inference p-values to account for variations in treatment assignment stemming from our randomization strategy and the network structure.

We begin by evaluating whether the intervention reduced the frequency of posts from harassers to victims and the extent of their use of toxic language, as measured through a third-party natural language processing (NLP) tool. Our findings indicate insignificant, small effects

¹We do not disclose the location and the identity of the NGO to avoid reprisals against the NGO.

on either outcome. Additionally, we examine whether harassers remained active users post-intervention, hypothesizing a decline in activity among treated harassers. However, contrary to our expectations – and consistent with Twitter’s puzzling response that the reported posts did not constitute harassment – all harassers remained active throughout the weeks following the intervention.

Next, we examine the intervention’s impact on several outcomes related to the posting activity (posts, replies to others’ posts) and follower engagement with the posts (replies, reposts, quoted reposts, and likes) of both harassers and victims. First, we focus on the harassers whose victims were assigned to the public messaging condition, allowing them to observe the treatment, and compare their activity to that of harassers whose victims were in the control condition. Second, we compare victims assigned to the public and private messaging conditions to those in the control condition and between each other. Throughout, we analyze results by pooling the sample across all weeks post-intervention while also segmenting the data into three time periods: the first (early), second (intermediate), and final (late) thirds of the sample. This approach allows us to explore the treatment effects’ dynamics over time and gain deeper insights into the mechanisms driving these effects.

Overall, the pooled-sample results indicate no treatment effects on the activity and engagement with the content of the harassers, but suggestive distinct treatment effects on victims depending on whether they were privately or publicly treated. Privately treated victims exhibit a suggestive increase in their activity (.09 standard deviations, $p = .11$), along with a 1.5% increase in the number of accounts following them ($p = .08$) and a corresponding rise in engagement from these followers with their content (.05 standard deviations, $p = .16$). Conversely, publicly treated victims experience a suggestive decrease in their activity (-.06 standard deviations, $p = .12$) and in follower engagement (-.06 standard deviations, $p = .02$). Importantly, the differences in effects on indexes across the private and public treatments are always statistically significant. All the results are also robust to accounting for cross-star spillovers, and we find no significant evidence of within-star spillovers.

These results collectively show that harassers were completely unaffected by the intervention. In turn, privately treated victims demonstrate higher levels of activity and associated engagement, while publicly treated victims exhibit relatively lower levels of activity and engagement. Distinguishing between early, intermediate, and late treatment effects, indicates that publicly-treated victims only reduce their activity substantively and significantly early after treatment (-.07 standard deviation, $p = .09$). Follower engagement exhibits, however, a persistent, significant drop. This persistent decline suggests that follower disengagement is driven by fear of harasser retaliation rather than the organic tendency of followers to interact less with victims who post less content.

These findings contrast with the outcomes of similar public interventions aimed at reducing harassment on social media in democratic contexts, where we observe a reduction in harassment and victims typically show increased activity. They thus highlight the importance of accounting for victim fear and follower fear-driven disengagement in such interventions in authoritarian contexts, thus offering significant insights for I/NGOs working in that context.

We contribute to several strands of literature. First, we naturally contribute to research evaluating the effectiveness of counterspeech used by I/NGOs to confront and reduce online hate speech. Previous work in democratic contexts evaluates the effectiveness of counterspeech that primes empathy or highlight the social cost of harassment in reducing retrospective or prospective harassment (Hangartner et al., 2021; Munger, 2017; Siegel and Badaan, 2020). In turn, we focus on an authoritarian context where such strategies are unlikely to work because hate speech usually questions victims’ values or loyalty to the country. Moreover, in addition to harasser behavior, we also focus on subsequent victim activity and associated follower engagement. We finally show that the effects of counterspeech may vary significantly depending on whether the intervention is private or public. In doing so, we also speak to the

research on the effects of media censorship by authoritarian governments (Hobbs and Roberts, 2018; Chen and Yang, 2019).

Second, we contribute to recent experimental work on content moderation in democratic contexts. The paper closest to ours is Jiménez Durán (2023), which shows that randomly reporting Twitter posts for violating rules against hateful speech in a democratic context increases the likelihood that Twitter removes them. While reporting does not affect harassers' activity, it does lead to greater activity among victims. In contrast, our study evaluates an intervention in an authoritarian context, showing that the effects on victim activity and follower engagement depend on whether victims are informed privately or publicly. In these settings, fear of retaliation plays a crucial role in shaping the behavior of both victims and their followers.

Third, we contribute to the experimental literature on social networks, where spillovers from treated to control units are a common threat to identification (Aridor et al., 2024). Some solutions include minimizing the overlap between participants (Jiménez Durán, 2023), or randomizing at the cluster level, with a large enough number of clusters (Donati et al., 2024; Enríquez et al., 2024; Larsen et al., 2023). These solutions may be impractical in highly clustered settings with both within- and across-network spillovers. We introduce a framework that evaluates the relevance of within-network spillovers and demonstrates robustness in accounting for across-network spillovers, offering a practical approach for such complex contexts.

The paper is structured as follows. Section 2 provides additional details about the authoritarian context of the intervention. In Section 3, we describe our experimental design, including the network decomposition algorithm used to minimize cross-network spillovers and optimize identification, detail the treatment conditions and the randomization procedure, and describe our estimation procedure and robustness checks. In Section 4, we present our hypotheses, the results, and robustness checks. Section 5 concludes.

2 Context

The setting we consider is a competitive authoritarian regime in a mid-income country. Regime Z does not ban any social media platform nor exerts any form of Internet censorship. The regime does, however, monitor social media and several individuals have been convicted of jail sentences of up to four years for posting anti-regime content. Furthermore, opposition figures and posts of anti-regime content are frequently targeted by posts that question their values and loyalty to the country.

Harassment attacks are typically carried out by accounts with questionable authenticity. These accounts often have few followers (a median of 10), provide minimal information (e.g., pseudonyms, no location, or account description), and are difficult to link to real individuals (see Table SA-2 for details). This pattern suggests that many of these accounts may be inauthentic and potentially orchestrated by government agents. However, a smaller subset of accounts attracts a significantly larger following (the top 10% have between 200 and 20,000 followers) and, despite remaining anonymous, appears to be operated by individual users.

In contrast, victims tend to have more authentic accounts, often displaying more detailed information and being more easily traceable to real-life individuals. They also command a larger audience, with the median victim having around 1,000 followers. Some victims are highly prominent, with the top 10% having between 200,000 and 2.5 million followers. These highly followed accounts are often prominent opposition figures who regularly face harassment. Less popular accounts generally fall into two categories: those frequently engaging in political discussions and attracting consistent harassment, and less politicized accounts that become targets after expressing opposition views.

3 Design and methods

3.1 The harassment network

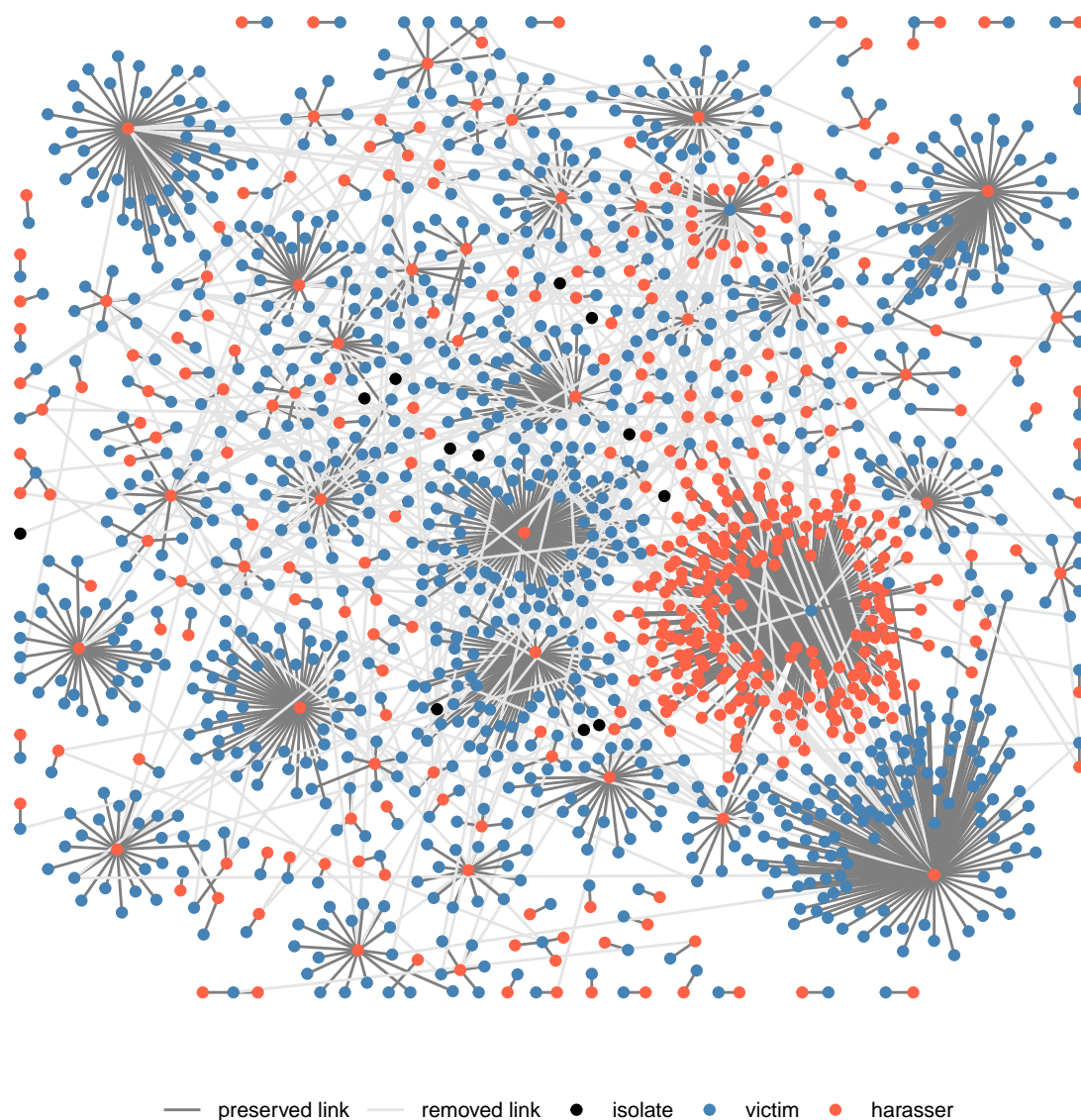
The NGO identified pairs of Twitter accounts where one user (the *harasser*) sent at least one harassment post in reply to another user (the *victim*). The collection of these pairs forms the harassment network. The NGO started with a small number of well-known victims of harassment (12) among the activists of country Z and a small number of well-known harassers (4). The NGO then initiated a snowballing procedure to identify additional victims and harassers. In other words, the NGO identified additional victims of those initial harassers and harassers of those initial victims, and so on.

To identify harassers and victims, the NGO used a combination of automated and manual procedures. To identify the victims of a given harasser, first, the NGO obtained a list of all reply posts (i.e., posts in response to a post by another user) sent by a given harasser, and their toxicity score through the Perspective API,² which uses machine learning models to estimate the probability that text uses toxic language. Second, among those posts whose toxicity score was greater than .9 (indicating a 90% probability of toxicity), they manually identified those that explicitly violated the platform’s regulations regarding hateful conduct and abusive behavior. The NGO used a custom-built interface to manually classify posts (Supplementary Appendix Figure SA-1). The interface displayed the post to be classified, as well as the platform’s rules surrounding hateful conduct and abusive behavior, with literal excerpts from the platform’s regulation. Lastly, the user to whom the post responded was labeled a victim. Likewise, to identify the harassers of a given victim, the NGO obtained a list of all reply posts to that user’s content and followed a similar procedure.

Figure 1 shows that the resulting network largely resembles a collection of harasser- and victim-stars: several highly active harassers may have many victims (harasser-stars), while few highly popular victims tend to attract many harassers (victim-stars).

²<https://perspectiveapi.com>

Figure 1: Harassment network.

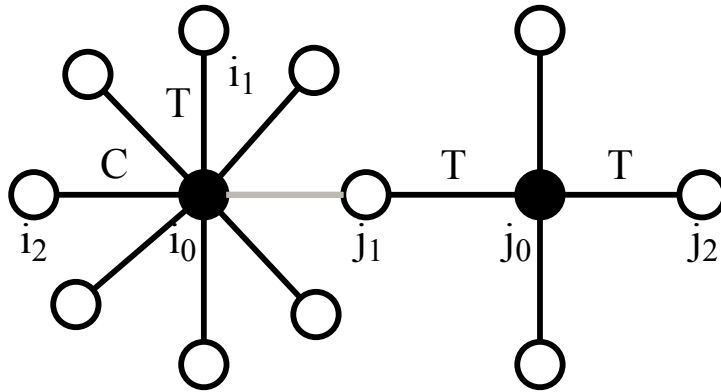


Notes. Nodes are colored according to their type after using the network decomposition algorithm.

3.2 Network decomposition algorithm

Ideal setting and rationale. By design, the NGO's intervention – namely, reporting the harasser's message to the victim to Twitter and posting a public or private message of support – jointly affects two actors – the harasser and the victim. Assessing the extent to which the intervention affects the activity of the victim and the harasser, as well as that of their followers, would ideally consider isolated pairs of harasser/victim, and compare treated pairs to untreated pairs. The setting is ideal in the sense that it eliminates spillovers. Since pairs are isolated, there cannot be interference across pairs.

Figure 2: Graph decomposition, spillovers, and treatment assignment.



Notes. This figure represents hypothetical stars, with victims in white and harassers in black. The decomposition algorithm removes the i_0 - j_1 link. Links i_0 - i_1 , j_0 - i_1 and j_0 - i_2 are assigned to treatment (T). Link i_0 - i_2 is assigned to control (C). Through node i_0 , treatment may spill over from the i_0 - i_1 link to i_2 (*within-star spillovers*) and to j_1 (*cross-star spillovers*).

Yet, Figure 1 shows that the harassment network resembles a collection of stars rather than isolated pairs. Stars still prove advantageous for analysis, as they afford some control over spillovers. Indeed, all spillovers must go through the central node. Consider Figure 2 and ignore the i_0 - j_1 link. In the left-hand side star, link i_0 - i_1 is treated while link i_0 - i_2 is not. Treatment may affect i_0 's behavior, which, in turn, may affect i_2 's behavior. However, all spillovers must go through i_0 . We could generalize the logic of our ideal setting to stars and compare treated stars to untreated stars. This approach would, however, be very costly in terms of power as there are many nodes but few stars. Instead, we do not examine the behavior of central nodes (i.e., i_0) and focus on the behavior of peripheral nodes (i.e., i_1, i_2). We then verify whether our results are robust to controlling for within-star spillovers. Our test, described formally in Section 3.4, compares treatment effects in stars to treatment effects in pairs. In the absence of spillovers, these treatment effects should be comparable.

Moreover, the network in Figure 1 is not exactly a collection of isolated stars: there are links across stars, which open the way to *cross-star spillovers*. In Figure 2, treating link i_0 - i_1 may affect i_0 's behavior, which, in turn, may affect j_1 's behavior. We, therefore, decomposed the network into a collection of stars, identifying cross-star links such as the i_0 - j_0 link in Figure 2, so that within and cross-star spillovers are clearly defined for each randomized treatment assignment. Later, we show that the results are robust to controlling for cross-star spillovers by comparing the treatment effects of nodes that have cross-star links (e.g., j_1) to treatment effects of nodes that do not have such cross-star links (e.g., j_2).

Star decomposition. We decompose the network into stars by iteratively removing sets of nodes that form stars from the network. We start from network g_0 , which is an undirected version of the network described above.






The algorithm starts with network g_0 , and operates as follows: at each iteration t , given network g_t , identify the node(s) I_t that have the highest number of neighbors that have only one connection on g_t . There may be several such nodes in case of ties. Note that each node $i \in I_t$ and her one-connection neighbors $N_i^1(g_t)$ on network g_t jointly form a star $S_i(g_t)$. Pick at random a node $i_t \in I_t$. Remove $S_{i_t}(g_t)$ from g_t and define $g_{t+1} = g_t - S_{i_t}(g_t)$ as the resulting graph. Repeat until period T in which g_T is empty. The collection $S_{i_0}(g_0), \dots, S_{i_T}(g_T)$ is the star decomposition of the network.

Using the network in Figure 2 as an example, we have $I_0(g_0) = i_0$. The star $S_{i_0}(g_0)$ includes all nodes with a black link to i_0 , because they are all leaves. We remove $S_{i_0}(g_0)$ from g_0 and

define $g_1 = g_0 - S_{i_0}(g_0)$. Now $I_1(g_1) = j_0$. Thus, the star $S_{j_0}(g_1)$ includes all remaining nodes. The algorithm removes link i_0-j_1 .

The top panel of Table 1 reports the outcome of the star decomposition algorithm on the harassment network (Figure 1), counting the number of stars by size and peripheral nodes in each. The algorithm led to the removal of 115 cross-star links.³

Table 1: Statistics about the intervention

		Harassers		Victims	
Panel A. Size		N	N stars	N	N stars
	2 (pair)	127	127	127	127
	3	12	6	24	12
	4	6	2	21	7
	5	8	2	8	2
	6 and more	238	2	840	38
Panel B. Treatment		All		Do not accept DMs	Accept DMs
Control		172		300	127
Public		212		388	108
Private		7		0	97
Total		391		688	332
		1020			

Notes. The top panel shows the distribution of stars by size and peripheral nodes of each type in each star size. E.g., there are 12 harassers that are peripheral nodes in stars of size 3, and there are six such stars. Note that pairs are reported twice, as central and peripheral nodes are symmetric in pairs. The bottom panel reports the number of peripheral nodes assigned to each treatment condition. For harassers, the private condition implies that their matching victim was assigned to the private condition. These are excluded from the analysis.

3.3 Treatment conditions and randomization

The NGO administered the following treatments (see Supplementary Appendix Figure SA-2 for a visual representation). In the *control* condition, no action was taken. In both the *public* and *private* conditions, the NGO reported to Twitter the latest harassment post that the harasser directed to the specific victim, then sent a message indicating that the post was reported, condemning the harasser’s behavior, and expressing support for the victim.

In the *public* condition, the following message was publicly posted as a reply to the incriminating post:

This post constitutes harassment and violates Twitter’s rules. Please [*@victim_handle*] know that you are not alone. I have reported this post to Twitter so they can take the necessary actions.

³By removing links, the algorithm may create isolated nodes, or turn nodes that were both harassers and victims (i.e., they sent a harassment post to one person and received a harassment post from another person) into either harassers or victims. Type reassignment was minimal (17 nodes, see Supplementary Appendix Table SA-3) for details).

In the *private* condition, the following message was privately sent as a direct message (DM) to the victim:

This post [`post_url`] constitutes harassment and violates Twitter’s rules. Please know that you are not alone, and I have reported it to Twitter so they can take the necessary actions.

We assisted the NGO in randomly assigning the nodes of the decomposed, star network to the different treatment arms. Since only users whose account accepts DMs can be sent DMs, we stratify treatment assignment by the size of the star and whether the victim accepts DMs.

We randomized treatment assignment as follows:

1. Assigned each node to treatment or control with a stratification that depends on the size of the star (i.e., number of spokes). We assigned half of the isolated pairs to treatment and the other half to control. Within each larger star, we assigned a pre-determined number of nodes to treatment and the remaining node to control.⁴
2. Among nodes assigned to treatment, we assigned all nodes in stars where the peripheral nodes are harassers or who do not accept DMs to the public condition. For the remaining nodes, peripheral victims who accept DMs, we assigned each node to the public condition with a probability of 0.5 and to the private condition with a probability of 0.5.

Recall that when we have stars, we focus on the treatment assignment of the peripheral nodes and exclude central nodes from the analysis. Moreover, we exclude harassers in pairs whose victim was assigned to the private message condition from the analysis since only the victim saw the message, and it is unclear how this treatment should affect harassers.

The bottom panel of Table 1 reports the number of nodes assigned to each treatment condition. Supplementary Appendix Tables SA-4, SA-5, and SA-6 respectively show balance for harassers, victims who do not accept DMs and victims who accept DMs.

3.4 Estimation and robustness

We estimate the following pre-registered specification using Ordinary Least Squares (OLS), separately for victims and harassers:

$$y_{is} = \alpha_s + y_{is,0}\beta_0 + y_{is,0}p_i'\beta_1 + x'_{is,0}\beta_2 + t'_i\gamma + \epsilon_{is}, \tag{1}$$

with y_{is} a post-treatment outcome measure for user i in star s , $y_{is,0}$ is the de-meanded pre-treatment outcome, p_i is a vector of propensity scores, $x_{is,0}$ is a vector of LASSO-selected pre-treatment covariates,⁵ t_i a vector of treatment indicators, and ϵ_{is} an error term. We make within-star comparisons by introducing star fixed-effect α_s . Since the model in equation (1) is estimated separately for victims and harassers, fixed effect α_s cannot be estimated for nodes that belong to isolated pairs. Consequently, we lump all such nodes in the same level α_0 . We report randomization inference p-values that account for uncertainty in treatment allocation across the network.

As pre-registered, we carry out one-sided hypothesis tests when the direction of the test has been pre-registered and the sign of the estimates is consistent with the pre-registered hypothesis, and carry out two-sided tests otherwise.

⁴Stars of size 3: 1 node to control, 1 node to treatment. Stars of size 4: 1 node to control, 2 nodes to treatment. Stars of size 5: 2 nodes to control, 2 nodes to treatment. Stars of size 6+: 2/5 nodes to control, 3/5 nodes to treatment.

⁵For each outcome, we use all pre-treatment covariates listed in Appendix A, and use LASSO to select those included in the model.

Testing for cross-star spillovers. We test whether our estimates are robust to controlling for cross-stars spillovers by estimating the following model:

$$y_{ist} = \alpha_s + y_{is,0}\beta_0 + y_{is,0}p'_i\beta_1 + x'_{is,0}\beta_2 + a_i\beta_2 + t'_i\gamma + \epsilon_{ist}, \quad (2)$$

where a_i indicates that node i has a cross-star link (see example depicted in Figure 2). If cross-star spillovers are negligible, then our estimates of γ should be robust to adding controls for spillovers.

Testing for within-star spillovers. If there are no within-star spillovers, then treatment effects for nodes pertaining to isolated pairs should be no different from treatment effects for nodes pertaining to larger stars. Let $g_s \in 1, 2, \dots$ be the number of peripheral nodes in star s . In an isolated pair, we have $g_s = 1$. We test for the relevance of within-star spillovers by adding an interaction term equation between treatment assignment and log star size to (1):

$$y_{ist} = \alpha_s + y_{is,0}\beta_0 + y_{is,0}p'_i\beta_1 + x'_{is,0}\beta_2 + t'_i\gamma + t'_i \log(g_s)\delta + \epsilon_{ist} \quad (3)$$

If within-star spillovers are negligible, then δ should not be significantly different from 0.

4 Hypotheses and results

4.1 Hypotheses

We pre-registered the following hypotheses.

The intervention should reduce harassment. We expected $\gamma < 0$ for all outcomes in the “Interaction” category (see Appendix A), both for the sample of harassers and that of victims. These outcomes include the number of posts sent by harassers to their victims, as well as their toxicity.

The intervention should reduce the activity of harassers. For the sample of harassers, we expected $\gamma < 0$ for all outcomes in the “Activity” category, including the extent to which their accounts were active, their posts and reply posts, and the number of users they followed.

The intervention should increase the activity of victims. We expected $\gamma > 0$ for all outcomes in the “Activity” category, for the sample of victims.

The intervention should decrease follower engagement with harassers. We expected $\gamma < 0$ for all outcomes in the “Engagement” category, including the number of followers and follower engagement with their posts (replies, reposts, quoted reposts, and likes), for the sample of harassers.

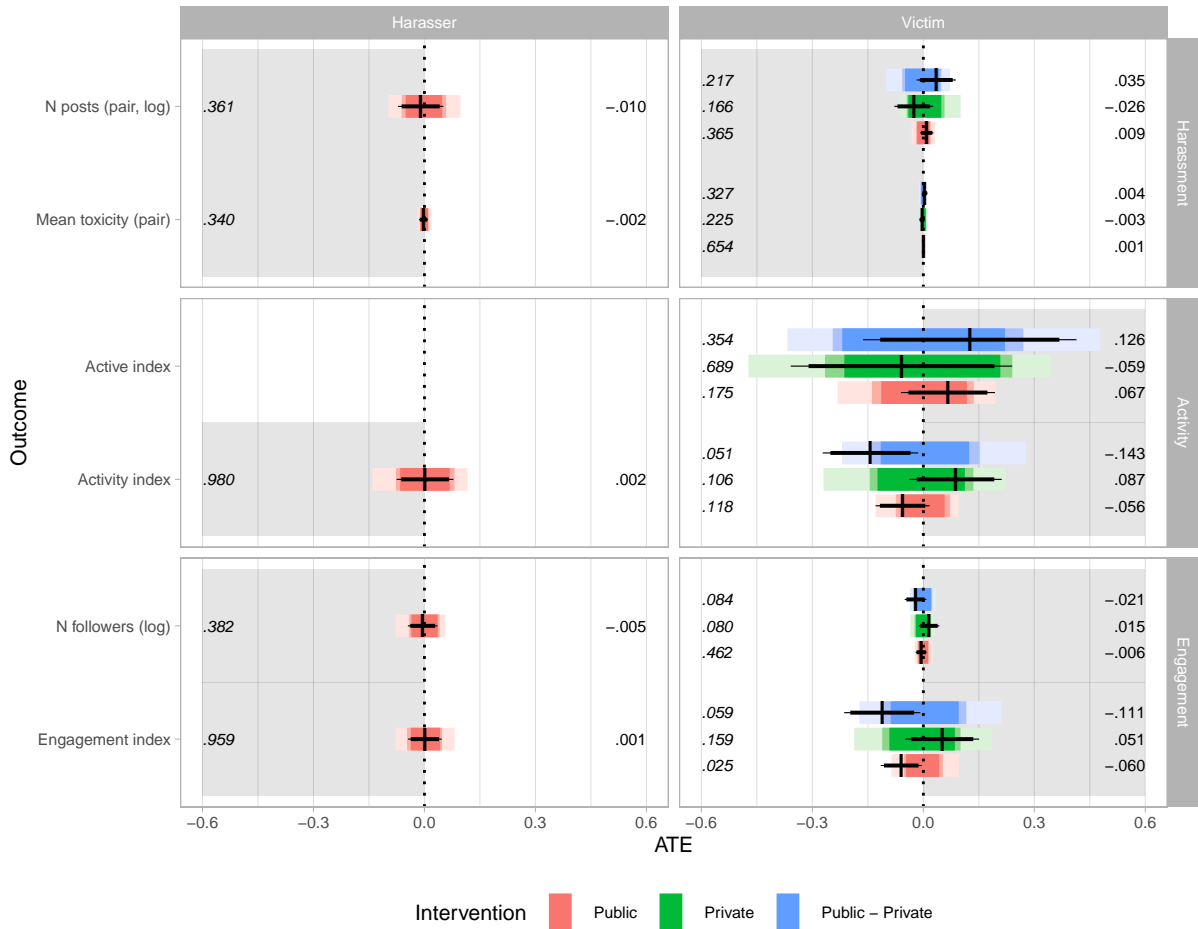
The intervention should increase follower engagement with victims. We expected $\gamma > 0$ for all outcomes in the “Engagement” category, for the sample of victims.

4.2 Results

Figure 3 presents our baseline pooled-sample results for harassers and victims.⁶ We focus our discussion on estimated coefficients reported on the right-hand side and randomization inference p-values reported on the left (more details in table note).

⁶See Supplementary Appendix Tables SA-7 and SA-8 for more details.

Figure 3: Pooled-sample results.



Notes. The shaded area corresponds to the direction of the pre-registered hypothesis. We report point estimates on the right-hand side and randomization inference p-values on the left-hand side. We report 1-sided p-values when the sign of the point estimate is consistent with the pre-registered hypothesis and 2-sided p-values when the sign of the point estimate is either inconsistent with the pre-registered hypothesis or when there was no pre-registered hypothesis. The vertical lines represent point estimates, and the horizontal lines represent frequentist confidence intervals of 90% (thick) and 95% (thin). The colored regions represent the distribution of randomization inference parameters under the null hypothesis, with the darkest and second darkest areas representing, respectively, the 90% and 95% confidence bands. “Public - Private” represents the difference in point estimates between the public and private treatment conditions, whose sign was not pre-registered. The “Active index” outcome could not be estimated for harassers due to insufficient variation.

The top-left panel shows no effect on the number posts from harassers to victims, nor their toxicity. The estimates in the top-right panel confirm this result, even when differentiating between victims who were treated publicly or privately. The effects are both insignificant and substantively small.

Turning to the middle-left panel, we see no results for whether harassers maintained active accounts post-intervention (“Active index”). This indicates that, contrary to our expectations and consistent with Twitter’s puzzling response that the reported posts did not constitute harassment, all harassers remained active throughout the weeks following the intervention, and thus, we lack variation. Also against our expectations, the middle-right panel shows no

treatment effects on whether both privately and publicly treated and victims maintained active accounts on Twitter.

Next, we turn to account activity, measured by posts and replies to others' posts, and the number of accounts followed, for both harassers and victims ("Activity index"), as well as follower engagement with their posts, measured by the number of followers ("N followers") and an index of their replies, reposts, quoted reposts, and likes ("Engagement index"). The middle- and bottom-left panels show no effects for harassers on these outcomes. These coefficients are all precisely estimated zeros.

In turn, the middle-right panel suggests distinct treatment effects on victims' activity depending on whether they were privately or publicly treated. Those privately treated exhibit a significant 0.09 standard deviation increase in activity (one-sided p-value of 0.11) relative to control victims. Disaggregated results in Supplementary Appendix Table SA-8 show that this effect is driven by a 19% increase in the number of posts (0.13 one-sided p-value) and a 25% significant increase in the number of replies to others' posts (0.08 one-sided p-value). In contrast, those publicly treated have a statistically insignificant but suggestive 0.06 standard deviation drop in activity (0.12 two-sided p-value), driven by a 13% drop in posts (0.15 two-sided p-value), and a 13% drop in replies (0.15 two-sided p-value). Importantly, the difference in the effects on indexes, posts, and replies across treatments are statistically significant (0.05, 0.09, and 0.04 two-sided p-values, respectively)

Likewise, the bottom-right panel suggests distinct treatment effects on follower engagement with victims depending on whether they were privately or publicly treated. Those privately treated exhibit a significant 1.5% increase in the number of followers (0.08 one-sided p-value) and an insignificant but suggestive .05 standard deviation increase in follower engagement (one-sided p-value of 0.16) relative to control victims. Disaggregated results in Supplementary Appendix Table SA-8 show that this effect is driven by a suggestive 29% increase in the number of replies to their posts (0.12 one-sided p-value). In contrast, those publicly treated have a negligible drop in the number of followers, and a significant .06 standard deviation drop in follower engagement (0.03 two-sided p-value), driven by drops in almost all engagement metrics: -26% replies to their posts (0.02 two-sided p-value), -15% quoted reposts (0.07 two-sided p-value), -25% likes (0.04 two-sided p-value), and -40% impressions (0.02 two-sided p-value)). The difference in the effects on the engagement index is statistically significant (0.06 two-sided p-value), as it is for all the other outcomes but the number of reposts.

These results collectively indicate that harassers were completely unaffected by the intervention. In turn, privately treated victims demonstrate higher levels of activity and associated engagement, while publicly treated victims exhibit relatively lower levels of activity and engagement. While the effects on privately treated victims were expected, those on publicly treated victims were not. We hypothesize several mechanisms driving these divergent treatment effects: (i) victim empowerment – a sense of increased support for the victim; (ii) victim fear – concerns about further scrutiny or potential repercussions; (iii) fear-driven disengagement – apprehension among followers that engaging with the victim's posts could make them targets as well; and (iv) activity-driven (dis)engagement – the organic tendency of followers to interact more (less) with victims that post more (fewer) content.

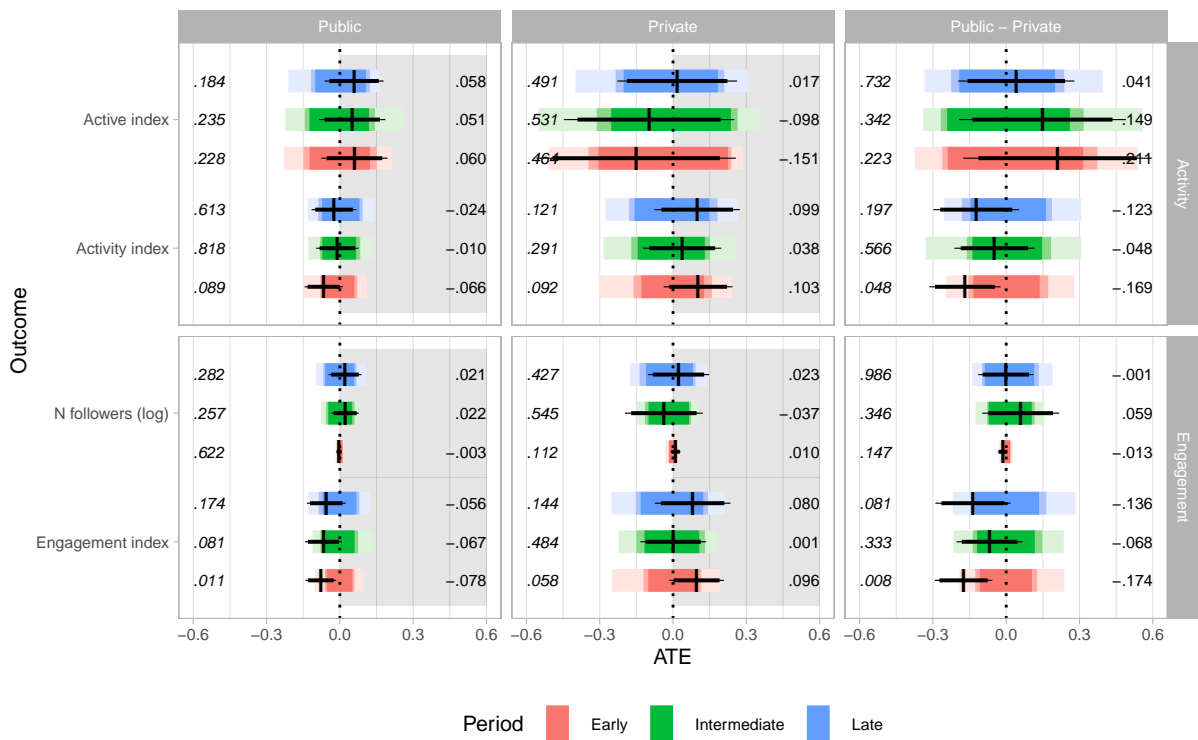
Our design allows disentangling these mechanisms. Both the public and private treatment condition may empower victims. Feelings of empowerment should increase victim activity and, in turn, follower engagement. This is consistent with our results on the private treatment condition. However, the public treatment condition is also observed by harassers and followers. As such, it may also exacerbate victim fear, causing them to reduce their activity and their followers to reduce their engagement. Follower disengagement may be either fear-driven or activity-driven. Fear-driven mechanisms cannot be at play in the private treatment condition, as the intervention is observed neither by harassers nor followers.

The observed increase in activity among privately treated victims is likely due to their greater

sense of empowerment compensating for any increased fear. However, the pooled-sample results cannot disentangle whether the effect on follower engagement is driven by fear- or activity-driven disengagement.

Turning to our split sample analysis, which distinguishes between early, intermediate, and late treatment effects,⁷ Figure 4 shows that publicly treated victims only reduce their activity substantively and significantly early after treatment (-0.07 standard deviations, $p = .09$) but that follower engagement with their activity exhibits a persistent, significant drop, thus pointing to fear-driven disengagement rather than activity-driven disengagement. Moreover, consistent with the pooled-sample results, victims assigned to the private treatment condition see a persistent effect on activity and associated engagement.

Figure 4: Dynamic effects, victims



Notes. The figure follows the same conventions as Figure 3. Effects for harassment outcomes are reported in Supplementary Appendix, Figure SA-3

4.3 Robustness

Consistent with our pre-registered analyses, we conduct a series of robustness checks to address concerns about spillovers across and within stars. First, to mitigate concerns of cross-star spillovers, following the specification in equation (2), Supplementary Appendix Tables SA-9 and SA-10 replicate Tables SA-7 and SA-8 while controlling for spillovers across stars. The results remain virtually unchanged.

Second, lessening the concern of within-star spillovers, Supplementary Appendix Tables SA-11 and SA-12 show that the interaction of treatment assignment and star size is generally a precisely estimated zero.

⁷Supplementary Appendix Figure SA-4 corroborate that harassers were completely unaffected by the intervention. Supplementary Appendix Figures SA-5 and SA-6 show that results are robust to splitting the post-treatment sample in two periods as opposed to three. Supplementary Appendix Figures SA-7 and SA-8 show that the results are robust to considering pre-treatment periods in a panel setting.

5 Conclusion

This study evaluates the effectiveness of counterspeech interventions in an authoritarian context, where harassment on social media often targets opposition figures with politically motivated attacks. Unlike in democratic settings, where counterspeech strategies have shown promise, our findings reveal starkly different dynamics in authoritarian regimes. Public interventions, despite their visibility, appear to suppress victim activity and follower engagement, likely driven by fear of retaliation, while private interventions provide a safer avenue for victims to remain active and engaged.

These results underscore the limitations of applying counterspeech strategies developed for democratic contexts to authoritarian settings. They also highlight the need for intervention designs that account for the heightened risks victims and their supporters face in such regimes. Harassers' persistent activity further emphasizes the challenges of curbing online harassment in environments with limited platform enforcement and political complicity. By addressing these challenges, this study contributes to the broader literature on counterspeech, content moderation, and social networks, offering practical insights for NGOs and policymakers.

References

- Aridor, Guy, Rafael Jiménez Durán, Ro'ee Levy, and Lena Song**, "Experiments on social media," *Forthcoming in the Handbook of Experimental Methods in the Social Sciences*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4893781, 2024.
- Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski**, "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment," *SSRN Working Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4307346, 2023.
- Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova**, "Social Media and Xenophobia: Theory and Evidence from Russia," *Working Paper*, https://drive.google.com/file/d/1oYpM6XQyge9JOMwWOMHR_Q2aJUxyZYWJ/view, 2024.
- Chen, Yuyu and David Y. Yang**, "The Impact of Media Censorship: 1984 or Brave New World?," *American Economic Review*, 2019, 109 (6), 2294–2332.
- Donati, Dante, Nandan Rao, Victor Orozco-Olvera, and Ana Maria Muñoz-Boudet**, "Can Facebook Ads Prevent Malaria? Two Field Experiments in India," *World Bank Policy Research Working Paper 10967*, <https://hdl.handle.net/10986/42412>, 2024.
- Enríquez, José Ramón, Horacio Larreguy, John Marshall, and Alberto Simpser**, "Mass Political Information on Social Media: Facebook Ads, Electorate Saturation, and Electoral Accountability in Mexico," *Journal of the European Economic Association*, 2024, 22 (4), 1678–1722.
- Guriev, Sergei and Daniel Treisman**, "Informational Autocrats," *Journal of Economic Perspectives*, 2019, 33 (4), 100–127.
- Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, Nicholas Bahrach, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Murias Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay**, "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment," *Proceedings of the National Academy of Sciences*, 2021, 118 (50), e2116310118.

- Hobbs, William R. and Margaret E. Roberts**, “How Sudden Censorship Can Increase Access to Information,” *American Political Science Review*, 2018, 112 (3), 621–636.
- Jiménez Durán, Rafael**, “The Economics of Content Moderation: Evidence from Hate Speech on Twitter,” *Working paper*, 2023.
- , **Karsten Müller, and Carlo Schwarz**, “The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany’s NetzDG,” *SSRN Working Paper*, 2023.
- Larsen, Bradley J., Timothy J. Ryan, Steven Greene, Marc J. Hetherington, Rahsaan Maxwell, and Steven Tadelis**, “Counter-stereotypical messaging and partisan cues: Moving the needle on vaccines in a polarized United States,” *Science Advances*, 2023, 9 (29), eadg9434.
- Müller, Karsten and Carlo Schwarz**, “Fanning the Flames of Hate: Social Media and Hate Crime,” *Journal of the European Economic Association*, 10 2020, 19 (4), 2131–2167.
- **and** – , “From Hashtag to Hate Crime: Twitter and Antiminority Sentiment,” *American Economic Journal: Applied Economics*, 2023, 15 (3), 270–312.
- Munger, Kevin**, “Tweetment effects on the tweeted: Experimentally reducing racist harassment,” *Political Behavior*, 2017, 39 (3), 629–649.
- Siegel, Alexandra A. and Vivienne Badaan**, “# No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online,” *American Political Science Review*, 2020, 114 (3), 837–855.
- Yildirim, Mustafa Mikdat, Jonathan Nagler, Richard Bonneau, and Joshua A. Tucker**, “Short of Suspension: How Suspension Warnings Can Reduce Hate Speech on Twitter,” *Perspectives on Politics*, 2023, 21 (2), 651–663.

Appendices

A Data dictionary

The indicators followed by a star (*) are available post-treatment (April 21 - June 6, 2023), but not pre-treatment (January 1, 2022 - April 20, 2023). Indicators of account activity and follower engagement are used as outcomes and pre-treatment covariates; indicators of account authenticity are used as pre-treatment covariates. All variables are winsorized at their 99th percentile. Indices of n variables correspond to the first factor of a factor analysis with $n - 1$ factors.

- Main outcomes
 - **N posts (pair)**: number of posts sent by the harasser to her victim (harasser outcome), or number of posts received by the victim from her harasser (victim outcome), each log-transformed.
 - **Mean toxicity**: mean toxicity score of posts sent by the harasser to her victim (harasser outcome), or number of toxic posts received by the victim from her harasser (victim outcome).
 - **Active index***: index of a binary indicator that equals 1 if the user's account was active every day in the post-treatment period*, 0 otherwise, and the number of days during which the user's account was active in the post-treatment period*.
 - **Activity index**: index of the number of posts, number of reply posts, and number of accounts followed by the user, each log-transformed.
 - **N followers**: number of accounts that follow the user, log-transformed.
 - **Engagement index**: index of the number of replies to, reposts of, quotes of, likes of, and impressions of the user's posts, each log-transformed.
- Control variables: we control for each of our outcomes, evaluated at pre-treatment (when available), as well as the individual variables that make up our indices, and indicators of account authenticity, including whether the profile includes a picture, description, URL, real name, and real location.