When Should Platforms Break Echo Chambers?

Mohamed Mostagir* James Siderius†

January 8, 2023

Abstract

Recent calls for regulation of social media platforms argue that they serve as conduits of extremism. Several platforms have responded by banning communities that peddle extreme or misleading ideas. These communities are usually echo chambers, consisting of users with similar ideologies repeating the same information to each other. This amplifies harmful beliefs and makes them more likely to metamorphose into dangerous offline actions. We develop a novel community formation model to show that this traditional view of echo chambers is incomplete, and that they sometimes can even lead to an overall reduction of harmful sentiment on the platform. Our model offers a nuanced understanding of these community dynamics and how they shape the structure of optimal interventions in non-trivial ways. For example, policies that successfully contain extremism in the short run can be the exact same policies that sow the seeds for extremism in the long run and vice versa. We provide several such insights that platforms and policymakers can use as a starting point for developing effective interventions that reduce extremism and misinformation.

1 Introduction

There is mounting evidence suggesting that recent conflicts like the US Capitol insurrection in January 2021 and the Defund the Police riots in Portland in 2020 had their origins in the echo chambers of online social media, where people with similar beliefs meet regularly to exchange information (see McEvoy (2021); Frenkel (2021)). The fact that these real-world actions are facilitated by virtual interactions has been documented in several contexts, e.g., Acemoglu et al. (2018); Tufekci (2017). As a result, social media platforms like Reddit and Twitter have found themselves at the center of multiple controversies and accusations of providing fertile grounds for the formation of echo chambers. These chambers amplify extremism and increase polarization, and a natural question is what can platforms do in order to eliminate or limit their harmful effects.

Platforms have attempted to address these echo chamber threats in various ways. For example, in 2020 Reddit completely shut down the subreddit r/The_Donald,[1] a right-leaning community where hundreds of thousands of users congregated and peddled misinformation and hate speech. Prior to the ban, Reddit employed a softer intervention, choosing instead to *quarantine* the subreddit, which

*University of Michigan Ross School of Business, mosta@umich.edu
†Massachusetts Institute of Technology, siderius@mit.edu

[1] In Reddit terminology, a community is referred to as a subreddit. (Also, see Appendix B for context and applicability of community-based social media on other platforms.)

keeps the community active but makes it difficult for users to access it or contribute content. Twitter routinely implements similar strategies, making it difficult to find controversial hash tags around which users aggregate and perpetuate dubious postings (Hwang and Lee, 2021). The assumption underlying these interventions — given the aforementioned literature — is that as online beliefs become more extreme, the chances that they spill over into the real world in the form of costly offline actions also increases. Such actions range from individual "lonewolf" acts like the infamous Pizzagate (Fisher et al., 2016) to collective acts like the ones mentioned in the beginning of the paper. This is both socially damaging and harmful to the platform's public image, and by implementing interventions like quarantining or banning communities, platforms are hoping to limit the amplification effects of echo chambers. This in turn prevents beliefs from becoming too extreme and decreases the likelihood of costly offline actions.

The space of interventions that platforms can implement is vast. Most of the attention is usually directed to informational interventions, e.g., censorship, because of how they interact with core values like freedom of expression. These interventions rely on controlling the information that people see. In this paper, we focus on policies that, at a fundamental level, can be described as *communication interventions*, where the platform does not interfere with the informational content but instead adjusts the ability of agents to communicate with one another, like the aforementioned ban and quarantine policies. If agents find it more difficult to congregate around dangerous ideas with like-minded individuals, then perhaps this would tune down the exaggerated beliefs that emerge from these interactions and the actions that could potentially ensue. Our goal is to study whether these interventions are effective in controlling the perceived harmful effects of echo chambers and to prescribe when and how they should be deployed, if at all.

Our starting point is a simple but novel model of community formation rooted in the empirical finding that people are more likely to join communities with similar beliefs (Marsden, 1987; Mosleh et al., 2021). As we show, this simple twist injects so much complexity and interesting behavior into the standard belief dynamics studied in the social learning literature. In that literature, agents update their beliefs by interacting with their neighbors and/or through occasional random interactions with others in their network. This modeling feature mirrored the nature of interactions at the time these models were developed, where interactions were mostly offline and people moved in relatively limited social circles, like their neighbors or coworkers, and had little control over the beliefs of others within these circles. These classic interaction dynamics do not capture the current reality where people can find and connect with anyone anywhere purely based on their beliefs. This requires a model that describes how communities form and beliefs evolve in these settings, which we provide in this paper and believe is one of its primary contributions.

A crucial point in studying these dynamical systems is that interventions like the ones mentioned above will alter belief dynamics and significantly shape the resulting communities in terms of size and belief composition. Indeed, this is the idea behind the intervention to begin with: if the platform can influence community structure and beliefs, then it can steer these communities in favorable directions, namely, towards moderate beliefs that are less likely to lead to costly offline actions. Our paper shows that these interventions can behave very differently in how communities look in the

short term versus in the long run. This implies that the nature of the intervention is completely dependent on the objective of the platform. As we demonstrate, shaping beliefs to influence a short-run outcome, like the results of an impending election or supporting a campaign to get people immediately vaccinated requires a very different intervention from a goal that aspires to shape beliefs in the long run. Interestingly, an intervention that perfectly addresses an immediate concern (for example, allows us to avoid the Capitol insurrection) can itself sow the seeds for even more extreme beliefs and costly offline actions in the distant future. Even more interesting is that, interventions that help reduce extremism in the long run can look completely baffling when viewed from a short-term lens, in the sense that they might seem unfair and targeting relatively innocent communities as opposed to extreme and problematic communities. This speaks to how politically fraught these interventions can be and how implementing effective policies can be challenging because of their unfavorable optics.

Contribution. Our paper makes the following contributions. On the technical side, we provide a novel model of community formation that is simple yet yields a plethora of rich outcomes. We show in Theorem 1 that this processes converges to a structure whose geometric properties are well understood. We then use this understanding to offer interesting insights about the nature of efficient interventions. We focus on *mild* and *strong* interventions throughout, paralleling the community quarantine and community ban discussed earlier. Theorem 2 shows that short-term interventions rely on a simple threshold structure: when the size of an extreme community (a community with dangerous beliefs) is small, the platform should ban it, but when it is too large, then the platform should not take any action that limits access to this community. This might seem paradoxical, since one would imagine that a large problematic community poses the most threat, but the reason is the following: as the (many) members of the large extremist community lose access to it, they start frequenting other (moderate) communities more. In doing so, they move the beliefs of these communities more towards extremism. In a sense, the echo chamber *protects* the rest of the communities from extreme beliefs by keeping these beliefs contained in the chamber.

We then turn our attention to long-run interventions. Despite the simplicity of our model, the evolution of the system of community membership and beliefs becomes quite complex. While we are able to provide a clean prescription for short-term interventions, a similar result for the long run is elusive. An important takeaway is that there is no "set it and forget it" intervention: an optimal short-term intervention can put the system on a trajectory that leads to communities becoming extreme in the long run. Proposition 4 provides conditions under which such interventions can be identified so that they are avoided or used with caution. In further highlighting this complexity, we offer some interesting insights about the nature of interventions that are effective in the long run. In particular, we define simple and complex interventions as these interventions that target the problematic community directly (simple) or that targets a community other than the problematic one (complex) in order to influence the problematic community. Indeed, the latter might be the optimal intervention in the long run. As noted, such complex interventions face barriers (e.g., political) to implementation beyond the technical difficulty of identifying the structure of the intervention. Fortunately, Proposition 5 shows that a simple intervention always exists for short-term objectives.

In addition, we provide conditions in that guarantee that the optimal intervention in the long run is

An important point when addressing problems like the one in this paper is the absence of an objective truth about what beliefs would be considered dangerous or are likely to lead to harmful actions. Throughout the paper, we assume that there is an exogenously-defined acceptable region of beliefs, such that whenever the beliefs of a community are within that region then there is no chance of costly action. In Section 6.4, we relax this and take a robust optimization approach to the problem when there is uncertainty about what this region is, and give a polynomial time algorithm to compute the optimal interventions in this case.

Related Literature. Our paper is related to two broad streams of literature. The first is the social learning and opinion formation literature, where most of the work studies conditions under which agents learn a true state of the world (Golub and Jackson, 2010). In these papers, agents eventually reach consensus and agree on the same opinion, even when they are more likely to interact with those with similar beliefs (Golub and Jackson, 2012), which is clearly not an outcome that is commonly observed in practice. To generate disagreement, previous work has resorted to imposing extra restrictions on the nature of communication between agents or in how they update their beliefs, e.g., in Hegselmann et al. (2002) and Blondel et al. (2009) agents stop communicating with anyone once their differences in opinions is beyond a certain threshold, or in Acemoglu et al. (2010) where some agents are stubborn and never change their opinions despite what they hear, or in Mostagir and Siderius (2022b); Mostagir et al. (2022) where an outsider manipulates information for a group of agents in order to generates disagreement. Our paper imposes no such restrictions – agents are free to join and interact in any communities they wish, yet disagreement and different beliefs still emerge as a natural outcome of the model.

Informational interventions to stop misinformation and extremism have been studied recently, e.g., Mostagir and Siderius (2022a) study censorship and equal coverage policies and when they do or do not work. As mentioned, the current paper focuses on interventions that affect community and network structures on the platform. There is a spurt of very recent empirical work that studies this problem in setups similar to the one we consider in this paper. Habib et al. (2019) show that one can predict the evolution of communities by studying previous bans and quarantines (of other communities) on Reddit, and suggest that this can be used for interim interventions. Agarwal et al. (2022) study "deplatforming," which is how members of a community respond when that community is shut down, e.g., they start their own alternative communities and/or join existing communities. Our model naturally extends to the study of deplatforming, which we do in Section 6.2. Mudambi and Viswanathan (2022) show evidence of negative spillovers (e.g., more verbal aggression) when communities are banned. Shen and Rose (2019) show how quarantining policies are viewed differently by users depending on their beliefs and affiliations.

The rest of the paper is organized as follows. Section 2 presents our model while Section 3 demonstrates the main results of the paper through a series of examples that show the wide range of interesting behavior arising from these deceptively simple dynamics. Section 4 formally derives the community structures of the model and describes their existence and uniqueness. Section 5 uses this understanding to examine the effectiveness of intervention policies in the short and long run. Several

modeling choices and extensions are left out of the main model to allow for parsimony, but lead to interesting insights which we explore in Section 6. Section 7 concludes the paper.

2 A Model of Community Formation

We consider a social media platform (e.g., Reddit) consisting of M communities (e.g., Reddit subreddits) that users can subscribe to.[2] Users form multi-dimensional beliefs/interests about d topics on the multi-dimensional spectrum $[0,1]^d$. Strong beliefs (or preferences) along a given dimension correspond to the extreme poles 0 and 1, and a moderate belief (or indifferent preference) corresponds to $1/2$. Beliefs can represent opinions on any number of topics, some of which may be divisive or controversial, and others which may be more innocuous.

Community Formation. Users arrive sequentially in discrete time $t = 1, 2, \ldots$. User t (at time t) is born with a belief vector $\mathbf{b}_t \sim H(\cdot)$, where H has a joint distribution over $[0,1]^d$, with continuous density h lower bounded by some constant $\underline{\mu} > 0$ over full support, and which satisfies single-peakedness.[3] Some dimensions may or may not be independent of each other. Subsequently, this user chooses a community $m_t \in \{1, \ldots, M\}$ in which to participate. Let $\mathcal{M}_{m,t}$ denote the set of users on page m at time t and call

$$\mathbf{b}^*_{m,t} = \begin{cases} \frac{1}{|\mathcal{M}_{m,t}|} \sum_{j \in \mathcal{M}_{m,t}} \mathbf{b}_j, & \text{if } |\mathcal{M}_{m,t}| > 0 \\ \frac{1}{2}\mathbf{1}, & \text{if } |\mathcal{M}_{m,t}| = 0 \end{cases}$$

as the average belief of agents in this community.[4] Each user prefers to engage in communities that agree with her belief, so agent t chooses community m that minimizes $||\mathbf{b}_t - \mathbf{b}^*_{m,t}||_2$.[5] At the same time, we assume users in the same community discuss with each other frequently and form a consensus at the average belief of participating users, $\mathbf{b}^*_{m,t}$.[6] Hence, at any point in time t, there are $|\mathcal{M}_{m,t}|$ users in community m all of which hold belief $\mathbf{b}^*_{m,t}$.

Costly Action. Negative societal impacts are measured by costly offline actions taken by platform users

[2] For our main results, we will assume the number of communities that users can join (M) is fixed for parsimony. With a natural change in the users' utility function that incorporates not only their distance to community beliefs but also participation (the size of the community), there is some finite M^* that arises endogenously in the steady-state community structure (see Proposition 7). Including this component of the utility function does not qualitatively affect any of our main results (which is why we exclude it), but is necessary when M is endogenous (to prevent every user from just starting their own community). We study this extension in detail in Section 6.2.

[3] Formally, fixing any $\mathbf{b}_{t,-j}$ (the beliefs of all dimensions except dimension j), the conditional distribution of H over dimension j always satisfies that density $\mathbb{P}_H[b_{t,j}|\mathbf{b}_{t,-j}]$ is increasing for all $b_{t,j} \leq \zeta^*_{t,j}$ and decreasing for all $b_{t,j} \geq \zeta^*_{t,j}$, for some $\zeta^*_{t,j} \in [0,1]$ that can depend on $\mathbf{b}_{t,-j}$. This eliminates bimodal preferences that can give rise to wonky community distributions.

[4] We assume unoccupied communities are initiated at a moderate belief vector $\frac{1}{2}\mathbf{1}$, which implies that users who do not partially match the beliefs of any community will start their own (when available), but otherwise will join an existing community. Many other initial configurations would lead to identical results, such as seeding each of the M communities with the first M agents to arrive.

[5] To most transparently deliver the insights of our model, we will focus on the special case where users join exactly one community. We discuss the more general multihoming formulation, where users can join and participate on many communities, in Section 6.3.

[6] Implicit in this assumption is that there is frequent exchange of ideas between all agents in a given community. Under mild assumptions, these beliefs will converge to a consensus belief roughly near the mean incoming belief of the population (e.g., see Golub and Jackson (2010)). This consensus belief will then evolve as new agents (with different incoming perspectives) join the community over time.

who occupy toxic communities propagating dangerous ideas. To formalize this, we define a region $\mathcal{R} \subset [0,1]^d$ known as the *acceptable* region, where agents cause zero offline conflict. For simplicity, we assume this region takes the form of a polyhedron $\mathcal{R} = \{\mathbf{b} : \mathbf{A}\mathbf{b} \leq \boldsymbol{\beta}\} \cap [0,1]^d$ (for some matrix $\mathbf{A} \in \mathbb{R}^{k \times d}$, vector $\boldsymbol{\beta} \in \mathbb{R}^k$, and number of constraints k), where beliefs $\mathbf{b}^*_{m,t} \in \mathcal{R}$ are considered "civil" and do not cause negative societal consequences. On the other hand, if $\mathbf{b}^*_{m,t} \notin \mathcal{R}$, then agents in community m engage in costly action from the actions incited within that community. This polyhedron represents some well-defined interior of the $[0,1]^d$ space, under the pretense that more passionate and extremist perspectives tend to be more likely to be the problematic groups.

Formally, we suppose there exists an increasing function $\mathcal{C}$ in the offline costly action participation, $\mu \equiv \frac{1}{N}\sum_{m=1}^{M} |\mathcal{M}_{m,t}| \cdot \mathbf{1}_{\mathbf{b}^*_{m,t} \notin \mathcal{R}}$. Some prototypical examples might include:

(i) a linear $\mathcal{C}$, which might represent the summation of one-time costly actions, for example, if each user with an extremist belief might take a dangerous action with some probability;

(ii) a threshold function $\mathcal{C}$, which might represent a collective action problem with a costly societal outcome, such as amassing sufficient support to storm the capitol building.

The platform is generally concerned about user participation that may result in costly offline action. This may stem from public backlash, legal responsibility, or a general concern for societal well-being from the actions of communities outside the acceptable region.

Platform Interventions. Platforms can implement either *mild* or *strong* interventions. Mild interventions correspond to content moderation policies that silo communities by making them more difficult to participate in (e.g, login walls) and by preventing visibility (e.g., posts cannot be shared outside the community). Strong interventions are aggressive content moderation policies that completely ban a given community and prevent similar content from being posted elsewhere. For reference, Reddit's content moderation interventions, as applied to the community r/The_Donald, are shown in Figure 16 in Appendix B. Following the same language as Reddit (but with applicability to broader community-based social media), we will often interchangeably refer to mild interventions as *quarantine policies* and strong interventions as *ban policies*.

We model these interventions as follows. Without loss of generality, let us assume the intervention is enacted for community 1. In the case of a mild intervention at time t, we assume that $\phi \in (0,1)$ fraction of the community 1 users move to the nearest adjacent community, i.e., the community $m^*_t = \arg\min_{2 \leq m \leq M} ||\mathbf{b}^*_{m,t} - \mathbf{b}^*_{1,t}||_2$. The other $1-\phi$ fraction of users remain in community 1. One can interpret ϕ as the fraction of "passive" users who leave the community after the extra hurdles are imposed by the mild intervention. In the case of a strong intervention, the entire community is shut down and all users from community 1 migrate to the adjacent community.

The platform finds it costly to intervene — arbitrarily or overly moderated content can lead to bad public relations or damage the reputation of the platform (e.g., free speech violations or political bias). It may also drive users to leave the platform in favor of a competitor or other outside option (e.g., leave social media altogether), costing the platform user engagement and ad revenue. While the platform cares about reducing costly offline action, we simultaneously assume that the platform faces a penalty for more aggressive interventions, which will discourage frivolous policies that have no meaningful

purpose. Formally, we assume the objective for the platform is to minimize $\mathcal{C}(\mu) + c \cdot (Q + B)$, where $\mathcal{C}(\mu)$ is the costly offline action with participation μ, Q is the number of quarantined communities, B is the number of banned communities. and $c \geq 0$ is the cost associated with an intervention.

Long-term Impacts of Interventions. We can also model the long-term impact of an intervention as new users who enter and decide which communities to join. In the case of a mild intervention, we assume the lack of visibility of content coupled with hurdles such as the login wall imply that a new user $t + t'$ will join community 1 with probability $1 - \phi' \leq 1 - \phi$ if $||\mathbf{b}_{t+t'} - \mathbf{b}^*_{1,t+t'}||_2 \leq ||\mathbf{b}_{t+t'} - \mathbf{b}^*_{m,t+t'}||_2$ for all m (i.e., community 1 is her preferred community), and with probability $\phi' \geq \phi$ will join the next-nearest community $m^*_{t+t'} = \arg\min_{2 \leq m \leq M} ||\mathbf{b}_{t+t'} - \mathbf{b}^*_{m,t+t'}||_2$ (i.e., she will join $m^*_{t+t'}$, her second preferred community). For a strong intervention (a ban), each new user joins her preferred community among $m \in \{2, \ldots, M\}$. To measure the long-term impacts of the intervention, we analyze the community structure at some time $T \gg t$.[7]

3 Illustration of Main Concepts

In this section, we present a demonstration of our model and explore the impact of different platform interventions on mitigating costly offline actions. In Sections 4 and 5 we formalize these insights and expand on them.

3.1 An Example of Community Formation

We consider a two-dimensional belief vector $\mathbf{b}_t$ for each incoming user t. The first dimension b^1_t represents the user's political ideology, with $b^1_t = 0$ indicating a user on the extreme left and $b^2_t = 1$ a user on the extreme right. The second dimension b^2_t represents the user's belief about fraud in the 2020 US election, which was won by the left-leaning candidate. We let $b^2_t = 0$ correspond to the belief that there was no election fraud and $b^2_t = 1$ correspond to the belief that there was certainly election fraud, with $b^2_t \in (0, 1)$ indicating some degree of uncertainty about the integrity of the election.

We assume the prior belief distribution (before users join the platform) looks like the heat map in Figure 1. Ideological preferences follow a truncated normal distribution around an average of $1/2$ (moderate) and for the most part, incoming users do not believe there was election fraud, even if their ideological beliefs lean more towards the right. However, naturally, there is some correlation between right-leaning ideology and belief in a fraudulent election.[8]

Costly action can arise in this setting from two forces. The first is just due to political extremism: there is evidence that as communities become too extreme (either on the left or the right), there is a higher likelihood of costly offline action (e.g., rioting), as documented by Shen and Rose (2019). The second force is due to the interaction between misinformation (the election was stolen) and political ideology (this negatively affects my preferred candidate). In our example, we assume costly action may

[7] A priori, it is not obvious that analyzing the community structure far into the future is well-defined, but as we establish in Theorem 1, there is indeed a limiting community structure that emerges which will be the focal analysis of long-term impacts.

[8] For a poll that empirically motivates these assumptions, see Durkee (2022).

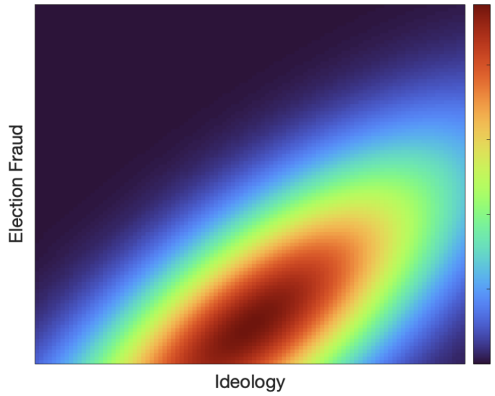


Figure 1. Arriving Prior Beliefs.

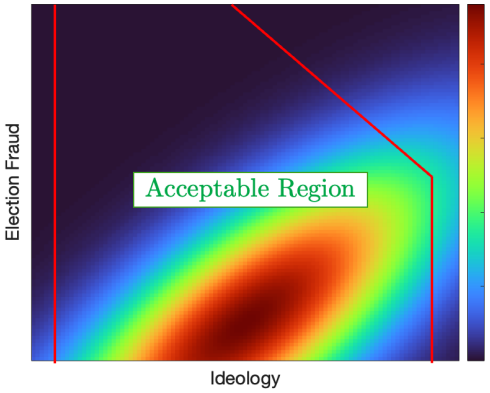


Figure 2. Costly Action Region.

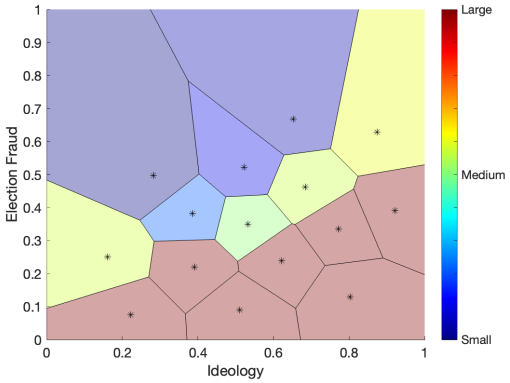


Figure 3. Communities and their Sentiments.

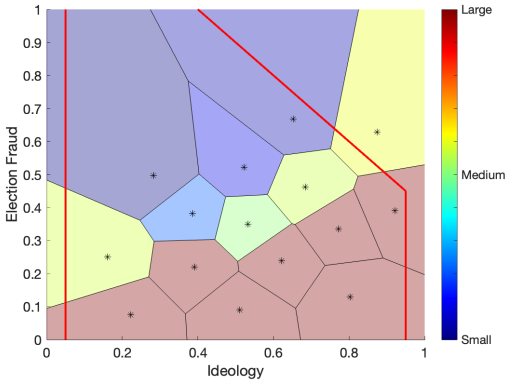


Figure 4. Costly Action Region.

also be taken for those with more moderate right-leaning ideologies, conditional on a strong belief in election fraud (the candidate the user supports was unfairly removed from office). These interactions allow us to define an *acceptable region* for beliefs, within which the emergence of costly action is less likely. The acceptable region $\mathcal{R}$ for our example can be seen in Figure 2. Note that in this example the region is asymmetric, accounting for the fact that a right-leaning ideology that is not too extreme can still lead to costly action when combined with a high belief in election fraud.

Under this belief distribution, online communities will converge to some equilibrium community structure, as depicted in Figure 3 (known as a *Voronoi diagram*). Each of the cells in the diagram represents one community, and the (prior) beliefs of the agents who will elect to join it. The asterisks represent the average sentiment of the community (known as its *barycenter*); because incoming beliefs are not drawn uniformly within each cell, it is possible for the average sentiment to be skewed away from its geometric center. Similarly, the "size" of the cell (the area or volume it occupies in the diagram) is not necessarily proportional to the size of its user base. Cells which are red represent more popular communities than those that are green or yellow, which in turn are more popular than those that are purple or blue. One can see that the largest communities represent a broad spectrum of ideological beliefs (ranging from fairly liberal to fairly conservative), but tend to have little belief in election fraud. Only smaller, fringe communities tend to believe in election fraud, and only one

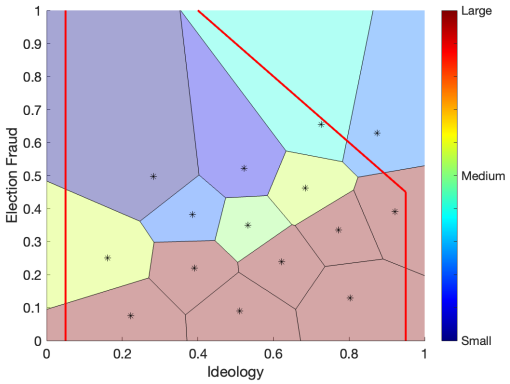


Figure 5. Mild intervention in the short term (quarantine).

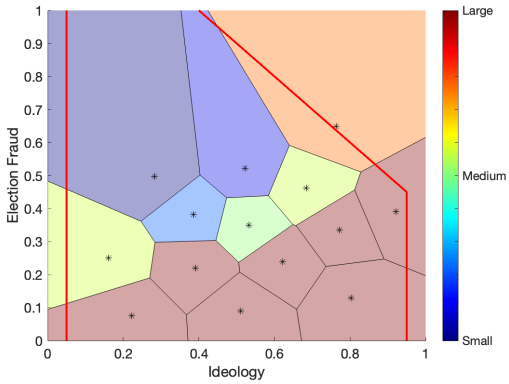


Figure 6. Strong intervention in the short term (ban).

community falls outside of $\mathcal{R}$, which is both strongly conservative and holds a firm conviction that the election was stolen (as shown in Figure 4). In our model, this community represents an online echo chamber that poses a threat in the form of societally costly offline action.

3.2 The Efficacy of Mild Interventions

Next, we consider a platform who wants to minimize costly action in the short term. Under the setting of Figure 4, there is one community inciting costly offline action, and the platform intends to limit the influence of the ideas in this community. We assume that $\phi = 1/2$, so a mild intervention (such as quarantining) has the short-term effect of displacing half of the users in this toxic community, whereas a strong intervention (such as a ban) displaces all of the users.

The results of this intervention are shown in Figures 5 and 6. The mild intervention shown in Figure 5 (known as a quarantine) results in a small exodus of users from the toxic community to an adjacent community (the more "mainstream" community). This leads to a small sentiment shift for the mainstream community, but which on the whole moderates many of the more extreme users by exchanging beliefs with the community that is less ideologically extreme. While the toxic community continues to exist, its user base is diminished as a consequence of the quarantine policy, depicted by the blue color (as opposed to yellow) relative to Figure 4. The quarantine policy ultimately reduces costly offline action and while the policy does not completely eradicate the problematic community, it effectively mitigates it by siloing a small group of extremists.

It is natural to think that in order to more effectively reduce offline costly action, the platform should take a more aggressive stance against the content and sentiment being shared in the problematic community. However, as we see in Figure 6, the strong intervention does more damage than no intervention at all (which in turn is worse than the quarantine policy). The authoritarian moderation policy that completely bans the toxic community leads to a large spillover into the mainstream community, resulting in a much larger population (orange instead of yellow) that still propagates harmful ideas and lies outside the acceptable region $\mathcal{R}$. In this sense, the intervention causes the toxic community to mass infect others, exacerbates extremist perspectives, and fosters

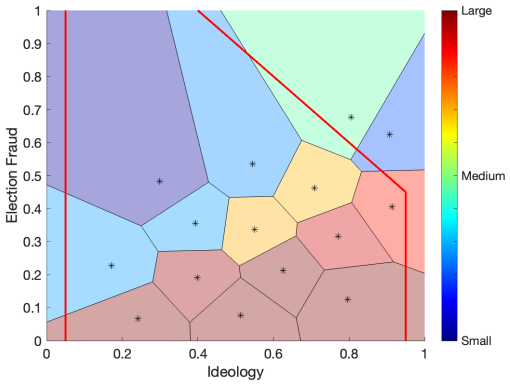


Figure 7. Mild intervention in the long term (quarantine).

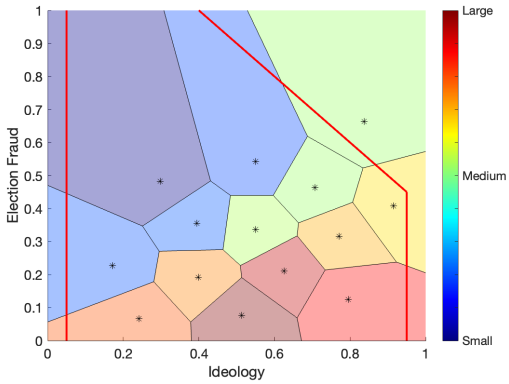


Figure 8. Strong intervention in the long term (ban).

more offline costly actions relative to no intervention whatsoever.

3.3 Short-term and Long-term Effects

In Section 3.2, we considered a platform whose objective was to design an effective intervention in the short term. Here, we instead look at the how community structure is affected in the long term following an intervention. Because an intervention results in a community structure that is no longer an equilibrium, it is critical for platforms to also study the properties of the new long-term Voronoi diagram once community equilibrium is reestablished. As in Section 3.2, Figure 4 serves as our baseline structure before any potential intervention, as before.

The effects of the interventions in the long term are shown in Figures 7 and 8, which can be compared to the short-term effects in Figures 5 and 6. In the short term, the quarantine policy is first-best, and in fact, the ban policy does strictly worse than no intervention at all:

$$\text{Quarantine} > \text{No Intervention} > \text{Ban} \qquad \text{(Short-Term Impact)}$$

We see that as the community landscape evolves over time, the efficacy of the policies actually reverse. The quarantine policy, which had been effective in the short term, ends up with undesirable consequences in the long term (even relative to the baseline of no intervention in Figure 4). The quarantine will drive new entering users of the platform to join other communities with broader appeal, slowly gravitating previously innocuous communities toward more polarized viewpoints. In other words, because the quarantine will now mostly attract those who are determined to spread misinformation and extremism, over time this community will be pulled more extreme. Simultaneously, the quarantine will send some entering extremists (who are potentially unaware of the quarantined community) to an adjacent community which eventually becomes problematic as well. As we show in Figure 7, the long-term equilibrium structure emerges with now *two* communities outside of the acceptable region $\mathcal{R}$, and the total population of both of these exceeds the original population of the one community in Figure 4.

For the strong intervention of a ban, the exact opposite happens. As we saw in Figure 6, the ban

initially backfires and leads to more immediate costly offline action by infecting a more moderate community with extremist perspectives. However, long term, we see this trend reverses (in Figure 8), and we end up with one problematic community but one that is much smaller than the one in the original community structure of Figure 4. This is because of how the equilibrium is reestablished following the ban with new incoming users. The problematic community after the short term becomes more ideologically extreme, attracting only (few) others that are similarly extreme. Eventually, this community transitions into more dangerous territory by gradually increasing its extremity, but while simultaneously losing its popularity. This paves the way for a new coalition to form with high conviction of election fraud but moderate ideological views, ultimately reducing the influence the problematic community has. Thus, the ban actually helps mitigate problems long term, as we summarize in:

$$\text{Ban} > \text{No Intervention} > \text{Quarantine} \qquad \text{(Long-Term Impact)}$$

The remainder of the paper generalizes this example to arbitrary beliefs and preferences, and provides formal results about optimal platform interventions under our model.

4 The Voronoi Structure of Communities

In this section, we first characterize the long-run distribution of users in the different communities as described in the sequential arrival model of Section 2. For this, we need to define some preliminaries. An M-cell *Voronoi diagram* of dimension d is uniquely represented by a set of barycenters $\{\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(M)}\}$ and a set of masses $\{\rho^{(1)}, \ldots, \rho^{(M)}\}$ corresponding to the (fraction of) population mass in each community (with $\sum_{k=1}^{M} \rho^{(k)} = 1$). The mth *cell* of the Voronoi diagram is defined as the region $\mathcal{A}^{(m)} = \{\mathbf{b} : ||\mathbf{b} - \mathbf{b}^{(m)}||_2 = \min_{m'} ||\mathbf{b} - \mathbf{b}^{(m')}||_2\}$ and the set of points $\mathcal{L}$ belonging to two (or more) distinct cells represents the *edges* of the Voronoi diagram.

4.1 Steady-State Community Equilibrium

We begin with a definition that specifies a special class of Voronoi diagrams of interest.

Definition 1. A Voronoi diagram is a *steady state* if:

(i) $\left(\int_{\mathbf{b}\in\mathcal{A}^{(m)}} h(\mathbf{b})\ d\mathbf{b}\right)^{-1} \int_{\mathbf{b}\in\mathcal{A}^{(m)}} \mathbf{b}\, h(\mathbf{b})\ d\mathbf{b} = \mathbf{b}^{*}_{m,\infty}$;

(ii) $\int_{\mathbf{b}\in\mathcal{A}^{(m)}} h(\mathbf{b})\ d\mathbf{b} = \rho^{(m)}$,

for all communities $m = 1, \ldots, M$.

The conditions for a steady-state community equilibrium are three-fold. First, the communities must form cells that obey a Voronoi structure with respect to their barycenters. In other words, a steady-state Voronoi diagram specifies a community structure consistent with the distribution of incoming beliefs that would indeed join that community (i.e., it must satisfy the properties of a Voronoi diagram). Second, the barycenter of each cell must be the conditional mean of underlying

beliefs (from distribution h) conditional on being drawn from within that Voronoi cell. Lastly, the community sizes of the Voronoi cells must reflect the density of that cell from underlying belief distribution h. Larger communities are not necessarily those with larger cells, but those with larger density concentration of incoming beliefs.

Proposition 1. *A steady-state Voronoi diagram always exists.*

Proposition 1 proves the existence of an equilibrium community distribution that depends only on the underlying belief density h. Under this community structure, incoming users self-select into communities that perpetually retain the community structure over time, both in terms of community beliefs and mass. These steady-state structures are the cornerstones of our community-based social media model, and identify the extent and intensity of online echo chambers. While the steady-state Voronoi condition of Definition 1 may seem fairly unrestrictive, in many cases these two conditions pin down a unique steady-state community structure, as shown in the next example.

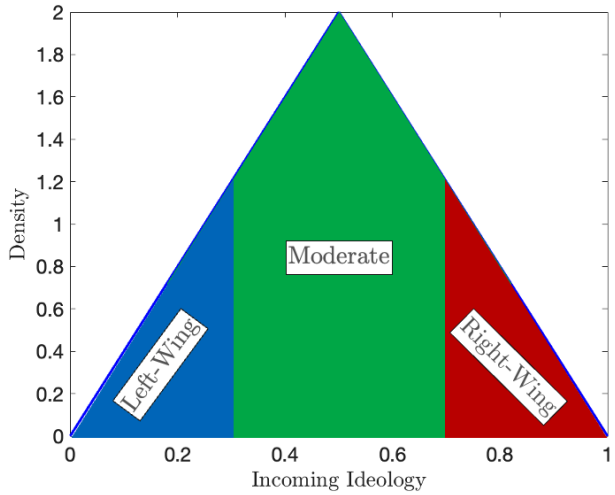


Figure 9. How communities segregate in Example 1 in the unique steady-state Voronoi diagram.

Example 1 (Steady-State Communities)**.** Consider a triangular distribution of ideological beliefs along the one-dimensional spectrum with density given by

$$h(b_t) = \begin{cases} 4b_t, & \text{if } 0 \leq b_t \leq 1/2 \\ 4 - 4b_t, & \text{if } 1/2 \leq b_t \leq 1 \end{cases}$$

as pictured in Figure 9. Under this distribution, more moderate ideological beliefs are more common, but more extreme beliefs on the left are mirrored by beliefs on the right with similar levels of extremism. Suppose there are $M = 3$ communities. Observe then that any steady-state Voronoi diagram is characterized by cutoffs (α, β) such that the Voronoi cells are determined by $\mathcal{A}^{(1)} = (0, \alpha)$, $\mathcal{A}^{(2)} = (\alpha, \beta)$ and $\mathcal{A}^{(3)} = (\beta, 1)$. Any steady-state Voronoi must satisfy:

$$\begin{cases} \alpha - \mathbb{E}[b_t \,|\, 0 \leq b_t \leq \alpha] &= \mathbb{E}[b_t \,|\, \alpha \leq b_t \leq \beta] - \alpha \\ \beta - \mathbb{E}[b_t \,|\, \alpha \leq b_t \leq \beta] &= \mathbb{E}[b_t \,|\, \beta \leq b_t \leq 1] - \beta \end{cases}$$

The first condition requires that the user with belief α be indifferent between joining community 1 and community 2, and the second condition requires that the user with belief β be indifferent between joining community 2 and community 3 in the steady-state community distribution.

These equations admit a unique solution of $\alpha = 3/10$ and $\beta = 7/10$; thus, the unique steady-state community distribution has a left-wing community consisting of 18% of the population (with incoming beliefs $b_t \in [0, 3/10]$), a right-wing community consisting of 18% of the population (with incoming beliefs $b_t \in [7/10, 1]$), and a moderate community consisting of 64% of the population (with incoming beliefs $b_t \in [3/10, 7/10]$). This community distribution is shown in Figure 9, with the largest community consisting of more moderate agents having diverse ideological backgrounds, and the two smaller fringe communities with more extreme leftists and rightists. ∎

4.2 Convergence and Uniqueness

Our next main result shows the importance of characterizing steady-state Voronoi diagrams, as it related to the dynamic stochastic process of Section 2.

Theorem 1. *As $t \to \infty$, the community structure converges to a steady-state Voronoi diagram almost surely. Formally, for some barycenters $\{\mathbf{b}^*_{m,\infty}\}_{m=1}^M$ of a steady-state Voronoi diagram, $\lim_{t\to\infty} ||\mathbf{b}^*_{m,t} - \mathbf{b}^*_{m,\infty}||_2 = 0$ almost surely for all m.*

Theorem 1 is an important result. It shows that a limiting community structure always emerges, and that it necessarily looks like one of the steady-state structures satisfying the conditions of Definition 1. In the context of Example 1, as new participants join the platform according the triangular ideological distribution with three established online communities, they will self-segregate into two smaller communities with more extreme beliefs, and one large moderate community, exactly as characterized in Figure 9. Our next result shows that for a single topic (e.g., ideology), this limiting community structure is in fact unique as well.

Proposition 2. *If $d = 1$, there exists unique steady-state Voronoi diagram and the community structure converges to this unique limiting Voronoi diagram almost surely.*

Proposition 2 provides a stronger result than Theorem 1 in the case of a single dimension of beliefs and preferences. A straightforward generalization of Proposition 2 is applicable when beliefs are independent across dimensions. In this event, each dimension can be decomposed and analyzed separately, leading to a unique community distribution as in the single-dimensional case covered by Proposition 2 (in the Cartesian-product space). This can be a useful simplifying assumption for platforms wanting to study problematic dimensions in the long run, which admit a single, well-defined equilibrium community structure.

However, with multiple topics (i.e., $d > 1$) where beliefs may be correlated across dimensions (as in the illustrative example of Section 3), it may be possible that multiple steady-state Voronoi diagrams exist. This multiplicity arises from the richness of potential community interests once interaction between beliefs become possible. Under these conditions, it may be possible that platforms with the same initial conditions evolve differently, some of which may be more societally acceptable than

others. In general, this means that constant interventions may be necessary, depending on how communities form stochastically over time.

5 Optimal Interventions

We next consider optimal interventions from the perspective of the platform. For this, we assume there is just a single problematic community $\tilde{m}$ where $b^*_{\tilde{m},t} \notin \mathcal{R}$ but with $b_{m,t} \in \mathcal{R}$ for all $m \neq \tilde{m}$.[9] The platform's goal is to minimize costly action in the short term, but the platform may also have long-term considerations about costly offline action.

5.1 Short-Term Interventions: A Geometric Interpretation

We focus on short-term interventions where we presume that the Voronoi structure of communities is already given to the platform. Because the community structure in higher dimensions may not be unique (per Section 4), we instead focus on the short-term intervention of a platform with the goal of decreasing costly offline action with perfect knowledge of the current community structure. For this, we fix the barycenters $\{\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(M)}\}$ of the current Voronoi diagram and consider how the relative popularity of communities affects the optimal intervention.

Size of the problematic community. A common consideration for platforms is to not only identify a problematic community, but also size it up to quantify the magnitude of the potential harm it causes. Our next result shows how the optimal intervention is affected by the size of the problematic community $\tilde{m}$ and its "closest" community $\hat{m}$.[10]

Theorem 2. *There exist* $0 \leq \rho_1 \leq \rho_2 \leq 1$ *such that:*

(i) If $\rho^{(\tilde{m})}/\rho^{(\hat{m})} < \rho_1$*, the optimal platform intervention is to ban community* $\tilde{m}$*;*

(ii) If $\rho_1 < \rho^{(\tilde{m})}/\rho^{(\hat{m})} < \rho_2$*, the optimal platform intervention is to quarantine community* $\tilde{m}$*;*

(iii) If $\rho^{(\tilde{m})}/\rho^{(\hat{m})} > \rho_2$*, the optimal platform intervention is to take no action.*

The intuition for Theorem 2 can be seen in Figure 10, using the example presented in Section 3.2. Community spillovers work both to the advantage and disadvantage of the platform's intervention. On one hand, the intervention moderates the extremist perspectives in the problematic community by breaking the echo chamber and forcing them to interact with more moderate platform users. On the other hand, the intervention "infects" more moderate users with the extremist perspectives of those in the echo chamber, possibly converting more moderate users to adopt these perspectives, and resulting in more costly offline action. Theorem 2 shows that the ratio of the community sizes determines which intervention is optimal, and that in the case of moderately-sized problematic communities, a mild intervention such as a quarantine may be first-best.

[9] Our insights generalize to cases where there are multiple problematic communities; however, this analysis involves considerably more machinery. Because this provides fewer direct insights than focusing on just the single-community case, we study this stylized version first, and consider an algorithmic solution to the more general problem in Section 6.4.

[10] The community $\hat{m}$ closest to $\tilde{m}$ is the one that solves $\hat{m} = \arg\min_m ||\mathbf{b}^*_m - \mathbf{b}^*_{\tilde{m}}||_2$, which is generically unique.

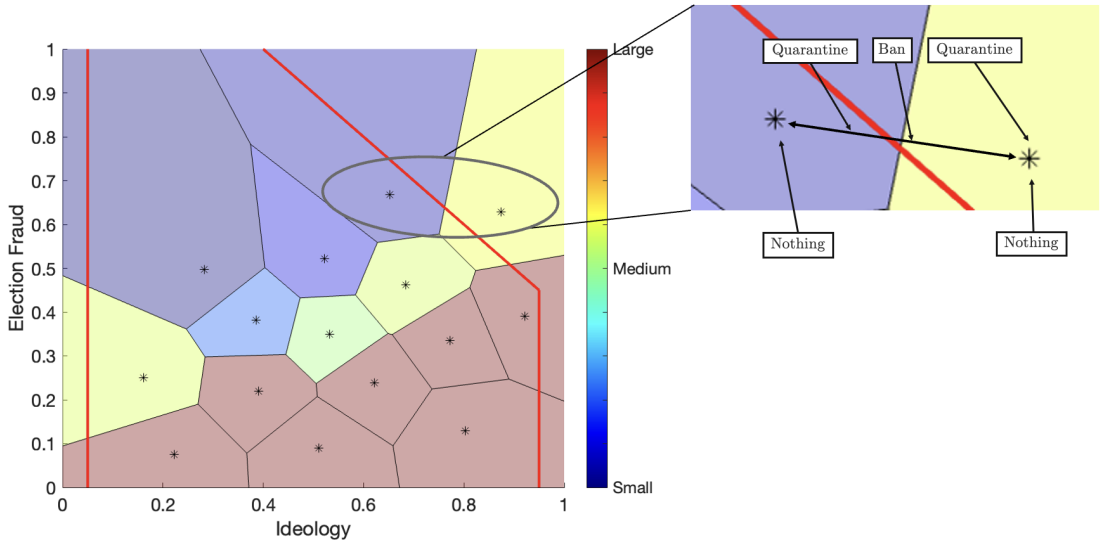


Figure 10. Intuition for Theorem 2: How will the ideological center of mass change after the short-run intervention?

The result is somewhat counterintuitive, as it goes against many current social media platform practices. Often, calls for regulation have been when toxic communities reach a critical mass that their offline actions mandate a response by the platform. It is exactly in these instances where strong interventions such as bans may be too late. Instead, the platform may need to turn to alternative solutions that are less authoritarian, and in more egregious cases, there may be in fact nothing the platform can do to alleviate the issue.

Tendency for costly offline action. Next, we can think about how short-term interventions are affected by the size of the acceptable region. When the size of the acceptable region shrinks, this corresponds to a situation where offline costly action is happening more frequently (all else equal), as relatively more moderate beliefs tend to also cause offline problems. We show that:

Proposition 3. *Let $\mathcal{R}' \subset \mathcal{R}$ such that $\tilde{m}$ is still the only problematic community. Then the intervention under $\mathcal{R}'$ is weaker than the intervention under $\mathcal{R}$.*[11]

As with Theorem 2, the result in Proposition 3 may go against conventional wisdom. With a worsening situation (shrinking $\mathcal{R}$), the platform will actually intervene *less*. The intuition can also be seen using Figure 10. As the barrier moves from right to left (corresponding to a shrinking $\mathcal{R}$) the optimal policy will move from a ban, to a quarantine, to no action. The reason is that with a worsening situation, the negative spillover effects begin to dominate the positive effects: the likelihood of having the problematic community influence a more mainstream one is steadily rising. There is thus a critical point where the problematic community is too far gone, and the platform cannot take any action to effectively save these agents.

[11] Here by "weaker" we of course mean that no action is a weaker policy than quarantining which in turn is weaker than banning.

Remark — A key assumption in Proposition 3 is that there is still only one problematic community even after $\mathcal{R}$ shrinks. When this assumption is violated, we get significantly richer optimal interventions that are non-monotone in the size of $\mathcal{R}$; for example, it is quite possible the optimal short-term intervention transitions from quarantining a community, to taking no action, to then banning it, as we continue to shrink the acceptable region. We study the subtleties of the optimal intervention when relaxing this assumption in Section 6.4.

5.2 Robust Interventions in the Long Run

While Theorem 2 fully characterizes the optimal intervention when the platform only cares about short-term goals, it is not generally the case that this intervention will lead to better long-term outcomes. As evident from Theorem 2, optimality of short-term interventions depends only on the *local* topology of the Voronoi diagram, i.e., the communities that are problematic or adjacent. In general, interventions that also lead to a reduction in cost in the long term depend much more on the *global* topology of the community structure. This makes a full characterization of optimality in the long term a much more challenging problem.

In this section, we look for effective short-term interventions that are also *robust*, in the sense that they will both correct the immediate issue and also ensure that the costly offline action associated with communities on the platform does not get worse in the long run. Formally, we say:

Definition 2. An intervention at problematic community $\tilde{m}$ at time t is *robust* if the expected offline costly participation μ does not increase following the intervention for any $T > t$.

In other words, robustness measures not only whether costly participation decreases immediately following the intervention, but whether it continues to remain at or below pre-intervention levels forever. This condition is much more demanding because the Voronoi diagram will adjust over time to re-establish an equilibrium distribution of communities. As we saw in the example presented in Section 3.3, it is likely for a successful short-term intervention to actually increase offline costly action in the long run, and vice-versa. Our concept of a robust short-term intervention represents a subset of effective short-term interventions that never increase expected costly action, even after recalibration of the community structure over time.

For this reason, we turn our attention to a necessary condition for robustness that also leverages only local properties of the community structure. We consider a short-term intervention on a problematic community $\tilde{m}$ that eliminates all costly offline action, i.e., that $\mu = 0$. Suppose that $\hat{m}$ is the community closest to $\tilde{m}$. Then we can define the new center of mass for community $\hat{m}$ as:

$$\bar{\mathbf{b}}^{(\hat{m})} = \left(\int_{\mathbf{b} \in \mathcal{A}^{(\hat{m})}} h(\mathbf{b})\, d\mathbf{b} + \phi' \int_{\mathcal{A}^{(\hat{m})}} h(\mathbf{b})\; d\mathbf{b} \right)^{-1} \left(\int_{\mathbf{b} \in \mathcal{A}^{(\hat{m})}} \mathbf{b}\, h(\mathbf{b})\, d\mathbf{b} + \phi' \int_{\mathcal{A}^{(\hat{m})}} \mathbf{b}\, h(\mathbf{b})\; d\mathbf{b} \right)$$

In other words, $\bar{\mathbf{b}}^{(\hat{m})}$ is the average short-term belief given that communities $\hat{m}$ and $\tilde{m}$ merge.[12]

[12] Implicit in the definition of $\bar{\mathbf{b}}^{(\hat{m})}$ is that the policy enacted is a quarantine policy; however, it can be generalized for ban policies as well by setting $\phi' = 0$.

Proposition 4. *An intervention is never robust if* $\bar{\mathbf{b}}^{(\hat{m})} \notin \mathcal{R}$ *and can only be robust if the ray* $\mathbf{b}^{(\hat{m})} \to \bar{\mathbf{b}}^{(\hat{m})}$ *intersects the boundary of a cell before intersecting the boundary* $\partial\mathcal{R}$ *of region* $\mathcal{R}$*.*

The intuition for the result relies on a positive feedback loop that generates a perpetually more extreme community. Recall that after an effective short-term quarantine intervention, the barycenter of the new community $\mathbf{b}^{(\hat{m})}$ may lie inside $\mathcal{R}$, but this will no longer be an equilibrium barycenter. If the recomputed center of mass $\bar{\mathbf{b}}^{(\hat{m})}$ lies outside $\mathcal{R}$, the incoming agents who join the community will induce a stochastic process that pulls the barycenter closer to $\bar{\mathbf{b}}^{(\hat{m})}$ over time. Simultaneously, the gradually more extreme barycenter will attract more extremist users as the sentiment of the community evolves; the resulting impact will be even more aggressive drift that exacerbates the problem. Thus, the requirement on $\bar{\mathbf{b}}^{(\hat{m})}$ provides a necessary condition on an effective short-term intervention that is simultaneously robust.

We briefly comment on the fact that sufficient conditions for robustness are much more difficult. One might imagine there exists a converse to Theorem 4, where if the ray $\mathbf{b}^{(\hat{m})} \to \bar{\mathbf{b}}^{(\hat{m})}$ is directed away from the boundary $\partial\mathcal{R}$, then by the same reasoning as before, the stochastic process that reestablishes the barycenter of the community in equilibrium will guarantee long-run stability. However, this reasoning is incomplete due to a "crowding out" effect that might induce previously innocuous communities to become more polarized and ultimately more problematic. The non-uniqueness of the steady-state Voronoi diagram amplifies this because it is nearly impossible to predict whether a candidate for a robust intervention will always lead to better future outcomes. Consequently, the essential platform problem of effective content moderation needs to be dynamic and ever-changing; however, Theorem 4 provides conditions under which we can identify candidate interventions that will not *necessarily* backfire in the long run.

5.3 Complex Interventions

In Sections 3.2 and 3.3, we focused on interventions that directly involved the toxic community. This naturally raises the question of whether it might ever be optimal (for reducing offline costly action) to intervene at communities that pose no direct cost, but affect the distribution of communities in richer ways. Toward this end, we can define a *simple intervention* as one that always intervenes at the problematic community $\tilde{m}$ (or does nothing at all). Conversely, a *complex intervention* is one that reduces offline costly action by intervening at some community $m \neq \tilde{m}$. We say a complex intervention is optimal if it reduces costly offline action (strictly) more so than any simple intervention would. Our next result establishes that:

Proposition 5. *For short-term objectives, there exists* $\bar{M} > 0$ *such that for all* $M > \bar{M}$*, there always exists an optimal intervention that is simple.*

Many platforms are concerned only with fixing echo chambers that are causing offline problems today. In this case, Proposition 5 guarantees there is always a solution in terms of simple interventions, where if the platform wants to optimally reduce costly offline action, it can always do so by intervening only at the community causing the costly action directly. This optimal intervention is then characterized by Theorem 2, which depends only on the relative populations of the problematic

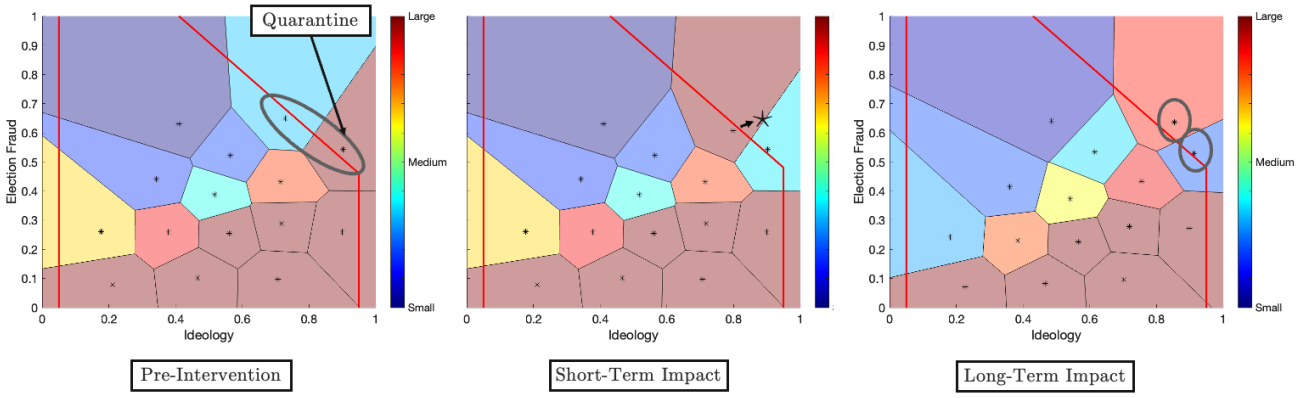


Figure 11. Long-term stochastic drift process for community $\hat{m}$ after the optimal short-term quarantine intervention at problematic community $\tilde{m}$.

community and its most adjacent one. In this sense, optimal short-term interventions only depend on local properties of the Voronoi structure.

On the other hand, as we discussed in Section 5.2, the outcome of long-run interventions depends much more on the global structure of the Voronoi diagram. As our next example shows, complex interventions can be optimal when long-term objectives are considered. This is exactly because the reestablishment of the equilibrium Voronoi diagram of communities can lead to a counterintuitive shuffling of the community structure. This may result in a platform (complex) intervention that can be more effective over time than any simple intervention enacted by a myopic platform.

Example 2 (Optimal Complex Intervention)**.** Suppose we are in the setting of Figure 11 (a different steady state Voronoi diagram than the one presented in Section 3, but for the same underlying belief distribution). The platform notices that there is a problematic community $\tilde{m}$ that, in the short term, could be partially corrected via a quarantine policy. This problematic community $\tilde{m}$ consists of right-wing extremists with a general disbelief in election integrity, but is not the community with the most election fraud misinformation. The closest community $\hat{m}$ is slightly more moderate on the ideological dimension, but holds a higher belief about election fraud. The short-term quarantine intervention would help reduce the prevalence and size of community $\tilde{m}$ by instilling more moderate ideological beliefs in those that leave to join community $\hat{m}$, leading to an effective short-term (simple) intervention.

However, using the same intuition underlying the condition in Theorem 4, this simple intervention will not be robust. In particular, over time the community will drift to become more ideologically extreme, as a solid fraction of incoming users are unaware of the quarantined community, and instead join the adjacent one, $\hat{m}$. This community will eventually also be problematic, holding both strong ideologies and strong beliefs in election fraud, resulting in a worse long-run outcome than no intervention whatsoever.

Instead, one can think about quarantining an innocuous community that is most ideologically extreme but which has much lower conviction that the election was stolen. This intervention is pictured in Figure 12. Because this is a complex intervention, it cannot beat the short-term

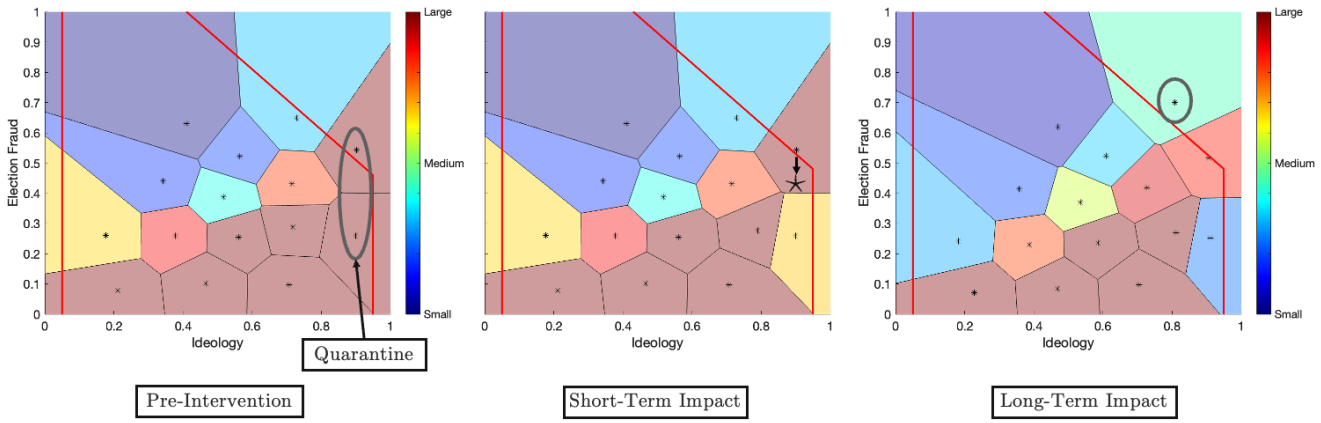


Figure 12. Long-term stochastic drift process for communities after an intervention at a relatively innocuous community $\hat{m}$.

intervention at community $\tilde{m}$ by Proposition 5. However, the quarantine intervention in this community will result in more right-wing extremists (but with little belief in election fraud) to join the problematic community $\tilde{m}$. This will result in a stochastic drift process that will slowly migrate the sentiment of community $\tilde{m}$ away from strong election fraud convictions, toward beliefs that fall within the acceptable region $\mathcal{R}$.

What emerges from the complex intervention then is a single problematic community that is relatively small (smaller than the pre-intervention size of $\tilde{m}$ or the size of the two problematic communities $\hat{m}$ and $\tilde{m}$ in the long run under the simple quarantine intervention). As a result, the optimal long-term intervention (which had a neutral effect in the short term) was a complex intervention at a community which at no point was the source of costly offline action. ∎

Example 2 shows how rich long-run dynamics can lead to new community structures over time, sometimes necessitating counterintuitive policies in the form of complex interventions. Complex interventions are difficult to motivate from the platform's perspective because it requires the platform to moderate content in a community where there is currently no problem directly stemming from the regulated community. At the same time, it useful for platforms to be aware that such interventions *can* be optimally effective long term, by limiting the interaction between extremism and misinformation by intervening in echo chambers that promote just one or the other.

6 Extensions

While Section 5 discusses the main insights from our model, there are many natural and salient extensions that can still be answered using the intuition developed. This section is organized as follows. In Section 6.1, we discuss *preemptive* interventions, which can be effective at steering community structure away from situations where platform interventions may eventually be too late to implement (as in Theorem 2(iii)). In Section 6.2, we modify the model to incorporate the possibility that users might start new communities or move platforms (e.g., from r/The_Donald to Parler), while in Section 6.3 we relax the assumption that users choose to only participate in a single community.

Finally, Section 6.4 uses a robust optimization framework to provide a more complete characterization of short-term optimal interventions in settings where many communities might be problematic and where there may be some uncertainty in the size of the acceptable region (e.g., through how one defines "misinformation" or the boundary where costly offline action becomes imminent).

6.1 Preemptive Interventions

Up until now, we have only considered short-term interventions (and their long-term implications) which act on steady-state Voronoi structures. However, it might be generally advantageous to anticipate problematic communities and intervene sooner rather than later to avoid potential future issues. Waiting for communities to stabilize may lead to a no-win situation, where Theorem 2(iii) kicks in and the platform's hands become tied (where no intervention is optimal). This motivates us to think about interventions that can moderate potentially problematic communities more proactively, before they hit a critical mass and reach an unsolvable state.

To formulate this, we assume that, as before, the platform wants to minimize costly offline action $\mathcal{C}(\mu)$ plus any intervention costs $c \cdot (Q + B)$ under the steady-state Voronoi community structure (where, as in Section 5, we assume there is one problematic community). However, we suppose the platform can intervene at some time t before this steady state has been reached, with the caveat that preemptive interventions at unproblematic communities are more costly, i.e., the platform faces cost $c' > c$ for quarantine and bans, respectively.[13] The platform minimizes this expression under an expectation of all possible steady-state Voronoi diagrams at $T \gg t$ following the intervention, while also being allowed to intervene at time T (a standard intervention of Section 5.1, with costs c_q and c_b). Our next result captures the interplay between the offline costly action function $\mathcal{C}$ and the efficacy of preemptive interventions. To keep the comparison fair across cost functions for the standard intervention (in steady state), we will assume that standard intervention costs are normalized at $c = 0$.[14]

Proposition 6. *Let $\bar{\mathcal{C}}$ and $\underline{\mathcal{C}}$ satisfy $\bar{\mathcal{C}}' > \underline{\mathcal{C}}'$ (pointwise). Then if a preemptive intervention is optimal under $\underline{\mathcal{C}}$, it is optimal under $\bar{\mathcal{C}}$.*

Phrased differently, Proposition 6 states that as costly offline action poses a greater threat (for all participation sizes), the platform needs to concern itself less with preemptive interventions. This may be slightly counterintuitive because as problems more easily manifest from online interactions in echo chambers, the platform can be more patient in when it intervenes.

The intuition is best seen by considering an example where the offline costly action function $\mathcal{C}$ represents a collective action problem. In situation (1), suppose the platform is worried about the formation of a gang, which imposes a cost of 1 if $\mu > 0.01$ and otherwise imposes a cost of zero. In situation (2), the platform is instead worried about the community building an arsenal of weapons

[13] This can be motivated in the same way as Section 5.3; interventions that target communities which are not directly causing problems can be difficult public relation moves that are hard to justify.

[14] Normalizing the standard intervention costs to 0 guarantees that the optimal short-term intervention is independent of $\mathcal{C}$ (for a proof see Appendix A). This provides a more direct comparison between how the offline action cost function affects just the platform's decision about preemptive interventions.

and tanks, which imposes a cost of 1 if $\mu > 0.1$ and otherwise imposes a cost of zero. Put differently, the costs of situation (1) and situation (2) are the same when $\mu < 0.01$ or $\mu > 0.1$, but the cost of situation (1) is strictly higher when $\mu \in (0.01, 0.1)$. Proposition 6 shows that a preemptive intervention will be more effective in situation (2). The reason is that if the low threshold $\mu = 0.01$ is where the problem first manifests itself, then the platform can always wait until this stage to ban or quarantine the community, with few side effects. However, if the high threshold $\mu = 0.1$ is where the problem first appears, it may be impossible to quarantine or ban the bad community without introducing large spillovers to other communities. This is exactly the scenario where a preemptive intervention may be optimal, despite the possible frictions associated with motivating such a policy.

6.2 Endogenous Community Origination and Deplatforming

Our baseline model of community formation assumes the number of communities is fixed at M. Instead, a more realistic (but less parsimonious) model is to allow users to start their own communities if none of the existing communities closely match their incoming preferences.

Here, we will extend the model to consider user incentives to start new communities. We will assume users value both how many other users participate in their community and how closely it matches their preference. Formally, we suppose that user t with belief $\mathbf{b}_t$ has utility function for joining community m given by $U_{m,t} = \varphi(|\mathcal{M}_{m,t}|, ||\mathbf{b}^*_{m,t} - \mathbf{b}_t||_2)$, where φ is strictly increasing in its first argument (the population of community m, $|\mathcal{M}_{m,t}|$, at time t), strictly decreasing in its second argument (the distance from incoming belief $\mathbf{b}_t$ and the community sentiment $\mathbf{b}^*_{m,t}$ at time t), continuous, and satisfies the conditions (i) $\lim_{k\to\infty} \varphi(k, \cdot) = \infty$, (ii) $\lim_{D\to 0} \varphi(k, D) = \infty$ for $k \geq 2$, and (iii) $\varphi(1, \cdot)$ is upper bounded.[15] Naturally, user t chooses the community m that maximizes her utility, i.e., $m^*_t = \arg\max_m U_{m,t}$. We let M_t denote the number of communities at time t; our next result shows that eventually, this number stabilizes at some constant M^*.

Proposition 7. *Almost surely, there exists some $T > 0$ such that $M_t = M_{t+1} = M^*$ for all $t \geq T$.*

Proposition 7 shows that it is largely without loss of generality to adopt our baseline model where the number of communities is fixed upfront. With richer (and more expressive) user preferences for community formation, there is an equilibrium number of communities that emerges which balances the tradeoff between social activity and community interest. One can study this richer model in the context of our simpler model after calibrating the number of communities that will form over time.

Another key assumption in our model is that users do not deplatform; that is, they never leave the platform following an intervention, but instead find an alternative home in another community on that same platform. We can apply this extension isomorphically to this setting. Historically, in the short term, fringe platforms tend to gain little traction. This is likely due to several frictions that push back against a swift and substantial exodus to another platform, including user familiarity, technological limitations to development, and the lack of other social media features (see Rogers

[15] With the reduced-form utility function we assume in the baseline model, every incoming agent would start their own community of one user (almost surely). This is, of course, unrealistic and neglects a key benefit of social media, which is to interact with other users. This version of the utility function will introduce a trade off between preference fit and more community activity.

(2020), Ali et al. (2021), and Jhaver et al. (2021)). However, in the long run, deplatforming can be a serious concern (e.g., Parler's user base grew by over 8000% in June 2022, but only after two years of the platform's existence).[16] Nothing about our model with endogenous community origination requires these communities to be co-located on the same platform; thus, one can study how deplatforming can impact the anticipated long-term effects of interventions through this richer model using identical analysis.

6.3 Multihoming: Participation in Many Communities

Another natural extension to our model is to suppose that users may actively participate in multiple communities simultaneously. One can model this through a *participation probability function*, P, which determines the probability that user t with belief $\mathbf{b}_t$ will join community m with sentiment $\mathbf{b}^*_{m,t}$. We will impose that P is strictly positive almost everywhere, and both weakly decreasing and continuous in $||\mathbf{b}^*_{m,t} - \mathbf{b}_t||_2$. Consistent with Definition 1, we can generalize our definition of steady state to this multihome setting.

Definition 3. A Voronoi diagram is a *multihome steady state* if:

(i) $\left(\int_{\mathbf{b}\in[0,1]^d} P(||\mathbf{b}^*_{m,\infty} - \mathbf{b}||_2)\, h(\mathbf{b})\; d\mathbf{b}\right)^{-1} \int_{\mathbf{b}\in[0,1]^d} \mathbf{b}\, P(||\mathbf{b}^*_{m,\infty} - \mathbf{b}||_2)\, h(\mathbf{b})\; d\mathbf{b} = \mathbf{b}^*_{m,\infty}$;

(ii) $\int_{\mathbf{b}\in[0,1]^d} P(||\mathbf{b}^*_{m,\infty} - \mathbf{b}||_2)\, h(\mathbf{b})\; d\mathbf{b} = \rho^{(m)}$,

for all communities $m = 1, \ldots, M$.

Definition 3 provides a more general notion of steady state where users may join multiple communities, but are more likely to join communities of nearer interest. For a few reasons, these steady states are more difficult to visualize compared to the Voronoi diagram steady states of our baseline model. First, the cells of the multihome diagram represent the incoming beliefs of users who are more likely to join that community instead of any other. However, because of the probabilistic nature of community formation, a user may decide to join slightly different communities than the one expected under the simpler, baseline model. Second, the steady-state diagram will not fully characterize the beliefs of individual users, because users may participate on multiple communities (and users who co-participate in one community may participate in different communities elsewhere). Instead, the barycenters of the multihoming Voronoi diagram capture the influence of the community on each participating user's average belief. In other words, each user's belief will be a convex combination of the belief barycenters of the communities she chooses to join. While the multihome model provides for richer community formation, the following corollary shows that this added generality leads to the same general equilibrium characterization.

Corollary 1. *There exists a multihome steady state; moreover, as $t \to \infty$, the community structure will converge to a multihome steady-state Voronoi diagram almost surely.*

[16] See https://appfigures.com/resources/insights/parler-beats-twitter-downloads for these statistics on Parler.

The main contribution of Corollary 1 is to demonstrate that the basic community formation model of Section 2 can be enriched without losing the same general properties. The main insights of Section 5 can also be applied in the context of Definition 3, under an appropriate transformation of the community structure to capture the full diversity of user beliefs.[17] Thus, the stylized assumption adopted in Section 2 that users joined only the single nearest community was largely for transparency and visual presentation of the results, but our main conclusions are not highly sensitive to this exact formulation.

6.4 Algorithmic Interventions and Optimizing under Uncertainty

To garner maximal insights about optimal platform interventions, in Section 5 we studied the case of a single problematic community and where the acceptable region $\mathcal{R}$ was known with certainty. Of course, it is possible for the platform to face multiple problematic communities simultaneously and have uncertainty associated with the region $\mathcal{R}$ that will cause offline costly action. We consider three different variants of this problem that span the frontier of potentially effective interventions:

(i) *Optimizing under Expectation* (EXP): The platform minimizes the objective by treating the acceptable region $\mathcal{R}$ as deterministic and equivalent to what it is in expectation. This is defined by the red lines in Figure 13.

(ii) *Optimizing under Worst Case* (WORST): The platform treats costly offline action as the worst it could be under uncertainty (i.e., the maximal $C(\mu)$ possible within the uncertainty range). This is defined by the blue lines in Figure 14 giving the smallest acceptable region possible.

(iii) *Optimizing under Best Case* (BEST): The platform treats costly offline action as the best it could be under uncertainty (i.e., the minimal $C(\mu)$ possible within the uncertainty range). This is defined by the blue lines in Figure 14 giving the largest acceptable region possible.

Optimizing under Expectation. For problem EXP, the platform chooses a subset $\chi_q \subset \{1,\ldots,M\}$ of communities to quarantine and a subset $\chi_b \subset \{1,\ldots,M\}$ of communities to ban, with $\chi_q \cap \chi_b = \varnothing$ and $\chi_{\text{nothing}} = \{1,\ldots,M\}\backslash(\chi_q \cup \chi_b) \neq \varnothing$. There are two cases to consider for an arbitrary community $m \in \chi_q \cup \chi_b$:

1. For community $m \in \chi_b$, let us enumerate an (ordered) list of the nearest communities to m that are in $\chi_{\text{nothing}} \cup \chi_q$. Let $(m^{(1)}, m^{(2)}, \ldots, m^{(k)})$ be the head of this list such that $m^{(k)}$ is the first community in χ_{nothing}. Each user in m sequentially draws a Bernoulli variable with probability $1-\phi'$; if successful, the user joins $m^{(1)}$, otherwise she repeats this process for $m^{(2)}$, and so on. Finally, if she is unsuccessful up until $m^{(k)}$, she simply joins community $m^{(k)}$ with probability 1.

[17] More formally, due to the fact there are $2^M - 1$ potential user beliefs after users have joined a non-empty subset of the M communities, the effects from platform interventions are slightly more nuanced. The platform can model this as $2^M - 1$ "pseudo-communities" which all have the same average belief, and then apply an intervention simultaneously to all pseudo-communities which contain the problematic community in the subset. After this transformation, one can analyze the efficacy of the intervention in the same way as in Section 5, and the full algorithmic characterization of the optimal short-term intervention follows immediately from the analysis in Section 6.4.

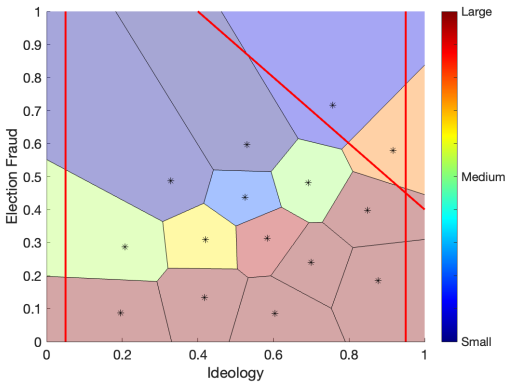


Figure 13. Estimate of Acceptable Region.

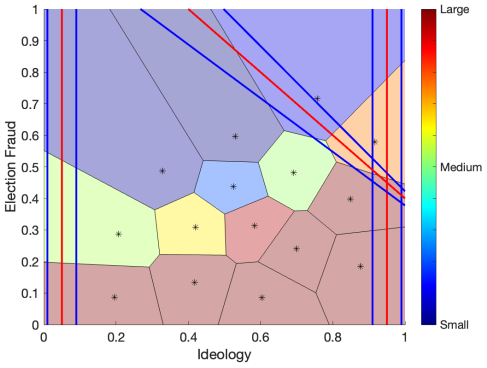


Figure 14. Acceptable Region under Uncertain Offline Costs.

2. For community $m \in \chi_q$, the process is the exact same as for $m \in \chi_b$ except the user first runs a Bernoulli trial with probability $1-\phi$; if successful, she remains in community m, otherwise continues identically to the case of a ban policy on community m.

The consequence of the platform's intervention(s) is a new Voronoi diagram with barycenters $\{\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(M)}\}$ and a set of masses $\{\rho^{(1)}, \ldots, \rho^{(M)}\}$. The platform's objective is to choose the intervention to optimize the objective:

$$\min_{\chi_q, \chi_b} \mathcal{C}\left(1 - \sum_{m=1}^{M} \mathbf{1}_{\mathbf{A}\mathbf{b}^{(m)} \leq \boldsymbol{\beta}} \cdot \rho^{(m)}\right) - \sum_{m=1}^{M} \left(c \cdot \mathbf{1}_{m \in \chi_q \cup \chi_b}\right)$$

which presumes knowledge of the acceptable region, $\mathcal{R} = \{\mathbf{b} : \mathbf{A}\mathbf{b} \leq \boldsymbol{\beta}\}$.

Optimizing under Worst Case. For problem WORST, we assume the acceptable region is not fully known, so while it is given by $\mathcal{R} = \{\mathbf{b} : \mathbf{A}\mathbf{b} \leq \boldsymbol{\beta}\}$, there is uncertainty in the sense that the platform only knows that $\mathbf{A}, \boldsymbol{\beta}$ satisfy $\mathbf{D}_i[\mathbf{A}]_i \leq \mathbf{f}_i$ for all i, and $\mathbf{C}\boldsymbol{\beta} \leq \boldsymbol{\alpha}$, where $\mathbf{D}_i \in \mathbb{R}^{\ell_i \times k}$, $\mathbf{f}_i \in \mathbb{R}^d$, $\mathbf{C} \in \mathbb{R}^{\ell' \times k}$, and $\boldsymbol{\alpha} \in \mathbb{R}^{\ell' \times 1}$ (for arbitrary dimensions ℓ_i, ℓ' for all i). Letting $\mathcal{U}_{\mathbf{A}} = \{\mathbf{A} : \mathbf{D}_i[\mathbf{A}]_i \leq \mathbf{f}_i \, \forall \, i\}$ and $\mathcal{U}_{\boldsymbol{\beta}} = \{\boldsymbol{\beta} : \mathbf{C}\boldsymbol{\beta} \leq \boldsymbol{\alpha}\}$ be the uncertainty sets associated with region $\mathcal{R}$, the platform solves:

$$\min_{\chi_q, \chi_b} \max_{\mathbf{A} \in \mathcal{U}_{\mathbf{A}}, \boldsymbol{\beta} \in \mathcal{U}_{\boldsymbol{\beta}}} \mathcal{C}\left(1 - \sum_{m=1}^{M} \mathbf{1}_{\mathbf{A}\mathbf{b}^{(m)} \leq \boldsymbol{\beta}} \cdot \rho^{(m)}\right) - \sum_{m=1}^{M} \left(c \cdot \mathbf{1}_{m \in \chi_q \cup \chi_b}\right)$$

Optimizing under Best Case. For problem BEST, one can similarly define the uncertainty sets $\mathcal{U}_{\mathbf{A}} = \{\mathbf{A} : \mathbf{D}_i[\mathbf{A}]_i \leq \mathbf{f}_i \, , \, \forall i\}$ and $\mathcal{U}_{\boldsymbol{\beta}} = \{\boldsymbol{\beta} : \mathbf{C}\boldsymbol{\beta} \leq \boldsymbol{\alpha}\}$; the platform then instead solves:

$$\min_{\chi_q, \chi_b} \min_{\mathbf{A} \in \mathcal{U}_{\mathbf{A}}, \boldsymbol{\beta} \in \mathcal{U}_{\boldsymbol{\beta}}} \mathcal{C}\left(1 - \sum_{m=1}^{M} \mathbf{1}_{\mathbf{A}\mathbf{b}^{(m)} \leq \boldsymbol{\beta}} \cdot \rho^{(m)}\right) - \sum_{m=1}^{M} \left(c \cdot \mathbf{1}_{m \in \chi_q \cup \chi_b}\right)$$

Our next result shows that all of these problems can in fact be solved in a tractable way:

Proposition 8. *There exist polynomial-time algorithms (in M) which solve the optimal short-term intervention in* EXP, WORST, *and* BEST.

The algorithm that obtains the polynomial-time bound is constructed in the proof of Proposition 8, located in Appendix A. The proof consists of two parts. First, we show that using the dual optimization problem (with a duality gap of zero), the platform can reduce both WORST and BEST (which are more general versions of EXP) to an optimization problem involving no uncertainty. This approach provides a tractable way of finding the appropriately-defined acceptable region under either the worst-case or best-case outlook. Second, we can write a scalable dynamic programming to find the optimal subsets χ_b and χ_q, which iteratively tries various combinations of interventions to minimize costly offline action.

Besides the potential practical applications of the algorithm given in Appendix A for Proposition 8, it can also provide insights related to the nuances of optimal interventions. For example, the intensity and number of interventions in the platform's optimal solution under EXP, WORST, BEST obeys no consistency monotonicity relationship. Despite the fact WORST is, in some sense, the most "risk averse", the subtle intuition about optimal interventions discussed in Section 5 carries over here. In particular, it might be that due to the potential to exacerbate costly action, the platform may take a more passive stand on content moderation under WORST than under EXP or even BEST. These counterintuitive outcomes underscores the importance of a general, tractable algorithm to compute the optimal intervention under uncertainty associated with how misinformation and extremism impact costly offline action.

7 Conclusion

Extremism, polarization, and misinformation are some of the most pressing social problems we face today. These problems directly stem from the advancements in networking technologies — a relatively recent phenomenon — that enable people to sort themselves into groups with similar beliefs. These beliefs then get amplified and can lead to a rapid growth of extremism, and one response is to try and reverse these effects through interventions that limit communication. These interventions usually take the shape of the policies we study in this paper, which curtail communication within groups or prevents it altogether based on their relative sizes and the information they circulate. This is a controversial approach that is nonetheless commonly used by social media platforms and advocated by some regulators and policymakers. Such approaches are constantly debated because of how directly they interact with the core organizing principles of society, and our paper contributes to this debate by trying to understand the short and long-term effects of these community interventions.

The central contribution of our paper is a simple communication model whose analysis yields unexpectedly rich outcomes. This model is unique in the literature because it does not lead to belief consensus while also endowing agents with complete autonomy over their communication choices. This allows us to analyze the interventions we are interested in by thinking of how they alter the dynamical system described by the community formation process. This model is likely to have applications beyond this particular setup, as it broadly captures the idea of how social networks endogenously emerge as a result of common beliefs and interests.

A Proofs

A.1 Auxiliary Lemmas

We prove a few geometric lemmas that will be useful throughout the remainder of the main proofs presented in Appendices A.2 to A.4.

Lemma A.1. *Consider a sequence of Voronoi diagrams, denoted by $V_1, V_2, \ldots$. Suppose that h has lower bounded support over $[0,1]^d$. Then there cannot exist two barycenters $\mathbf{b}_n^{(m)}$ and $\mathbf{b}_n^{(m')}$ for $m \neq m'$ with $\liminf_{n\to\infty} ||\mathbf{b}_n^{(m)} - \mathbf{b}_n^{(m')}||_2 = 0$.*

Proof of Lemma A.1. By definition, every Voronoi diagram partitions the d-dimensional space into cells $(\mathcal{A}_n^{(1)}, \ldots, \mathcal{A}_n^{(M)})$ with $\mathcal{A}_n^{(m)} \cap \mathcal{A}_n^{(m')} \neq \varnothing$ for all $m \neq m'$. Suppose, by way of contradiction, that $\liminf_{n\to\infty} ||\mathbf{b}_n^{(1)} - \mathbf{b}_n^{(2)}||_2 = 0$ and consider the set of all communities $m \in \{2, 3, \ldots, k\}$ where $\liminf_{n\to\infty} ||\mathbf{b}_n^{(1)} - \mathbf{b}_n^{(m)}||_2 = 0$ (it is without loss to call $m = 1$ and $m' = 2$, and suppose all other communities which have approaching barycenters to be communities $\{2, 3, \ldots, k\}$ for some $k \geq 2$). Let us call the common barycenter of all of these communities in the limit infimum to be $\tilde{\mathbf{b}}$.

We claim that the cells $(\mathcal{A}_n^{(1)}, \ldots, \mathcal{A}_n^{(k)})$ must have vanishing Lebesgue measure in $\mathbb{R}^d$ (in the limit infimum sense). If not, there is some cell associated with community ℓ with lower bounded measure $\delta > 0$ in $\mathbb{R}^d$, i.e., $\mathcal{A}_n^{(\ell)}$ has lower bounded measure for all n. This implies that there exists a radius $r > 0$ such that for all n, there is a ball B_r with radius r such that $B_r \subset \mathcal{A}_n^{(\ell)}$. Moreover, for any $\varepsilon > 0$, there exists some other cell $\mathcal{A}_n^{(\ell')}$ (with community $\ell' \neq \ell$) which has a barycenter satisfying $||\mathbf{b}_n^{(\ell')} - \tilde{\mathbf{b}}||_2 \leq \varepsilon$ at some index n, so it must be the case that $\tilde{\mathbf{b}}$ lies at most ε from an edge of $\mathcal{A}_n^{(\ell)}$. Notice then that then distance between the barycenter of community ℓ, $\mathbf{b}_n^{(\ell)}$, and the common limit point $\tilde{\mathbf{b}}$ must satisfy $||\mathbf{b}_n^{(\ell)} - \tilde{\mathbf{b}}||_2 \geq \frac{\underline{\mu}\pi^{d/2}(r-\varepsilon)^d}{\Gamma(d/2+1)}$, where $\underline{\mu}$ is the lower support of distribution h. For large enough n, we have sufficiently small ε such that $||\mathbf{b}_n^{(\ell)} - \tilde{\mathbf{b}}||_2$ is bounded from below (e.g., taking n large enough such that $\varepsilon \leq r/2$). This is a contradiction, so the Lebesgue measure of all cells $(\mathcal{A}_n^{(1)}, \ldots, \mathcal{A}_n^{(k)})$ must vanish.

If all of $(\mathcal{A}_n^{(1)}, \ldots, \mathcal{A}_n^{(k)})$ have vanishing Lebesgue measure, however, and this list is exhaustive of all cells with barycenters such that $\liminf_{n\to\infty} ||\mathbf{b}_n^{(m)} - \tilde{\mathbf{b}}||_2 = 0$, then there exists an improper Voronoi diagram for some n. For all other cells m' not on this list, it must be the case that $\liminf_{n\to\infty} ||\mathbf{b}_n^{(m')} - \tilde{\mathbf{b}}||_2 > 0$, and in particular because the number of communities is finite (i.e., $M < \infty$) there is some $\eta > 0$ such that $\liminf_{n\to\infty} \max_{m>k} ||\mathbf{b}_n^{(m')} - \tilde{\mathbf{b}}||_2 > \eta$. However, then any $\mathbf{b}$ satisfying $||\mathbf{b} - \tilde{\mathbf{b}}||_2 < \eta/4$ is closer to one of the communities in the list $(\mathcal{A}_n^{(1)}, \ldots, \mathcal{A}_n^{(k)})$ than any community not on the list. However, this is a contradiction because such a region has positive Lebesgue measure in $\mathcal{R}^d$. Therefore, our original premise that there can exist two distinct communities $m \neq m'$ with $\liminf_{n\to\infty} ||\mathbf{b}_n^{(m)} - \mathbf{b}_n^{(m')}||_2 = 0$ must be false. ∎

Lemma A.2. *Let us define the ray $\mathbf{b}^{(\tilde{m})} \to \mathbf{b}^{(\hat{m})}$ where $\hat{m}$ is the adjacent community to a problematic community $\tilde{m}$. This ray either (i) lies entirely outside of $\mathcal{R}$ (and $\mathbf{b}^{(\hat{m})} \notin \mathcal{R}$), (ii) intersects $\partial\mathcal{R}$ exactly twice (and $\mathbf{b}^{(\hat{m})} \notin \mathcal{R}$), (iii) intersects $\partial\mathcal{R}$ exactly once (and $\mathbf{b}^{(\hat{m})} \in \mathcal{R}$), or (iv) intersects $\partial\mathcal{R}$ on an interval (with $\mathbf{b}^{(\hat{m})}$ in that interval).*

Proof of Lemma A.2. Note that by definition of $\mathcal{R}$ (i.e., $\{\mathbf{b} : \mathbf{A}\mathbf{b} \leq \boldsymbol{\beta}\}$), the acceptable region is a convex set. We consider two cases separately:

(a) $\underline{\mathbf{b}^{(\hat{m})} \notin \mathcal{R}}$: We are either in case (i), or there exists some point along the ray $\mathbf{b}^{(\tilde{m})} \to \mathbf{b}^{(\hat{m})}$ (call it $\mathbf{p}$) such that $\mathbf{p} \in \mathcal{R}$. Observe that $\mathbf{b}^{(\tilde{m})} \to \mathbf{b}^{\hat{m}}$ must intersect $\partial\mathcal{R}$ at least twice. This is because the ray $\mathbf{b}^{(\tilde{m})} \to \mathbf{p}$ has $\mathbf{b}^{(\tilde{m})} \notin \mathcal{R}$ but $\mathbf{p} \in \mathcal{R}$ (so this ray intersects $\partial\mathcal{R}$ at least once) and the ray $\mathbf{p} \to \mathbf{b}^{(\hat{m})}$ has $\mathbf{p} \in \mathcal{R}$ but $\mathbf{b}^{(\hat{m})} \notin \mathcal{R}$ (so also intersects $\partial\mathcal{R}$ at least once). Simultaneously, observe that $\mathbf{b}^{(\tilde{m})} \to \mathbf{b}^{(\hat{m})}$ intersects $\partial\mathcal{R}$ at most twice. To see this, suppose there are at least three intersections, and in particular, let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the first two intersections on $\mathbf{p} \to \mathbf{b}^{(\hat{m})}$. Note that $\mathbf{x}_1 \in \mathcal{R}$ and $\mathbf{x}_2 \in \mathcal{R}$, but any convex combination of them is outside of $\mathcal{R}$, contradicting the convexity of $\mathcal{R}$.

(b) $\underline{\mathbf{b}^{(\hat{m})} \in \mathcal{R}}$: Because $\mathbf{b}^{(\tilde{m})} \notin \mathcal{R}$ but $\mathbf{b}^{(\hat{m})} \in \mathcal{R}$, the ray $\mathbf{b}^{(\tilde{m})} \to \mathbf{b}^{(\hat{m})}$ must intersect $\partial\mathcal{R}$ at least once. We show that if it intersects $\partial\mathcal{R}$ more than once, it must intersect it on an interval containing $\mathbf{b}^{(\hat{m})}$. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two distinct intersections; because $\mathcal{R}$ is convex, it must be the case that all convex combinations of $\mathbf{x}_1$ and $\mathbf{x}_2$ are also in $\mathcal{R}$. Because the ray $\mathbf{x}_2 \to \mathbf{b}^{(\hat{m})}$ must also lie entirely within $\mathcal{R}$ (given $\mathbf{b}^{(\hat{m})} \in \mathcal{R}$ and otherwise $\mathcal{R}$ would violate convexity), there must exist an entire interval $\mathbf{x}_2 \to \mathbf{b}^{(\hat{m})} \in \partial\mathcal{R}$. Moreover, taking this interval to be maximal, we see that no other points in $\mathcal{R}$ lie on this ray and that indeed it contains $\mathbf{b}^{(\hat{m})}$. ∎

Lemma A.3 (Atomic Packing)**.** *For every Voronoi diagram with M cells, there exist $0 < \underline{\gamma} < \bar{\gamma} < 1$ such that every Voronoi cell $\mathcal{A}^{(m)}$ satisfies $\underline{\gamma}/M \leq \lambda_{\mathcal{L}}(\mathcal{A}^{(m)}) \leq \bar{\gamma}/M$ with high probability as $t \to \infty$ (where $\lambda_{\mathcal{L}}$ is the Lebesgue measure with respect to $\mathbb{R}^d$).*

Proof of Lemma A.3. It is sufficient to consider H as the uniform distribution, and note there is an isomorphism between Voronoi diagrams (based on the mapping between probability distributions of $\mathbf{b}_t$) given that h is lower bounded by density $\underline{\mu} > 0$ and upper bounded by density $\bar{\mu} < \infty$. Under the uniform distribution, the Lebesgue measure of each cell can then be determined by the density of an atomic packing (of non-overlapping spheres) of maximal density within $\mathbb{R}^d$. Via the Minkowski-Hlawka theorem (see Conway and Sloane (2013)), we know that a lower bound for the density of the packing is $\zeta(d)/2^{d-1}$, where $\zeta(d)$ is the Riemann zeta function, which guarantees each Voronoi cell has a ball inscribed with a hypervolume lower bounded by $\zeta(d)/(2^{d-1} \cdot M)$, which in turn lower bounds the Lebesgue measure of the cell itself. Simultaneously, the upper bound of Cohn and Elkies (2003) guarantees that every Voronoi cell has a ball inscribed with radius upper bounded by $2^{-\omega d}/M$ for some constant $\omega > 0$. The result of Hui (2023) immediately implies that the Voronoi cell has a hypervolume upper bounded by $d \cdot 2^{-\omega d+1}/M$, which proves the claim for any fixed d. ∎

A.2 Proofs from Section 4

Proof of Proposition 1. The set of average community beliefs are represented by barycenters $(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})$ (i.e., vectors in $[0,1]^d$). Provided that $\mathbf{b}^{(m)} \neq \mathbf{b}^{(m')}$ for all $m \neq m'$ ("non-identical barycenters"), this uniquely determines the distribution of communities (i.e., the "cells") in the Voronoi diagram (but not their masses), up to a set of measure 0 in $\mathbb{R}^d$ (the "edges"). Given non-identical barycenters, we can define $V : (\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)}) \mapsto (\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(M)})$ to be the mapping of barycenters to cells, which is a partition of $[0,1]^d$ (i.e., $\mathcal{A}^{(m)} \cap \mathcal{A}^{(m')} = \varnothing, \forall\, m \neq m'$, with $\bigcup_{m=1}^{M} \mathcal{A}^{(m)} = [0,1]^d$) such that $\mathcal{A}^{(m)} = \{\mathbf{b} \in [0,1]^d : ||\mathbf{b} - \mathbf{b}^{(m)}||_2 = \arg\min_{m'} ||\mathbf{b} - \mathbf{b}^{(m')}||_2\}$. For non-unique barycenters, we can drop the dimension of M down until we have uniqueness of the barycenters,

and apply the same mapping, but where it is possible that $\mathcal{A}^{(m)} = \mathcal{A}^{(m')}$ for some pair $m \neq m'$. We will often write $V^{(m)}(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})$ to denote the mth cell given the barycenters $(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})$.

Let us define another mapping $\phi : [0,1]^{m \times d} \to [0,1]^{m \times d}$ which takes a set of barycenters $(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})$ (with well-defined Voronoi cells) and maps them to an alternative set of barycenters $(\tilde{\mathbf{b}}^{(1)}, \ldots, \tilde{\mathbf{b}}^{(M)})$ by setting:

$$\begin{aligned}\tilde{\mathbf{b}}^{(m)} &= \mathbb{E}_H\left[\mathbf{b} \,\middle|\, \mathbf{b} \in V^{(m)}(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})\right] + \varepsilon m \\ &= \left(\int_{\mathbf{b} \in V^{(m)}(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})} h(\mathbf{b})\, d\mathbf{b}\right)^{-1} \left(\int_{\mathbf{b} \in V^{(m)}(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})} \mathbf{b}\, h(\mathbf{b})\, d\mathbf{b}\right) + \varepsilon m\end{aligned}$$

where $\mathbb{E}_H$ is the conditional expectation with respect to distribution H and $\varepsilon > 0$ is a constant, but small, non-zero number. For any component of ϕ that maps above 1 (resp. below 0), we assume it maps to exactly 1 (resp. exactly 0), so that indeed ϕ is well-defined because it maps $[0,1]^{m \times d}$ into $[0,1]^{m \times d}$ (i.e., we simply cap all components above 1 at 1 and floor all components below 0 at 0).[18]

First, we argue that ϕ is continuous. Note that the Voronoi mapping V from barycenters to cells is continuous because the Euclidean distance is continuous. Moreover, an integral over continuous bounds (given that density h is integrable) is always continuous. Lastly, note that $\int_{\mathbf{b} \in V^{(m)}(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})} h(\mathbf{b})\, d\mathbf{b} \neq 0$ can be proven by induction on the total number of communities M. In the case of $M = 1$, this expression is equal to exactly 1. For $M > 1$, if the closest any two barycenters are is $\delta > 0$, then this integral is lower bounded by $\pi^{d/2}\delta^d / \Gamma(d/2+1) > 0$ (i.e., the volume of the d-dimensional sphere of radius δ). If two barycenters are exactly identical, then one can reduce this to the case of $M-1$ communities, which by the inductive hypothesis lower bounds the integral.

It is also immediate that $[0,1]^{d \times m}$ is compact and convex. Thus, applying Brouwer's fixed point theorem establishes that there exists a set of barycenters $(\mathbf{b}_*^{(1)}, \ldots, \mathbf{b}_*^{(M)})$ such that $(\mathbf{b}_*^{(1)}, \ldots, \mathbf{b}_*^{(M)}) = \phi(\mathbf{b}_*^{(1)}, \ldots, \mathbf{b}_*^{(M)})$. Moreover, for any $\varepsilon > 0$, we note that any fixed point must yield unique barycenters (i.e., $\mathbf{b}_*^{(m)} \neq \mathbf{b}_*^{(m')}$ for all $m \neq m'$) by construction. At the same time, as $\varepsilon \to 0$, it cannot be that $\mathbf{b}_*^{(m)} \to \mathbf{b}_*^{(m')}$ for $m \neq m'$ because h has lower bounded positive density over all of $[0,1]^d$ and any fixed point corresponds to unique barycenters and distinct Voronoi cells per Lemma A.1. Finally, taking $\varepsilon \to 0$, and setting every mass $\rho^{(m)} = \int_{\mathbf{b} \in V^{(m)}(\mathbf{b}_*^{(1)}, \ldots, \mathbf{b}_*^{(M)})} h(\mathbf{b})\, d\mathbf{b}$ equal to the density of cell m under the limiting fixed point, we obtain a steady-state Voronoi structure with non-identical barycenters. ∎

Proof of Theorem 1. The proof consists of two parts. The first is to show that every barycenter $\mathbf{b}_{m,t}^*$ must converge to some $\mathbf{b}_{m,\infty}^*$ almost surely. The second is to establish that the only such candidates for limit points are the fixed points of the map ϕ in the proof of Proposition 1.

Let us consider the stochastic process of evolving Voronoi diagrams $V_1, V_2, \ldots, V_t, \ldots$ that leads to population masses $(\rho_t^{(1)}, \ldots, \rho_t^{(m)})$ at time t. Note that by Lemma A.1, it must be the case that $\liminf_{t \to \infty} \min_m \rho_t^{(m)} > 0$ almost surely (i.e., all communities contain a non-vanishing fraction of the population); otherwise, there would be a Voronoi cell associated with community m^* with vanishing Lebesgue measure in $\mathbb{R}^d$ as $t \to \infty$. As a result, because $t \to \infty$ and $\lim_{t \to \infty} \min_m |\mathcal{M}_{m,t}| = \infty$ is implied given $\liminf_{t \to \infty} \min_m \rho_t^{(m)} > 0$, we must obtain $\lim_{t \to \infty} \mathbb{E}[||\mathbf{b}_{m,t}^* - \mathbf{b}_{m,t+1}^*||_2] = 0$ almost surely for

[18] Such a transformation does not impact the continuity of ϕ.

all communities m. Applying Markov's inequality guarantees that in fact $||\mathbf{b}^*_{m,t} - \mathbf{b}^*_{m,t+1}||_2$ converges almost surely to 0, which implies that $\lambda_{\mathcal{L}}\left(\mathcal{A}^{(m)}_t \backslash \mathcal{A}^{(m)}_{t+1} \cup \mathcal{A}^{(m)}_{t+1} \backslash \mathcal{A}^{(m)}_t\right) \overset{\text{a.s.}}{\to} 0$, where $\lambda_{\mathcal{L}}$ is the Lebesgue measure. Consequently, for

$$\bar{\mathbf{b}}_{m,t-1} = \left(\int_{\mathbf{b}\in\mathcal{A}^{(m)}_{t-1}} h(\mathbf{b})\, d\mathbf{b}\right)^{-1} \int_{\mathbf{b}\in\mathcal{A}^{(m)}_{t-1}} \mathbf{b}\, h(\mathbf{b})\, d\mathbf{b}$$

the stochastic process $\mathcal{Z}_{m,t} = ||\mathbf{b}^*_{m,t} - \bar{\mathbf{b}}_{m,t-1}||_2$, with respect to the obvious filtration, is eventually a supermartingale. Via the martingale convergence theorem, $\mathbf{b}^*_{m,t} \overset{\text{a.s.}}{\to} \mathbf{b}^*_{m,\infty}$ for all m, for some set of barycenters $(\mathbf{b}^*_{1,\infty}, \ldots, \mathbf{b}^*_{M,\infty})$ (which are constant but not necessarily deterministic).

By way of contradiction, suppose that for some community m, $\mathbf{b}^*_{m,t}$ does not converge to a fixed point of ϕ, so in particular, $\bar{\mathbf{b}}^*_m \equiv \left(\int_{\mathbf{b}\in\mathcal{A}^{(m)}_*} h(\mathbf{b})\, d\mathbf{b}\right)^{-1} \int_{\mathbf{b}\in\mathcal{A}^{(m)}} \mathbf{b}\, h(\mathbf{b})\, d\mathbf{b} \neq \mathbf{b}^*_{m,\infty}$ where $\mathcal{A}^{(m)}_* = V^{(m)}(\mathbf{b}^*_{1,\infty}, \ldots, \mathbf{b}^*_{M,\infty})$ (the map defined in the proof of Proposition 1). For $\gamma \equiv \liminf_{t\to\infty} \min_m \rho^{(m)}_t > 0$, we have that $\mathbb{E}[||\bar{\mathbf{b}}^*_m - \mathbf{b}^*_{m,t}||_2] \geq \gamma \cdot ||\bar{\mathbf{b}}^*_m - \mathbf{b}^*_{m,\infty}||_2/t$. In particular, $\sum_{t=1}^{\infty} \mathbb{E}[||\bar{\mathbf{b}}^*_m - \mathbf{b}^*_{m,t}||_2] = \infty$, and because $\mathbb{E}[\bar{\mathbf{b}}^*_m - \mathbf{b}^*_{m,t}]$ is a ray pointing in the same direction as $\bar{\mathbf{b}}^*_m - \mathbf{b}^*_{m,\infty}$ (which does not depend on t), we have that $\lim_{t\to\infty} \mathbf{b}^*_{m,t} \notin [0,1]^d$, a contradiction. Thus, we must have $\bar{\mathbf{b}}^*_m = \mathbf{b}^*_{m,\infty}$ for all m. ∎

Proof of Proposition 2. Observe that when $d = 1$, an isomorphic way of expressing the Voronoi diagram is via a sequence of cutoffs $(\alpha_1, \alpha_2, \ldots, \alpha_{M-1})$ where community m consists of the cell $\{b : \alpha_{m-1} \leq b \leq \alpha_m\}$ with the convention that $\alpha_0 = 0$ and $\alpha_M = 1$. The corresponding barycenters are then $b^*_{m,\infty} \equiv \left(\int_{\alpha_{m-1}}^{\alpha_m} h(b)\, db\right)^{-1} \int_{\alpha_{m-1}}^{\alpha_m} b\, h(b)\, db$ with population masses $\int_{\alpha_{m-1}}^{\alpha_m} h(b)\, db$. This translates into an isomorphic map $\tilde{\phi}$, which instead of mapping barycenters to barycenters, maps the cutoff partition of $[0,1]$. Applying Proposition 1, we know there exists at least one fixed point of the map $\tilde{\phi}$. We will use Kellogg (1976) to show that in fact this fixed point is unique, which in conjunction with Theorem 1, establishes the claim.

To apply Kellogg (1976), we need to prove the map $\tilde{\phi}$ is differentiable, its Jacobian $\nabla\tilde{\phi}$ has no eigenvalue of 1, and there are no fixed points on the boundary of $[0,1]^{m-1}$ (recall that $m-1$ cutoffs unambiguously pin down the one-dimensional Voronoi diagram under $\tilde{\phi}$). The proof of Lemma A.1 shows that no barycenter of a steady-state Voronoi diagram can lie on the boundary, which by the assumption that the density h is lower bounded on $[0,1]$, implies that none of the cutoffs $(\alpha_1, \ldots, \alpha_{M-1})$ will be equal to 0 or 1. At the same time, Lemma A.1 guarantees $\alpha_m \neq \alpha_{m+1}$ for any m, so $\int_{\alpha_{m-1}}^{\alpha^m} h(b)\, db > 0$, and taking ε in the map of ϕ from Proposition 1 to be sufficiently small, we guarantee that $\nabla\tilde{\phi}$ is well-defined (i.e., $\tilde{\phi}$ is differentiable). Notice that the cutoffs in the limiting fixed point of $\tilde{\phi}$ must satisfy:

$$\begin{aligned} &\alpha_m - \mathbb{E}_H[b_t \,|\, \alpha_{m-1} \leq b_t \leq \alpha_m] = \mathbb{E}_H[b_t \,|\, \alpha_m \leq b \leq \alpha_{m+1}] - \alpha_m \\ \implies \quad &2\alpha_m = \mathbb{E}_H[b_t \,|\, \alpha_{m-1} \leq b_t \leq \alpha_m] + \mathbb{E}_H[b_t \,|\, \alpha_m \leq b \leq \alpha_{m+1}]\,, \end{aligned}$$

for all $1 \leq m \leq M-1$. However, notice the derivative with respect to α_m on the LHS is exactly 2, whereas the derivative with the respect to the RHS is strictly less than 2 given than h has lower bounded support over all of $[0,1]$. Thus, the eigenvalues of $\nabla\tilde{\phi}$ lie strictly within the unit circle, and

Brouwer admits a unique fixed point via the sole theorem in Kellogg (1976). ∎

A.3 Proofs from Section 5

Proof of Theorem 2. First, we prove the theorem in the case that $c = 0$, and then generalize to the case where $c \geq 0$. Note that a ban policy results in a new community $\hat{m}'$ that supplants both $\hat{m}$ and $\tilde{m}$, and has an average sentiment

$$\mathbf{b}^{(\hat{m}')} = \frac{\rho^{(\tilde{m})}\mathbf{b}^{(\tilde{m})} + \rho^{(\hat{m})}\mathbf{b}^{(\hat{m})}}{\rho^{(\tilde{m})} + \rho^{(\hat{m})}}.$$

A quarantine policy, on the other hand, leaves both community $\tilde{m}$ and $\hat{m}$ as before, but with potentially different beliefs and populations. The population of community $\tilde{m}$ drops to $(1-\phi)\rho^{(\tilde{m})}$, while still holding belief $\mathbf{b}^{(\tilde{m})}$, whereas the population of community $\hat{m}$ increases to $\rho^{(\hat{m})} + \phi\rho^{(\tilde{m})}$, while holding the new belief $(\rho^{(\hat{m})}\mathbf{b}^{(\hat{m})} + \phi\rho^{(\tilde{m})}\mathbf{b}^{(\tilde{m})})/(\rho^{(\hat{m})} + \phi\rho^{(\tilde{m})})$.

Observe that the community beliefs of community $\hat{m}'$ in the case of a ban and community $\hat{m}$ in the case of a quarantine lie on the ray between $\mathbf{b}^{(\tilde{m})}$ and $\mathbf{b}^{(\hat{m})}$. Because there is a single problematic community $\hat{m}$, there are two possible scenarios (i.e., scenarios (iii) and (iv) from Lemma A.2) that can exist for how communities $(\tilde{m}, \hat{m})$ relate to the acceptable region $\mathcal{R}$. In both cases, there is a cutoff γ^* such that belief $\gamma\mathbf{b}^{(\tilde{m})} + (1-\gamma)\mathbf{b}^{(\hat{m})} \in \mathcal{R}$ if and only if $\gamma < \gamma^*$. With some algebraic rearrangement, one can show that the ban yields a $\gamma_b = \frac{\rho^{(\tilde{m})}/\rho^{(\hat{m})}}{1+\rho^{(\tilde{m})}/\rho^{(\hat{m})}}$ and the quarantine yields a $\gamma_q = \frac{\phi\rho^{(\tilde{m})}/\rho^{(\hat{m})}}{1+\phi\rho^{(\tilde{m})}/\rho^{(\hat{m})}}$ for the beliefs of merged community $\hat{m}'$ and adapted community $\hat{m}$ under ban and quarantine policies, respectively.

Note that because $\phi < 1$, $\gamma_q < \gamma_b$. If $\gamma_b < \gamma^*$, the ban policy is optimal because it reduces the costly offline action to $\mu = 0$, whereas the quarantine policy and no policy interventions leave $\mu > 0$. This occurs whenever $\rho^{(\tilde{m})}/\rho^{(\hat{m})} < \gamma^*/(1-\gamma^*) \equiv \rho_1$. If $\gamma_b > \gamma^*$, banning community $\tilde{m}$ is dominated by no policy, because in this case the costly offline participation increases from $\mu = \rho^{(\tilde{m})}$ to $\mu' = \rho^{(\tilde{m})} + \rho^{(\hat{m})}$. If $\gamma_q < \gamma^*$, the quarantine policy dominates no policy because it reduces costly offline participation from $\mu = \rho^{(\tilde{m})}$ to $\mu' = (1-\phi)\rho^{(\tilde{m})}$, and also dominates the ban policy whenever $\gamma_q < \gamma^* < \gamma_b$ because it reduces offline costly action instead of increasing it. This happens exactly whenever $\rho_1 < \rho^{(\tilde{m})}/\rho^{(\hat{m})} < \gamma^*/(\phi(1-\gamma^*)) \equiv \rho_2$. Finally, whenever $\gamma_b > \gamma^*$ and $\gamma_q > \gamma^*$ (i.e., when $\rho^{(\tilde{m})}/\rho^{(\hat{m})} > \rho_2$), the quarantine policy and ban policy both increase offline costly participation to $\mu' = \rho^{(\tilde{m})} + \rho^{(\hat{m})} > \rho^{(\tilde{m})} = \mu$, and no policy is the most effective.

Finally, we can easily generalize to the case of $c \geq 0$ by comparing to the reduction in costly offline action conditional on each policy being optimal, $\Delta_b = \mathcal{C}(\rho^{(\tilde{m})})$ and $\Delta_q = \mathcal{C}(\rho^{(\tilde{m})}) - \mathcal{C}((1-\phi)\rho^{(\tilde{m})}) < \Delta_b$. If $\Delta_b > \Delta_q > c$, the result stands as is. If $\Delta_q < \Delta_b < c$, the result is trivial by setting $\rho_1 = \rho_2 = 0$. If $\Delta_q < c < \Delta_b$, the result holds by amending ρ_2 to be equal to ρ_1. ∎

Proof of Proposition 3. Decreasing the size of the acceptable region to $\mathcal{R}' \subset \mathcal{R}$ while retaining the property that there is only one problematic community is equivalent to decreasing γ^* in region $\mathcal{R}$ (from the proof of Theorem 2, denoted $\gamma^*_{\mathcal{R}}$) to some $\gamma^*_{\mathcal{R}'} < \gamma^*_{\mathcal{R}}$. If the optimal intervention is to ban community $\tilde{m}$, the claim holds vacuously because shrinking $\mathcal{R}$ cannot make the intervention stronger. If the optimal intervention to quarantine community $\tilde{m}$, then $\gamma_q < \gamma^*_{\mathcal{R}} < \gamma_b$, with $\gamma_q < \gamma^*_{\mathcal{R}'} < \gamma_b$ (which means the quarantine intervention is still optimal) or $\gamma^*_{\mathcal{R}'} < \gamma_q$ (which means no intervention

is optimal, which is a weaker intervention). Finally, if $\gamma^*_{\mathcal{R}'} < \gamma^*_{\mathcal{R}} < \gamma_q$, then no intervention is always optimal in both regimes. ∎

Proof of Proposition 4. Let us consider the map ϕ from Proposition 1 restricted to the acceptable region space of $\mathcal{R}^m$ for all communities $m = 1, \ldots, M$ (i.e., defined as $\phi_{\mathcal{R}} : \mathcal{R}^m \to \mathcal{R}^m$ where $\mathcal{R} \subset [0,1]^d$). Notice that the same conditions of Brouwer's fixed point theorem apply to the mapping $\phi_{\mathcal{R}}$ (given that $\mathcal{R}$ is also compact and convex), so there should exist a fixed point of barycenters $(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(M)})$ all within $\mathcal{R}$, satisfying $\left(\int_{\mathbf{b}\in\mathcal{A}^{(m)}} h(\mathbf{b})\, d\mathbf{b}\right)^{-1} \int_{\mathbf{b}\in\mathcal{A}^{(m)}} \mathbf{b}\, h(\mathbf{b})\, d\mathbf{b} = \mathbf{b}^{(m)}$ (and with corresponding Voronoi cells $(\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(M)})$), given that the current set of barycenters lies in $\mathcal{R}$ by assumption. At the same time, this would suggest there is a fixed point of the standard mapping $\phi : [0,1]^d \to [0,1]^d$ under Proposition 1, and this must be a steady state under the standard conditions of Definition 1. However, this leads to a contradiction because we know $\bar{\mathbf{b}}^{(\hat{m})} \notin \mathcal{R}$, invalidating that this can exist as a steady state (and a fixed point under ϕ). For the second claim of the Proposition, note that if $\mathbf{b}^{(\hat{m})} \to \bar{\mathbf{b}}^{(\hat{m})}$ intersects the boundary $\partial\mathcal{R}$ of the acceptable region $\mathcal{R}$, then by Lemma A.2, it must be that $\bar{\mathbf{b}}^{(\hat{m})} \notin \mathcal{R}$, which is an equivalent condition that guarantees the intervention is not robust. ∎

Proof of Proposition 5. The proof consists of two parts. First, by single-peakedness of H, we show that the community $\hat{m}$ adjacent to $\tilde{m}$ is also the only community whose adjacent community can be the problematic community $\tilde{m}$ (i.e., there exists no other community m' with $\tilde{m}$ as its adjacent community, besides potentially $\hat{m}$). Second, we show that if there is an optimal ban (resp. quarantine) policy, there is a simple optimal ban (resp. quarantine) policy, given that with the possible exception of $\hat{m}$, no community is adjacent to $\tilde{m}$.

First, for some $\bar{m} \neq \hat{m}$, suppose that $\bar{m}$ is the adjacent community to $\tilde{m}$. Notice then that the maximal radius ball B_r contained in $\mathcal{A}^{(\bar{m})}$ is tangent to the shared edge of $\mathcal{A}^{(\tilde{m})}$ and $\mathcal{A}^{(\hat{m})}$, but is not tangent to any other shared edges. Call this tangent point $\mathbf{y}$; observe that along the ray $\mathbf{b}^{(\bar{m})} \to \mathbf{y}$, the density h must be monotonically decreasing. To see this, note that using Lemma A.3, no Voronoi cell has a Lebesgue measure greater than $\underline{\gamma}/M$. Thus, for sufficiently large $\bar{M}$ (and $M > \bar{M}$), the peak of H along the line $\mathbf{b}^{(\bar{m})} \leftrightarrow \mathbf{y}$ must precede $\mathbf{b}^{(\bar{m})}$. Consequently, the conditional mean of $\mathbf{b}_t$ given that it lies in B_r and along the line $\mathbf{b}^{(\bar{m})} \leftrightarrow \mathbf{y}$ must precede $\mathbf{b}^{(\bar{m})}$. Moreover, because B_r is tangent at $\mathbf{y}$, the line $\mathbf{b}^{(\bar{m})} \leftrightarrow \mathbf{y}$ does not extend past $\mathbf{y}$ conditional on being in $\mathcal{A}^{(\bar{m})}$, but may extend outside of B_r on the side that precedes $\mathbf{b}^{(\bar{m})}$. As a result, the conditional mean $\mathbb{E}[\mathbf{b}_t \,|\, \mathbf{b}_t \in \mathcal{A}^{(\bar{m})} \cap \mathbf{b}^{(\bar{m})} \leftrightarrow \mathbf{y}]$ must precede $\mathbf{b}^{(\bar{m})}$, which we can denote as $\mathbf{x}$. However, given that $\tilde{m}$ is not adjacent to $\bar{m}$, we know that $\mathbb{E}[\mathbf{b}_t \,|\, \mathbf{b}_t \in \mathcal{A}^{(\bar{m})}]$ must lie in the open half-space containing $\mathbf{x}$ and with $\mathbf{b}^{(\bar{m})}$ as a limit point, which is a contradiction that $\mathbf{b}^{(\bar{m})}$ is the barycenter of $\mathcal{A}^{(\bar{m})}$. Hence, any such candidate community $\bar{m}$ must be adjacent to the problematic community $\tilde{m}$, and in fact be $\hat{m}$.

Second, suppose the optimal intervention is a ban policy. Given the previous intermediate result, there are only two candidates for communities to ban (as banning any other community leads to no change in the costly offline participation μ by convexity of $\mathcal{R}$):

1. The problematic community $\tilde{m}$: If the problematic community $\tilde{m}$ is optimal to ban, this is a simple intervention, establishing the claim.

2. The adjacent community $\hat{m}$: There are two subcases if the adjacent community $\hat{m}$ is banned. First is that $\tilde{m}$ is also adjacent to $\hat{m}$, which leads to the same outcome (in terms of costly participation μ) as if the simple intervention of just banning the problematic community $\tilde{m}$ had been enacted. The second is that $\tilde{m}$ is adjacent to some other community ℓ, with $\mathbf{b}^{(\ell)} \in \mathcal{R}$ as its barycenter (by assumption that $\tilde{m}$ is the unique problematic community). By convexity of $\mathcal{R}$, there is no impact on μ from this intervention.

Suppose the optimal intervention is a quarantine policy. Once again, given the previous intermediate result, there are only two candidates for communities to quarantine:

1. The problematic community $\tilde{m}$: If the problematic community is optimal to quarantine, this is a simple intervention, establishing the claim.

2. The adjacent community $\hat{m}$: There are two subcases if the adjacent community is quarantined. First is that $\tilde{m}$ is also adjacent to $\hat{m}$. A quarantine at $\tilde{m}$ is only potentially optimal if it reduces costly offline action beyond no policy, which implies that $(\mathbf{b}^{(\tilde{m})}\rho^{(\tilde{m})} + \phi\mathbf{b}^{(\hat{m})}\rho^{(\hat{m})})/(\rho^{(\tilde{m})} + \phi\rho^{(\hat{m})}) \in \mathcal{R}$. By Lemma A.2, however, we know that $(\mathbf{b}^{(\tilde{m})}\rho^{(\tilde{m})} + \mathbf{b}^{(\hat{m})}\rho^{(\hat{m})})/(\rho^{(\tilde{m})} + \rho^{(\hat{m})}) \in \mathcal{R}$ given that $\mathbf{b}^{(\hat{m})} \in \mathcal{R}$. But then, of course, a ban policy on community $\tilde{m}$ is also optimal and simple. The second case is that $\tilde{m}$ is adjacent to some other community ℓ, with $\mathbf{b}^{(\ell)} \in \mathcal{R}$ as its barycenter (by assumption that $\tilde{m}$ is the unique problematic community). By convexity of $\mathcal{R}$, there is no impact on μ from this intervention. ∎

A.4 Proofs from Section 6

Proof of Proposition 6. First, notice that when $c_b = c_q = 0$, then the standard intervention of Theorem 2 depends only on the ratio $\rho^{(\tilde{m})}/\rho^{(\hat{m})}$ and not on $\mathcal{C}$, because $\Delta_b > \Delta_q > 0$. Consequently, when the Voronoi diagram reaches steady state (via Theorem 1) and the platform takes the optimal short-term intervention, the offline costly participation will either be $\mu = 0$ (optimal intervention is a ban by Theorem 2(i)), $\mu = (1-\phi)\rho^{(\tilde{m})}$ (optimal intervention is a quarantine by Theorem 2(ii)), or $\mu = \rho^{(\tilde{m})}$ (optimal intervention is to do nothing by Theorem 2(iii)). This will be true regardless of the platform's cost function (i.e., whether it is $\bar{\mathcal{C}}$ or $\underline{\mathcal{C}}$). Hence, there is some random variable μ^* that determines the platform's costly offline participation, should she choose to forgo a preemptive intervention and possibly intervene at a later date $T \gg t$ after establishing steady state.

Let us denote by σ_{SS} the probability distribution over all steady states conditional on the Voronoi diagram at some time $T \gg t$, following an intervention at time t (which is, again, guaranteed to exist by Theorem 1 almost surely). Notice that for any fixed steady-state Voronoi diagram V following an intervention at t, there is some costly offline participation, call it μ_V, which obviously has no dependence on the platform's cost function. Let us consider the expression $(\bar{\mathcal{C}}(\mu_V) - \bar{\mathcal{C}}(\mu^*)) - (\underline{\mathcal{C}}(\mu_V) - \underline{\mathcal{C}}(\mu^*)) \geq 0$ for all μ_V, μ^*, given that $\bar{\mathcal{C}}' \geq \underline{\mathcal{C}}'$ pointwise, by assumption. Applying the expectation operator shows that $\mathbb{E}_{\sigma_{SS},\mu^*}[\bar{\mathcal{C}}(\mu_V) - \bar{\mathcal{C}}(\mu^*)] \geq \mathbb{E}_{\sigma_{SS},\mu^*}[\underline{\mathcal{C}}(\mu_V) - \underline{\mathcal{C}}(\mu^*)]$. A preemptive intervention is only optimal under $\underline{\mathcal{C}}$ if $\mathbb{E}_{\sigma_{SS},\mu^*}[\underline{\mathcal{C}}(\mu_V) - \underline{\mathcal{C}}(\mu^*)] \geq 0$, which would immediately imply that $\mathbb{E}_{\sigma_{SS},\mu^*}[\bar{\mathcal{C}}(\mu_V) - \bar{\mathcal{C}}(\mu^*)] \geq 0$, and so the preemptive intervention is also optimal under $\bar{\mathcal{C}}$. ∎

Proof of Proposition 7. First, notice that the number of communities can only increase because users who commit to starting their own community never leave. This implies that the number of currently active communities at time t, M_t, is monotonically increasing over time, and either diverges or converges to some $M^* < \infty$. We establish the claim by showing that, almost surely, the number of communities cannot diverge, i.e., $\lim_{t\to\infty} M_t \neq \infty$.

We will derive a contradiction if the population of all communities is upper bounded with positive probability (i.e., there exists $n > 0$ such that $\limsup_{t\to\infty} \max_m |\mathcal{M}_{m,t}| \leq n$ with probability $\delta > 0$). Pick an arbitrary community m^* (say, the community started by the first agent). Notice that by Lemma A.3, the Lebesgue measure of cell m^* is lower bounded by $\underline{\gamma}/M$, which in turn is lower bounded by $\underline{\gamma}/t$. Given that density h is lower bounded by density $\underline{\mu}$ and upper bounded by density $\bar{\mu}$, one can inscribe a ball B of volume $(\underline{\mu}\cdot\underline{\gamma})/(\bar{\mu}\cdot t)$ inside cell $\mathcal{A}^{(m^*)}$ around m^*'s barycenter, $\mathbf{b}^{(m^*)}$. Because the population of all communities is upper bounded (so $\varphi(\cdot, D)$ is upper bounded for any fixed $D > 0$) and φ is continuous with $\lim_{D\to 0}\varphi(\cdot, D) = \infty$ (by assumption), there exists a smaller ball $B' \subset B \subset \mathcal{A}^{(m^*)}$ of volume $C\cdot(\underline{\mu}\cdot\underline{\gamma})/(\bar{\mu}\cdot t)$ for $C < 1$ such that all agents born with $\mathbf{b}_t \in B'$ join community m^*, and ex-ante, this occurs with at least probability $(C\underline{\mu}^2\cdot\underline{\gamma})/(\bar{\mu}\cdot t)$, and naturally it is assumed agents are drawn independently across time. But of course, $\sum_{t=1}^{\infty}(C\underline{\mu}^2\cdot\underline{\gamma})/(\bar{\mu}\cdot t) = \infty$, so by the Borel-Cantelli lemma, the probability that an infinite number of incoming agents join community m^* is 1. This is a contradiction, so the population of community m^* must grow infinitely large with probability 1.

Finally, recall that $\lim_{k\to\infty}\varphi(k,\cdot) = \infty$, $\varphi(1,\cdot)$ is upper bounded, and by our previous intermediate result, we know that at least one community's population tends toward infinity almost surely. Hence, almost surely, eventually (for some $t' > 0$) joining community m^* leads to a higher $U_{m,t'}$ than starting one's own community for all $t \geq t'$, regardless of the agent's incoming belief $\mathbf{b}_{t'}$. Thus, for all $t' \geq t$, $M_{t'} = M_{t'+1} = M^*$, as the claim establishes. ∎

Proof of Proposition 8. We will first show that there exists a polynomial-time algorithm to solve WORST. Note that it is without loss of generality to assume $\boldsymbol{\beta}$ is known with certainty, because one can simply solve $\min \boldsymbol{\beta}$ subject to $\mathbf{C}\boldsymbol{\beta} \leq \boldsymbol{\alpha}$, which only depends on the constraints k and not on the number of communities M. Thus, we will focus solely on uncertainty with respect to $\mathbf{A}$, i.e., that $\mathcal{U}_A = \{\mathbf{A} : \mathbf{D}_i[\mathbf{A}]_i \leq \mathbf{f}_i\,\forall\, i\}$, given no uncertainty about $\boldsymbol{\beta}$. Note that our original optimization problem can be rewritten as:

$$\begin{aligned}
\min_{\chi_q,\,\chi_b}\ & \mathcal{C}\left(1 - \sum_{m=1}^{M} \mathbf{1}_{\mathbf{A}\mathbf{b}^{(m)} \leq \boldsymbol{\beta}}\cdot\rho^{(m)}\right) - \sum_{m=1}^{M}\left(c\cdot\mathbf{1}_{m\in\chi_q\cup\chi_b}\right) \\
\text{s.t.}\ & \max_{[\mathbf{A}]_i}[\mathbf{A}]_i\mathbf{b} \leq \boldsymbol{\beta}_i \\
& \mathbf{D}_i[\mathbf{A}]_i \leq \mathbf{f}_i\ \forall\ i
\end{aligned}$$

We can convert the inner optimization problem (in the constraints) to its dual problem which will give

us a pair of minimization problems. This dual problem is given by

$$\begin{aligned}
&\min_{\mathbf{q}_i} \ \mathbf{q}_i^T \mathbf{f}_i \\
&\text{s.t.} \ \ \mathbf{D}_i^T \mathbf{q}_i = \mathbf{b}, \ \forall\, i \\
&\qquad \mathbf{q}_i^T \mathbf{f}_i \leq \boldsymbol{\beta}_i \ \forall\, i \\
&\qquad \mathbf{q}_i \geq 0, \ \forall\, i
\end{aligned}$$

By strong duality, both problems have the same optimal solution so we can replace the original problem with

$$\begin{aligned}
&\min_{\chi_q, \chi_b} \ \mathcal{C}\left(1 - \sum_{m=1}^{M} \mathbf{1}_{\mathbf{A}\mathbf{b}^{(m)} \leq \boldsymbol{\beta}} \cdot \rho^{(m)}\right) - \sum_{m=1}^{M} \left(c \cdot \mathbf{1}_{m \in \chi_q \cup \chi_b}\right) \\
&\text{s.t.} \ \ \min_{\mathbf{q}_i} \ \mathbf{q}_i^T \mathbf{f}_i \\
&\qquad \mathbf{D}_i^T \mathbf{q}_i = \mathbf{b}, \ \forall\, i \\
&\qquad \mathbf{q}_i^T \mathbf{f}_i \leq \boldsymbol{\beta}_i \ \forall\, i \\
&\qquad \mathbf{q}_i \geq 0, \ \forall\, i
\end{aligned}$$

But naturally, this can be simply rewritten as

$$\begin{aligned}
&\min_{\chi_q, \chi_b, \mathbf{q}_i} \ \mathcal{C}\left(1 - \sum_{m=1}^{M} \mathbf{1}_{\mathbf{A}\mathbf{b}^{(m)} \leq \boldsymbol{\beta}} \cdot \rho^{(m)}\right) - \sum_{m=1}^{M} \left(c \cdot \mathbf{1}_{m \in \chi_q \cup \chi_b}\right) \\
&\text{s.t.} \ \ \mathbf{q}_i^T \mathbf{f}_i \leq \boldsymbol{\beta}_i \ \forall\, i \\
&\qquad \mathbf{D}_i^T \mathbf{q}_i = \mathbf{b}, \ \forall\, i \\
&\qquad \mathbf{q}_i \geq 0, \ \forall\, i
\end{aligned} \tag{1}$$

which is now a deterministic optimization problem under no uncertainty.

We will formulate a dynamic program that can solve this in polynomial time with respect to the number of communities, M. We focus solely on quarantine policies without loss of generality – if ban is permitted, the same algorithm would work by doubling the communities from M to $2M$ and matching each community with its counterpart in the larger space, but where a quarantine policy can only be used for communities $1 \leq m \leq M$ and a ban policy can only be used for communities $M+1 \leq m \leq 2M$, and where interventions in communities that are M distance apart are disallowed. For this version, we build an $M \times M$ grid and proceed by backward induction. In each square of this grid, we keep the following metrics:

(i) The current subset of communities χ_q with quarantine interventions.

(ii) The Voronoi diagram V associated with such policies, which is computable in polynomial time (see Fortune (1986)).

(iii) The offline costly participation μ associated with V as well its cost $\mathcal{C}(\mu)$.

Starting in the final column of this grid, we consider quarantine policies that only affect a single community. Each row in this column of the grid corresponds to a community that is quarantined, and where no action is taken on any other community. For any grid square, it requires polynomial time to compute all of the metrics conditional on each of the communities (out of M) being quarantined and no others. Also note that it takes polynomial time to confirm whether the constraints of (1) are satisfied, and one can assign a costly action of $+\infty$ for all grid squares where not these are not satisfied. Thus, conditional on only being able to intervene at one community, it requires polynomial time to find the optimal intervention and the results of all possible quarantine interventions can be stored in the final column of the grid.

Given the platform can conduct multiple simultaneous interventions, we note that the optimal short-term intervention can be solved via backward induction using the grid. At every column ℓ from the final column of the grid (with column $\ell = 1$ being the last column), for each community m potentially being quarantined, the platform can compute the Voronoi diagram $V_{m,\ell}$ associated with least cost under $\mathcal{C}$ given that community m is quarantined and that there are $\ell - 1$ other (optimally) communities quarantined with their Voronoi diagrams given by the grid square in column $\ell - 1$, and conditional on the constraints in (1) being satisfied. The optimal solution is then a grid search for the least cost, with the optimal intervention being the χ_q that involves some number ℓ of quarantine interventions.

Notice that if we can solve WORST in polynomial time, it immediately implies we can solve EXP and BEST in polynomial time. For EXP, this can be seen by introducing no uncertainty in $\mathcal{R}$, i.e., by setting

$$\mathbf{D}_i = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & 0 \\ 0 & \cdots & 0 & -1 \end{pmatrix}$$

and $\mathbf{f}_i = [\mathbf{A}]_i$ for all i. Similarly, for BEST, it is easy to simply add the constraints $\mathbf{D}_i[\mathbf{A}]_i \leq \mathbf{f}_i$ (for all i) and $\mathbf{C}\boldsymbol{\beta} \leq \boldsymbol{\alpha}$ to the existing constraints of the original problem. This is the standard min-min problem which is of the form in the algorithm for WORST, and the additional constraints can be validated (or invalidated) in polynomial time. Thus, nearly the same algorithm as WORST can be applied identically to solve EXP and BEST also in polynomial time. ∎

B Supplementary Material

In this Appendix, we provide partial context for a reader about social media platforms, which might supply a more direct mapping between the model and the current social media landscape. The social media platform, Reddit, serves as the quintessential example of the social media platform we describe in this paper. Reddit is known for being segregated into "subreddits," thought of as the communities of our model. Many of the more popular communities, some of which are innocuous and others that include more dangerous discussions, are seen in Figure 15.

Other social media platforms, including Facebook (now "Meta") and Twitter, also have community-

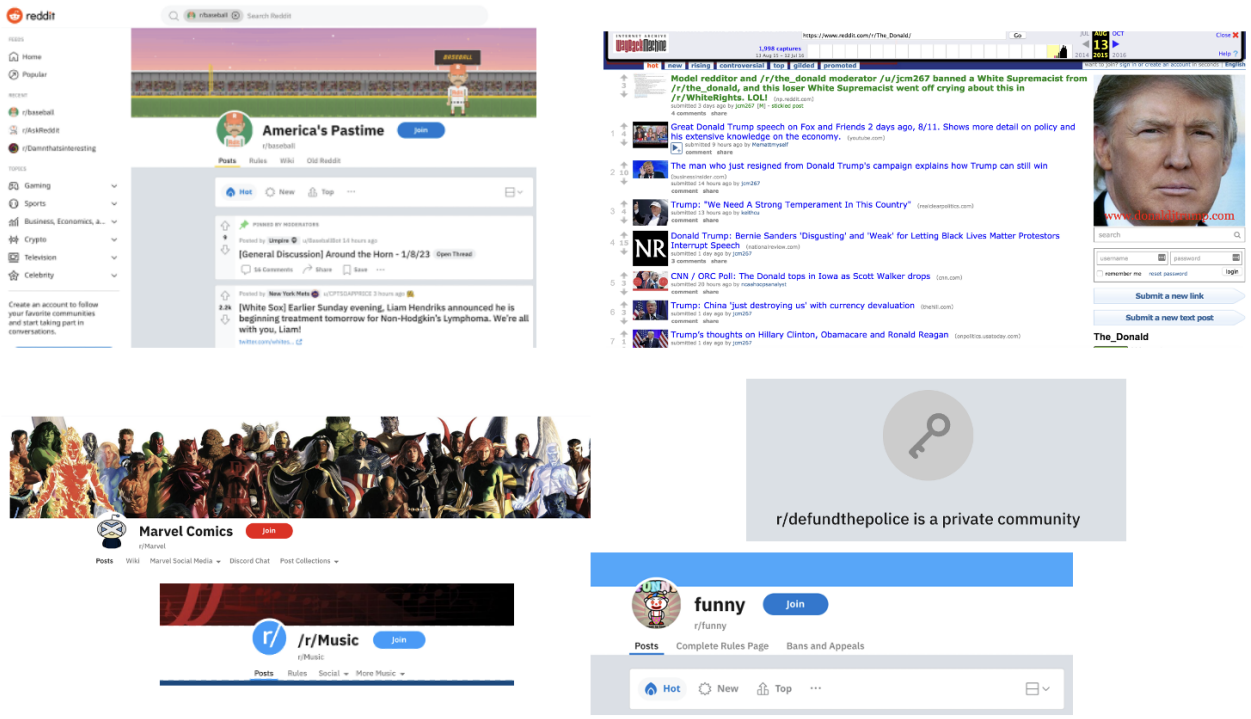
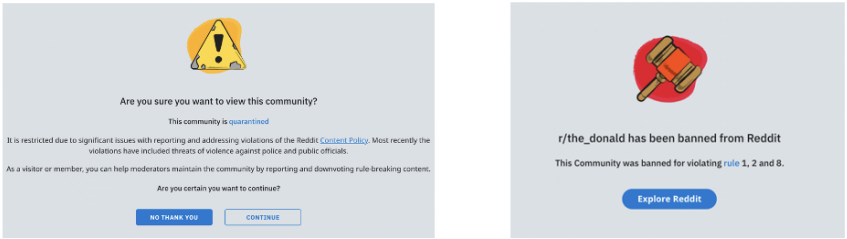


Figure 15. Examples of Reddit Communities.

based features on their platforms. For the former, these communities are often classified as "pages" and for the latter, they are often categorized based on their "hashtags" (e.g., dangerous Twitter communities around ideas such as #PizzaGate). As a result, the results of our model appeal broadly to many social media platforms.

Finally, we note that the typical interventions we consider in our model are motivated by the current content moderation strategies that exist on platforms like Reddit. For example, our main motivation for the *strong* and *mild* interventions, are Reddit's ban and quarantine policies enacted on the community r/The_Donald in 2019 and 2020 (see Figure 16).



Quarantined June 26, 2019 Banned June 29, 2020

Figure 16. Reddit Interventions for r/The_Donald.

References

Acemoglu, Daron, Tarek A Hassan, and Ahmed Tahoun (2018), "The power of the street: Evidence from egypt's arab spring." *The Review of Financial Studies*, 31, 1–42.

Acemoglu, Daron, Asuman Ozdaglar, and Ali ParandehGheibi (2010), "Spread of (mis)information in social networks." *Games and Economic Behavior*, 70, 194–227.

Agarwal, Saharsh, Uttara M Ananthakrishnan, and Catherine E Tucker (2022), "Deplatforming and the control of misinformation: Evidence from parler." *Available at SSRN*.

Ali, Shiza, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini (2021), "Understanding the effect of deplatforming on social networks." In *13th ACM Web Science Conference 2021*, 187–195.

Blondel, Vincent D, Julien M Hendrickx, and John N Tsitsiklis (2009), "On krause's multi-agent consensus model with state-dependent connectivity." *IEEE transactions on Automatic Control*, 54, 2586–2597.

Cohn, Henry and Noam Elkies (2003), "New upper bounds on sphere packings." *Annals of Mathematics*, 689–714.

Conway, John Horton and Neil James Alexander Sloane (2013), *Sphere packings, lattices and groups*, volume 290. Springer Science & Business Media.

Durkee, Alison (2022), "Most republicans believe midterms were 'free and fair,' poll finds as fraud fears fall flat."

Fisher, Marc, John Woodrow Cox, and Peter Hermann (2016), "Pizzagate: From rumor, to hashtag, to gunfire in dc." *Washington Post*.

Fortune, Steven (1986), "A sweepline algorithm for voronoi diagrams." In *Proceedings of the second annual symposium on Computational geometry*, 313–322.

Frenkel, Sheera (2021), "The storming of capitol hill was organized on social media." *The New York Times*, 6, 2021.

Golub, Benjamin and Matthew O Jackson (2010), "Naive learning in social networks and the wisdom of crowds." *American Economic Journal: Microeconomics*, 2, 112–49.

Golub, Benjamin and Matthew O Jackson (2012), "How homophily affects the speed of learning and best-response dynamics." *The Quarterly Journal of Economics*, 127, 1287–1338.

Habib, Hussam, Maaz Bin Musa, Fareed Zaffar, and Rishab Nithyanand (2019), "To act or react: Investigating proactive strategies for online community moderation." *arXiv preprint arXiv:1906.11932*.

Hegselmann, Rainer, Ulrich Krause, et al. (2002), "Opinion dynamics and bounded confidence models, analysis, and simulation." *Journal of artificial societies and social simulation*, 5.

Hui, Achille (2023), "Largest ball guaranteed to fit in a bounded polyhedron of volume v." Mathematics Stack Exchange, URL https://math.stackexchange.com/q/4612559.

Hwang, Elina H and Stephanie Lee (2021), "A nudge to credible information as a countermeasure to misinformation: Evidence from twitter." *Available at SSRN*.

Jhaver, Shagun, Christian Boylston, Diyi Yang, and Amy Bruckman (2021), "Evaluating the effectiveness of deplatforming as a moderation strategy on twitter." *Proceedings of the ACM on Human-Computer Interaction*, 5, 1–30.

Kellogg, R Bruce (1976), "Uniqueness in the schauder fixed point theorem." *Proceedings of the American Mathematical Society*, 60, 207–210.

Marsden, Peter V (1987), "Core discussion networks of americans." *American sociological review*, 122–131.

McEvoy, Jemima (2021), "Capitol attack was planned openly online for weeks—police still weren't ready."

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G. Rand (2021), "Shared partisanship dramatically increases social tie formation in a twitter field experiment." *Proceedings of the National Academy of Sciences*, 118.

Mostagir, Mohamed, Asu Ozdaglar, and James Siderius (2022), "When is society susceptible to manipulation?" *Management Science, Forthcoming*.

Mostagir, Mohamed and James Siderius (2022a), "Naive and bayesian learning with misinformation policies." Technical report, Working paper, University of Michigan and Massachusetts Institute of Technology.

Mostagir, Mohamed and James Siderius (2022b), "Social inequality and the spread of misinformation." *Management Science, Forthcoming*.

Mudambi, Maya and Siva Viswanathan (2022), "Prominence reduction versus banning: An empirical investigation of content moderation strategies in online platforms."

Rogers, Richard (2020), "Deplatforming: Following extreme internet celebrities to telegram and alternative social media." *European Journal of Communication*, 35, 213–229.

Shen, Qinlan and Carolyn Rose (2019), "The discourse of online content moderation: Investigating polarized user responses to changes in reddit's quarantine policy." In *Proceedings of the Third Workshop on Abusive Language Online*, 58–69.

Tufekci, Zeynep (2017), *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.