Price of Censorship of a Policy removing Misinformation in a Social Network

Jérémy Petithomme, Corinne Touati, Christophe Bravard

Price of Censorship of a Policy removing Misinformation in a Social Network

Jeremy Petithomme
Inria and Grenoble University
Grenoble, France
jeremy.petithomme@inria.fr

Corinne Touati
Inria and Grenoble University
Grenoble, France
corinne.touati@inria.fr

Christophe Bravard
Grenoble University
Grenoble, France
christophe.bravard@univ-grenoble-alpes.fr

Abstract

The proliferation of misinformation on social media has become a significant concern, particularly in the realms of political discourse and public health. Censorship policies have emerged as a solution to limit the spread of misinformation. However, although censorship reduces the proportion of misinformation disseminated, it also creates an implied truth effect which skews the perception of less reliable information. This paper investigates the impact of censorship policies in an online social network model where agents sequentially observe an article and decide whether to share it with others or not. We measure the impact of censorship in the *virality* of articles containing misinformation and observe that while censorship can effectively reduce the spread of misinformation, it also allows less reliable articles to spread over the network. Specifically, we quantify the "price of censorship", a variation of the price of anarchy, associated with these censorship policies using a formal model that incorporates agents' beliefs, network structure, and content reliability. Unlike usual frameworks of resources allocation games in commutation networks, we show that the price of censorship is unbounded and we exhibit minimal limit case scenarios.

I. Introduction

The circulation of false information on social media platforms has become a pervasive problem, driven by factors such as users' interest in the virality of shared content rather than its truthfulness (see Pennycook [1]), political motivations, and the algorithmic design of social networks. Misinformation has been particularly problematic in contexts such as the COVID-19 pandemic (see Mian and Khan [2]) and political elections, where it can be used as a political weapon and exacerbate societal divisions. Echo chambers (see Jamieson and Cappella [3], Cinelli et al. [4]), which arise when users communicate and share content with like-minded individuals, have been identified in Törnberg [5] as conducive environments for the spread of misinformation. In addition, social platforms' algorithms create filter bubbles (see Pariser [6]), promoting content within a sphere of like-minded individuals, in order to maximize engagement. These filter bubbles also contribute to increasing the virality of untrustworthy content.

Several strategies have been proposed to address these challenges. Common solutions include censorship and fact-checking to combat misinformation. However, these methods can sometimes have unintended consequences and be counter-productive. For example, Ecker et al. [7] and Swire-Thompson et al. [8] show that fact-checking can inadvertently increase the spread of misinformation under certain conditions; Pennycook et al. [9] experimentally show that censorship in a Bayesian belief-updating model can lead to an "implied truth effect", meaning that agents are more likely to believe in the veracity of certain articles when a censorship policy is applied.

In contrast to these approaches, in this article we formally study the impact of censorship policies. We base our model on the seminal work of Acemoglu et al. [10] in which agents in a social network have prior beliefs about a state of the world. We consider the diffusion of an article in the social network. This article has a level of reliability r, which models its likelywood to convey misinformation. Censorship is modeled by considering a regulator removing a proportion $\delta \in [0; 1]$ of articles containing misinformation.

Our first major contribution is to consider the concept of "price of censorship", similar to some price of anarchy (Roughgarden [11]) in the context of censorship in social networks. While the price of anarchy has been extensively used in resources allocation problems (such as transportation system, communication systems and various variants of congestion games), we integrate its use in the fight against fake-news, to quantify how much this policy can increase the dissemination of some misinformation messages within a social network. In other words, we determine the maximum value of $\frac{A}{B}$, where A is the number of agents reached by an article that contains misinformation under a censorship policy, and B is the proportion of agents reached by this same article without any censorship. We further provide minimalist examples exhibiting limit-case scenarios. They can be used as toy-examples for further case-analyses and studies.

We further show that for any level of censorship δ, the price of censorship is unbounded, that is, there are articles containing misinformation that are disseminated to all agents when a censorship policy is applied ($\delta > 0$), while this dissemination does not occur in the absence of censorship ($\delta = 0$). This contrasts with allocation games where the price of anarchy is generally bounded. These findings highlight the need for any censorship policy to be carefully tailored to the specific social media platform and user profiles before

implementation.

Our study is divided into three parts. In Section II, we present the model setup. In Section III, we establish thresholds of reliability that an article must meet in order to be disseminated in both uncensored and censored situations. Finally, in Section IV, we illustrate this "paradox" with a minimalistic example.

II. Model setup and strategies

Our model is based on the seminal work of Acemoglu et al. [10]. Let θ be the state of the world, taken in the set $\Theta = \{L, R\}$. For example, $\theta = L$ or $\theta = R$ represents the situation where the left-wing or right-wing candidate is considered more qualified for political office, respectively. We consider a set of $N+1$ agents having prior (ideological) beliefs about θ. Agent i's prior that $\theta = R$ is denoted by b_i.

Sharing network

We consider an oriented graph with $N+1$ vertices, representing the agents. The neighborhood $\mathcal{N}_i$ of agent i is the set of agents connected to agent i directly by an outgoing link in the sharing network. As deployed in most actual infrastructures, the social media algorithm determines the sharing network to maximize content sharing.

Article generation

An article is a three dimensional vector (r, m, ν), where $r \in [0,1]$ is the reliability of the article, $m \in \{L, R\}$ the message it vehiculates about the state of the world and $\nu \in \{\mathcal{T}, \mathcal{M}\}$ indicates whether its content is truthful or contains misinformation.

At the beginning of the game, the article (r, m, ν) is generated as follows:

1) The article receives a reliability score $r \in [0,1]$.
2) The veracity of the article is drawn randomly with $\nu = \mathcal{T}$ with probability r and $\nu = \mathcal{M}$ with probability $1-r$.
3) Finally, if $\nu = \mathcal{T}$, then the message generated is $m = \theta$ with probability $p = 1$. Conversely, if $\nu = \mathcal{M}$, the message generated is $m = \theta$ with probability $q = 1/2$.

Observe that the message of a truthful article always contains the true state of the world. Meanwhile, an article containing misinformation can be identified as just noise (or propaganda) and its message can be either true or false without distinction.[1]

While (r, m) is common knowledge to all agents receiving the message, ν is unknown to all agents. We assume that agents are Bayesian and update their beliefs about ν using Bayes' rule, given their beliefs about θ and the observed data (r, m).

[1]A lower value of q would not change the results of the paper, we choose $q = 1/2$ to simplify the calculations.

Social media behavior

After receiving an article, an agent i has three possible actions $a_i \in \{\mathcal{S}, \mathcal{I}, \mathcal{D}\}$

- Share ($\mathcal{S}$): the agent shares the article, and their direct neighbors receive it.
- Ignore ($\mathcal{I}$): the agent ignores the article.
- Dislike ($\mathcal{D}$): the agent dislikes the article, expressing disagreement, believing it contains misinformation.

News diffusion

Consider that time is discrete. At $t = 0$, the article is sent to a random agent. At each time step, when agent i receives an article, they choose $a_i \in \{\mathcal{S}, \mathcal{I}, \mathcal{D}\}$. If $a_i = \mathcal{S}$, agent i shares the article and sends it to all their neighbors $j \in \mathcal{N}_i$. If $a_i = \mathcal{I}$, agent i ignores it and the article does not reach their neighbors. Finally, if $a_i = \mathcal{D}$, agent i dislikes it, and the article does not reach their neighbors. Moreover, in the latter case, a negative feedback is sent to all their neighbors who have shared the article.

Obviously, at time $t = 0$, if the first agent receiving the message (called the seed) chooses action $\mathcal{I}$ or $\mathcal{D}$, the message does not propagate through the network. However, if the first agent takes action $\mathcal{S}$, dissemination begins.

Agent's payoff

The payoff of agent i is given by

$$U_i = \begin{cases} 0 & \text{if } a_i = \mathcal{I}, \\ \tilde{u}\mathbf{1}_{\mathcal{M}} - \tilde{c} & \text{if } a_i = \mathcal{D}, \\ u\mathbf{1}_{\mathcal{T}} - c\mathbf{1}_{\mathcal{M}} + \kappa S_i - dD_i & \text{if } a_i = \mathcal{S}, \end{cases}$$

where u, c, κ, d, $\tilde{u}$ and $\tilde{c}$ are positive parameters, $\mathbf{1}_{\mathcal{T}} = 1$ if the article is truthful, $\mathbf{1}_{\mathcal{T}} = 0$ otherwise, and $\mathbf{1}_{\mathcal{M}} = 1$ if the article contains misinformation, $\mathbf{1}_{\mathcal{M}} = 0$ otherwise[2]. Finally, $S_i = |\{j \in \mathcal{N}_i, a_j = \mathcal{S}\}|$ is the number of i's neighbors also sharing the article, and $D_i = |\{j \in \mathcal{N}_i, a_j = \mathcal{D}\}|$ is the number of i's neighbors disliking the article.

To prevent the existence of equilibria in which all agents share every article, we assume that κ is upper bounded:[3]

$$\kappa < \max\left(\frac{\tilde{u} - \tilde{c} + c}{N}, \frac{c}{N}\right). \tag{1}$$

Let us provide some intuition behind the payoff function.

1) The agent has neither an advantage, nor a disadvantage in ignoring an article: it costs them nothing.
2) The agent is rewarded with $\tilde{u}$ if they report misinformation, but disliking incurs a constant cost $\tilde{c}$.
3) The payoff when the agent shares an article can be divided into two parts:
   - their interest in the veracity of the article, represented by $U_i^{(1)} = u\mathbf{1}_{\mathcal{T}} - c\mathbf{1}_{\mathcal{M}}$, and

[2]Note that $\mathbf{1}_{\mathcal{M}} = 1 - \mathbf{1}_{\mathcal{T}}$ and thus the utility for sharing could be rewritten using only $\mathbf{1}_{\mathcal{T}}$, but for simplicity and intuitive understanding of the payoff, we prefer to write it this way.

[3]This value ensures that $U_i(\mathcal{S})$ is strictly lower than either $U_i(\mathcal{I})$ or $U_i(\mathcal{D})$ for an article with reliability 0.

- the externalities of the actions of their neighbors, represented by $U_i^{(2)} = \kappa S_i - dD_i$.

The agent's incentives regarding the veracity of the article come from their desire to share content that is both of high quality and aligns with their beliefs. Clearly, $U_i^{(1)} = u$ if agent i shares reliable information, and $U_i^{(1)} = -c$ if they share misinformation. The externalities resulting from the actions chosen by the agent's neighbors reflect the agent's desire to "create buzz", regardless of the article's veracity. Thus, $U_i^{(2)}$ depends on the number of the neighbors of agent i who also share the article and the number of their neighbors who dislike the article. Specifically, the more the article is shared by the neighbors, the greater the reward for the agent; conversely, the more the neighbors dislike the article, the greater the penalty for the agent.

Note that the content of an article with a reliability score of $r = 0$ is not necessarily false. Indeed, such an article is true with probability $1/2$. However, agents are not interested in sharing an article just because it could be true. They want to share articles that are *truthful*, i.e., high-quality, well-sourced, and honest.

Strategies and equilibria

Without loss of generality, we fix the article's message as $m = R$ for the rest of the paper. Recall that b_i denotes agent i's prior that $\theta = R$. Hence, a prior belief b_i close to 1 indicates a greater propensity to share the article for player i. When agent i receives an article with reliability r and message $m = R$, they compute their (*ex-post*) belief that the article is truthful, π_i, according to Bayes' rule:

$$\begin{aligned}\pi_i &= \mathbf{P}[\nu = \mathcal{T} | m = R, r] \\ &= \frac{b_i \cdot r}{\frac{1}{2}(1-r) + b_i \cdot r}. \end{aligned} \tag{2}$$

It is worth noting that π_i is not an update of b_i, but the *ex-post* belief of the agent about ν, the truthfulness of the article, based on their prior b_i.

In the following, we focus on Bayesian Nash equilibria, which we will simply refer to as "equilibria". Agent i is said to use a *cutoff strategy* if there exist $b_i^*(r)$ and $b_i^{**}(r)$ such that if $b_i < b_i^*(r)$, the agent dislikes the article ($\mathcal{D}$); if $b_i^*(r) < b_i < b_i^{**}(r)$, the agent ignores the article ($\mathcal{I}$); and if $b_i^{**}(r) < b_i$, the agent shares the article ($\mathcal{S}$). Formally,

$$a_i = \begin{cases} \mathcal{D} & \text{if} \quad b_i < b_i^*(r), \\ \mathcal{I} & \text{if} \quad b_i^*(r) < b_i \leq b_i^{**}(r), \\ \mathcal{S} & \text{if} \quad b_i^{**}(r) \leq b_i. \end{cases} \tag{3}$$

Cutoff-strategies reflect that an agent with a prior close to the message observed is more likely to share it, while an agent whose prior belief is distant from the message is more inclined to reject it. Furthermore, b_i^* and b_i^{**} depend on i as each agent takes their neighbors into account to make their decision (part $U_i^{(2)}$).

Acemoglu et al. [10] shows that there exists a Bayesian-Nash equilibrium and all equilibria are in these cutoff strategies. Thus, an equilibrium can be represented as a vector $(b_1^*, b_1^{**}, \ldots, b_{N+1}^*, b_{N+1}^{**})$. Finally, Acemoglu et al. [10] shows that the set of equilibrium vectors forms a lattice, allowing us to define a *most sharing equilibrium*, where the bounds b_i^* and b_i^{**} are the lowest for each user, among all equilibria. Moreover, Acemoglu et al. [10] shows that in a clique, all agents have the same cutoff-strategies.

III. Model Analysis

Since we are looking for a worst case scenario, our goal is to have the entire population to share the article when a censorship policy is applied, while they do not share the message in the absence of such a policy. Thus, an echo chamber, where all users have similar beliefs is the most favorable situations for this to happen. Therefore, unlike Acemoglu et al. [10], we consider a population with symmetric beliefs about θ, similar to what might occur in an echo chamber (see Cinelli et al. [4]). We assume that each agent has the same belief and set for every i, $b_i = 1 - \tau$, with $\tau > 0$.

Moreover, Acemoglu et al. [10] shows that the optimal strategy to maximize the number of shares in the network is to create two "islands": one where no one is connected and another where everyone is connected (the filter bubble). In other words, the network consists of a clique and isolated agents. Consequently, agents not connected to the filter bubble have no impact on the dissemination of information. We thus focus on the filter bubble and unlike Acemoglu et al. [10], we consider the complete graph as the sharing network, i.e. $\mathcal{N}_i = \{1, \ldots, i-1, i+1, \ldots, N+1\}$ for all i.[4]

Since all agents are in a clique, they all have the same cutoff-strategy (Eq. (3)) at the most sharing equilibrium. Thus, for a given article, they will all choose the same action. We formalize this in the following lemma.

Lemma 1. *Let i be the seed agent. At the most sharing equilibrium, $a_j = a_i$ for all $j \in \{1, \ldots, N+1\}$.*

The payoff for agent i when they choose to share, and every other user also shares, is

$$U_i(\mathcal{S}) = u\mathbf{1}_{\mathcal{T}} - c\mathbf{1}_{\mathcal{M}} + \kappa \cdot N - d \cdot 0 = u\mathbf{1}_{\mathcal{T}} - c\mathbf{1}_{\mathcal{M}} + \kappa \cdot N.$$

We aim to determine the value of r for which the most sharing equilibrium is characterized by every agent choosing to share ($\mathcal{S}$). It is equivalent to determine for which value of r we have

$$u\mathbf{1}_{\mathcal{T}} - c\mathbf{1}_{\mathcal{M}} + \kappa \cdot N \geq \max(U_i(\mathcal{I}), U_i(\mathcal{D})).$$

Thus, for simplicity of notation, we set $S_i = N$ and $D_i = 0$ for the remainder of the paper. Formally, our next sections will be about determining for which value of r we have $U_i(\mathcal{S}) \geq \max(U_i(\mathcal{I}), U_i(\mathcal{D}))$. When this inequality holds, it means that $(a_1 = \mathcal{S}, a_2 = \mathcal{S}, \ldots, a_{N+1} = \mathcal{S})$ is an equilibrium, and thus the most sharing equilibrium. Otherwise, the most sharing

[4]Observe that now, all agents are identical, thus the choice of the seed agent is irrelevant.

equilibrium is either $(a_1 = \mathcal{I}, a_2 = \mathcal{I}, \ldots, a_n = \mathcal{I})$ or $(a_1 = \mathcal{D}, a_2 = \mathcal{D}, \ldots, a_n = \mathcal{D})$, and the article does not circulate through the social media.

A. Without censorship

Recall the value of π_i for any agent:

$$\pi_i = \frac{b_i \cdot r}{\frac{1}{2}(1-r) + b_i \cdot r} = \frac{2(1-\tau)\cdot r}{1 - r + 2(1-\tau)\cdot r}.$$

Note that

1) π_i is strictly increasing with r, as agents believe more in the truthfulness of an article with high reliability,
2) $\pi_i \to 0$ when $r \to 0$, as agents are certain of the nature of an article when $r = 0$ (misinformation),
3) and $\pi_i \to 1$ when $r \to 1$, as agents are certain of the nature of an article when $r = 1$ (truthfulness).

Let us compute the expected payoff of each user, considering their action. We have

$$\begin{aligned} & U_i(\mathcal{D}) && = \tilde{u}(1-\pi_i) - \tilde{c}, \\ & U_i(\mathcal{I}) && = 0, \\ \text{and} \quad & U_i(\mathcal{S}) && = u\pi_i - c(1-\pi_i) + N\cdot\kappa \\ & && = (u+c)\pi_i - c + N\cdot\kappa. \end{aligned}$$

Remark 1. Observe that $U_i(\mathcal{D})$, $U_i(\mathcal{S})$ and $U_i(\mathcal{I})$ are all functions of π_i, which in turns is a function of r. However, to avoid cumbersome notations, in the remaining of the paper, we simply write it as $U_i(\mathcal{D})$ or $U_i(\mathcal{D})(r)$ instead of $U_i(\mathcal{D})(\pi_i(r))$, when there is no ambiguity. We do the same for $U_i(\mathcal{S})$ and $U_i(\mathcal{I})$.

Payoff properties

We briefly study the payoffs $U_i(\mathcal{D})$ and $U_i(\mathcal{S})$ as functions of r with the following propositions.

Proposition 2. *The payoff function resulting from action $\mathcal{D}$ (disliking)[5] satisfies:*

1) $U_i(\mathcal{D})$ is strictly decreasing with r and $U_i(\mathcal{D})(1) < 0$.
2) If $U_i(\mathcal{D})(0) > 0$, then there exists a unique $r_\mathcal{D}$ such that $U_i(\mathcal{D})(r_\mathcal{D}) = 0$ and

$$r_\mathcal{D} = \frac{1}{A_0(1-\tau)+1}, \tag{4}$$

with

$$A_0 = \frac{2\tilde{c}}{\tilde{u}-\tilde{c}}.$$

The first part of Proposition 2 is consistent with the intuition that an agent is less inclined to dislike a more reliable article. Furthermore, this proposition also demonstrates that an agent will never dislikes an article with reliability 1, i.e., which is truthful with probability 1.

When $U_i(\mathcal{D})(0) > 0$, we denote by $r_\mathcal{D}$ the unique value such that $U_i(\mathcal{D})(r_\mathcal{D}) = 0$, otherwise we set $r_\mathcal{D} = 0$.

[5]Note that it does not depend on the other agents' actions.

Proposition 3. *The payoff function of player i when they choose $\mathcal{S}$ (sharing), and all other players adopt the same strategy satisfies:*

1) $U_i(\mathcal{S})$ is strictly increasing with r and $U_i(\mathcal{S})(1) > 0$.
2) If $U_i(\mathcal{S})(0) < 0$, then there exists a unique $r_\mathcal{S}$ such that $U_i(\mathcal{S})(r_\mathcal{S}) = 0$ and

$$r_\mathcal{S} = \frac{1}{A_1(1-\tau)+1}, \tag{5}$$

with

$$A_1 = \frac{2u + 2N\kappa}{c - N\kappa}.$$

The first part of Proposition 3 aligns with the intuition that an agent is more inclined to share a more reliable article. Moreover, this proposition establishes that an agent will always share an article with reliability 1, i.e., an article that is truthful with probability 1. This behavior is explained by the fact that the agent is in an echo chamber, aware that each of their neighbors will also share the article, thereby rewarding the agent for sharing it too.

When $U_i(\mathcal{S})(0) < 0$, we denote by $r_\mathcal{S}$ the unique value such that $U_i(\mathcal{S})(r_\mathcal{S}) = 0$, otherwise we set $r_\mathcal{S} = 0$.

Case analysis

Next, we divide our study into two cases according to the relative values of $r_\mathcal{D}$ and $r_\mathcal{S}$.

1) The case $r_\mathcal{D} \le r_\mathcal{S}$, illustrated in
   a) Figure 1 where $r_\mathcal{D} = 0$, and
   b) Figure 2 where $r_\mathcal{D} > 0$.
2) The case $r_\mathcal{D} > r_\mathcal{S}$, illustrated in Figure 3.

Let us discuss each of these cases.

Case 1a: When $r_\mathcal{D} = 0$, we have $U_i(\mathcal{D}) \le 0$ for all r. Note that this happens if and only if $\tilde{u} \le \tilde{c}$, that is, when the cost to dislike an article is higher than the maximal potential reward for calling out misinformation. Then, every agent is only focused on whether their payoff for sharing an article is greater than or equal to 0, their payoff for ignoring it. Thus, the whole population shares the article if and only if $U_i(\mathcal{S}) \ge U_i(\mathcal{I}) = 0$, i.e., if and only if the article is sufficiently reliable $r \ge r_S$.

Case 1b: When $r_\mathcal{D} > 0$ and $r_\mathcal{D} \le r_\mathcal{S}$, the payoff for disliking $U_i(\mathcal{D})$ drops below zero at a lower value of r than the value at which $U_i(\mathcal{S})$ exceeds 0. Therefore, an agent decides to share an article based only on whether her sharing payoff is greater than or equal to her ignoring payoff. Consequently, as in Case 1a, the entire population shares the article if and only if $U_i(\mathcal{S}) \ge U_i(\mathcal{I}) = 0$, i.e., if and only if the article has reliability $r \ge r_S$.

Case 2: When $r_\mathcal{D} > r_\mathcal{S}$, the payoff for disliking, $U_i(\mathcal{D})$, falls below zero at a higher value of r than the value at which $U_i(\mathcal{S})$ rises above zero. Therefore, an agent does not necessarily share an article if $U_i(\mathcal{S}) > 0$; instead, the agent focuses on whether their payoff for sharing the article is greater than or equal to their payoff for disliking it. The entire

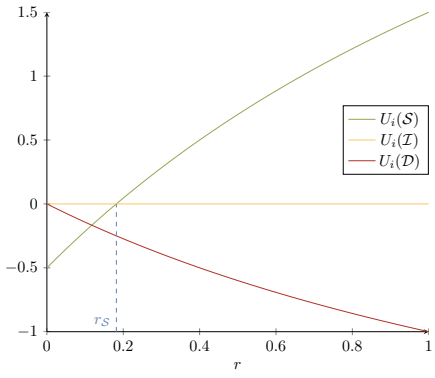

Fig. 1. Case $r_{\mathcal{D}} = 0$ (Case 1a)

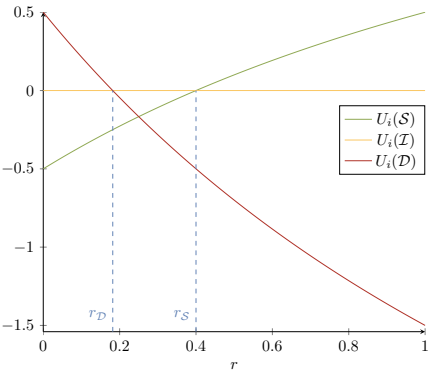

Fig. 2. Case $r_{\mathcal{D}} \leq r_{\mathcal{S}}$ (Case 1b)

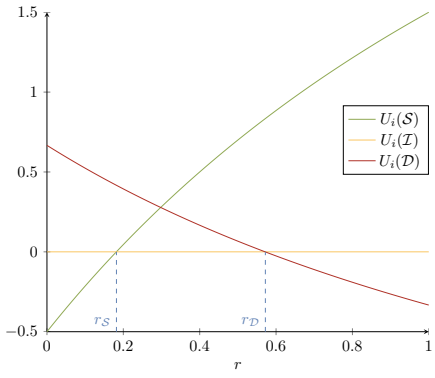

Fig. 3. Case $r_{\mathcal{D}} > r_{\mathcal{S}}$ (Case 2)

Fig. 4. Shapes of utility functions and optimal strategies of agents upon receiving an article carrying message $m = R$ as a function of its reliability r.

population shares the article if and only if $U_i(\mathcal{S}) \geq U_i(\mathcal{D})$.

The following proposition formalizes these claims and establishes some thresholds on the reliability r of an article, in each of the cases, so that the entire population shares it.

Proposition 4. *We have*

(i) If $r_{\mathcal{D}} \leq r_{\mathcal{S}}$, then every agent shares an article (r, R) at the most sharing equilibrium if and only if

$$r \geq r_{\mathcal{S}}. \tag{6}$$

(ii) If $r_{\mathcal{D}} > r_{\mathcal{S}}$, then

a) If $U_i(\mathcal{D})(0) > U_i(\mathcal{S})(0)$, then every agent shares an article (r, R) at the most sharing equilibrium if and only if

$$r \geq \frac{1}{A_{2a}(1-\tau)+1}, \tag{7}$$

with

$$A_{2a} = \frac{2u + 2N\kappa + 2\tilde{c}}{\tilde{u} + c - N\kappa - \tilde{c}}.$$

b) If $U_i(\mathcal{D})(0) \leq U_i(\mathcal{S})(0)$, everybody shares any article.

Remark 2. Case (ii)b of the Proposition never occurs under the assumption on κ given by Eq. (1).

Note that Case 1 corresponds to situation (i) of Proposition 4, and that Case 2 corresponds to situation (ii).

We have provided the necessary and sufficient conditions for an article's reliability to determine whether it will be shared across the entire network or not shared by any agent in the absence of censorship. In the following section, we examine what happens in the presence of censorship.

B. With censorship

Censorship policy

We assume that a regulator has the capability to detect an article containing misinformation and censor it before it is introduced into the network (or alternatively, tag it as misinformation as in Clayton et al. [12]). To model this, we assume that the regulator can detect an article containing misinformation with probability $\delta \in (0, 1)$. Additionally, we assume that an article containing reliable content will never be falsely detected as misinformation (i.e., there are no false positives).

In the rest of the paper, we identify some threshold on the reliability r of an article that determine whether it will be shared, based on the model parameters and the level of censorship. We show that, depending on the parameters, a threshold exists where no article below it is shared, while all articles above it are shared by all agents.

As empirically documented in Pennycook et al. [9], removing a portion of the misinformation leads to an "implied truth effect" on the agents. Indeed, knowing that a regulator censors a portion of the misinformation, an article that reaches a user is more likely to be truthful. This implied truth effect is mathematically represented in our model by replacing r with $\phi(r)$ representing the probability for an article which has not been censured to be truthful.

Recall that there are no false positives, and therefore the probability that an article is not censored given that it is truthful equals 1 and that the probability that an article is not censored given that it contains misinformation is $1-\delta$. Bayes' rule directly leads to

$$\phi(r) = \frac{r}{r + (1-\delta)(1-r)}.$$

The agents' beliefs π_i are then modified in the following way:

$$\pi_i = \frac{b_i \phi(r)}{\frac{1}{2}(1-\phi(r)) + b_i \phi(r)} = \frac{2(1-\tau)r}{(1-\delta)(1-r) + 2(1-\tau)r}.$$

As expected, the implied truth effect leads to an increase in users' belief in the sincerity of an article.

Note that, as in the case with no censorship,

1) π_i is strictly increasing with r,
2) $\pi_i \to 0$ when $r \to 0$,

3) and $\pi_i \to 1$ when $r \to 1$.

In addition, π_i is increasing with δ, as a non-censored article is more likely to be true.

Remark 3. Observe that with censorship, $U_i(\mathcal{D}), U_i(\mathcal{I})$ and $U_i(\mathcal{S})$ are not only functions of r but also functions of δ. Again, to avoid cumbersome notations, and to distinguish them from the payoffs without censorship, we denote by $U_i^\delta(\mathcal{D})$ or $U_i^\delta(\mathcal{D})(r)$ the associated utilities for disliking the article instead of $U_i^\delta(\mathcal{D})(\pi_i(r,\delta))$, when there is no ambiguity. We do the same for $U_i^\delta(\mathcal{S})$ and $U_i^\delta(\mathcal{I})$.

Payoff properties

Similar to the case without censorship, we derive a function analysis to study the payoffs $U_i(\mathcal{D})$ and $U_i(\mathcal{S})$ as functions of r with the following propositions.

Proposition 5. *The payoff function resulting from action $\mathcal{D}$ (disliking) satisfies:*

1) $U_i^\delta(\mathcal{D})$ is strictly decreasing with r and $U_i^\delta(\mathcal{D})(1) < 0$.
2) If $U_i^\delta(\mathcal{D})(0) > 0$, then there exists a unique $r_\mathcal{D}^\delta$ such that $U_i^\delta(\mathcal{D})$, and

$$r_\mathcal{D}^\delta = \frac{1-\delta}{A_0(1-\tau)+1-\delta} \tag{8}$$

where A_0 is the same as in Proposition 2.

By Lemma A given in the appendix, we have $r_\mathcal{D}^\delta < r_\mathcal{D}$ if $r_\mathcal{D} \neq 0$. Intuitively, this is consistent with the fact that an agent is less inclined to dislike an article that has been checked and could have been censored. The entire curve of $U_i(\mathcal{D})$ shifts downwards, and the value of $r_\mathcal{D}$ shifts closer to 0.

When $U_i^\delta(\mathcal{D})(0) > 0$, we denote by $r_\mathcal{D}^\delta$ the unique value such that $U_i^\delta(\mathcal{D})(r_\mathcal{D}^\delta) = 0$, otherwise we set $r_\mathcal{D}^\delta = 0$.

Proposition 6. *The payoff function resulting from action $\mathcal{S}$ (sharing) when all other players adopt the same strategy satisfies:*

1) $U_i^\delta(\mathcal{S})$ is strictly increasing with r, and $U_i^\delta(\mathcal{S})(1) > 0$.
2) If $U_i^\delta(\mathcal{S})(0) < 0$, then there exists a unique $r_\mathcal{S}^\delta$ such that $U_i^\delta(\mathcal{S})(0) = 0$, and

$$r_\mathcal{S}^\delta = \frac{1-\delta}{A_1(1-\tau)+1-\delta} \tag{9}$$

where A_1 is the same as in Proposition 3.

Again, by Lemma A given in the appendix, we have $r_\mathcal{S}^\delta < r_\mathcal{S}$ if $r_\mathcal{S} \neq 0$. Intuitively, this is consistent with the fact that an agent is more inclined to share an article that has been checked and could have been censored. The entire curve of $U_i(\mathcal{S})$ shifts upwards, and the value of $r_\mathcal{S}$ shifts closer to 1.

When $U_i^\delta(\mathcal{S})(0) < 0$, we denote by $r_\mathcal{S}^\delta$ the unique value such that $U_i^\delta(\mathcal{S})(r_\mathcal{S}^\delta) = 0$, otherwise we set $r_\mathcal{S}^\delta = 0$.

Note that $U_i(\mathcal{D})$ and $U_i(\mathcal{S})$ have the same properties as in the case with no censorship. This is not surprising since the censoring does not change the positive impact of a better reliability on agents' beliefs.

Case analysis

Now that we have examined the properties of utility functions, we divide our study into two cases, depending on the relative values of $r_\mathcal{D}^\delta$ and $r_\mathcal{S}^\delta$:

1^δ) The case $r_\mathcal{D}^\delta \leq r_\mathcal{S}^\delta$,
2^δ) The case $r_\mathcal{D}^\delta > r_\mathcal{S}^\delta$.

These cases are similar to cases 1 and 2 from the situation without censorship.

We conduct a similar analysis as in in Section III-A, replacing $r_\mathcal{X}$ with $r_\mathcal{X}^\delta$ and $U_i(\mathcal{X})$ with $U_i^\delta(\mathcal{X})$, where $\mathcal{X} \in \{\mathcal{D}, \mathcal{I}, \mathcal{S}\}$. For conciseness, we omit illustrative figures, which are similar to Figures 1, 2 and 3. This leads to the necessary and sufficient conditions for the most sharing equilibrium being the one where all agents share the article, summarized in the following proposition.

Proposition 7. *We have*

(i) If $r_\mathcal{D}^\delta \leq r_\mathcal{S}^\delta$, then every user shares the article (r, R) at the most sharing equilibrium if and only if

$$r \geq r_\mathcal{S}^\delta. \tag{10}$$

(ii) If $r_\mathcal{D}^\delta > r_\mathcal{S}^\delta$, then

a) If $U_i^\delta(\mathcal{D})(0) > U_i^\delta(\mathcal{S})(0)$, then every user shares an article (r, R) at the most sharing equilibrium if and only if

$$r \geq \frac{1-\delta}{A_{2a}(1-\tau)+1-\delta}, \tag{11}$$

where A_{2a} is the same as in Proposition 4.(ii)a.

b) If $U_i^\delta(\mathcal{D})(0) \leq U_i^\delta(\mathcal{S})(0)$, then every user shares any article at the most sharing equilibrium.

Note that as in the situation without censorship, Case 1^δ corresponds to situation (i) of Proposition 7, and that Case 2^δ corresponds to situation (ii).

In this section, we provided the conditions on the reliability r of an article to spread through the network, when censorship is introduced. The analysis was similar to the case without censorship since the system behaves as if the value of r were modified. We have established through utility analysis that an article containing misinformation hat passes through censorship is more likely to be shared across the platform than without censorship.

C. A small censorship can spread infinitely more misinformation

We have determined in Sections III-A and III-B the different thresholds of an article's reliability for it to be shared (which we will refer to as *sharing thresholds*), both in the case without censorship and in the case with censorship. In this section, we show that, for any given level of censorship δ, it is *always* possible to find an article containing misinformation that is not shared if no censorship policy is applied, but is shared by *all* agents once the censorship policy δ is applied.

Price of Censorship

We define the price of censorship of a policy as the maximum value of $N_{\mathcal{S}}^{\delta}(r)/N_{\mathcal{S}}(r)$, where $N_{\mathcal{S}}^{\delta}(r)$ is the number of agents reached by an article with misinformation of reliability r that has passed through a censorship policy of level δ, and $N_{\mathcal{S}}(r)$ is the proportion of agents reached by the same article without any censorship.

Our main result on the price of censorship is stated in the following theorem.

Theorem 8. *For any level of censorship δ, there exists $r > 0$ such that, given any article (r, R), if no censorship policy is applied, the article is not shared by anyone. However, if a censorship policy at or above level δ is applied and the article is not censored, then it is shared by everyone.*

Proof of Theorem 8. First, we introduce Lemma 9, that states the monotone behavior of censoring.

Lemma 9. *Let $\delta > 0$ and $r > 0$ such that an article (R, r) is not shared with no censorship, but shared among the entire population with a censorship of level δ. Then, for any censorship $\delta' \geq \delta$, the article is also shared among the entire population.*

Lemma 9 reflects the fact that a greater level of censorship creates a more significant implied truth effect. Thanks to this lemma, it is sufficient to show that there exists a reliability level r such that (i) any article (r, R) is not shared when no censorship policy is applied, and (ii) if a censorship policy at level δ is applied, any article (r, R) which is not censured is shared by all agents.

Let us fix $\delta > 0$. Formally, we have to find, for each case of Proposition 4, a value of r under the sharing threshold, but above the sharing threshold in the corresponding case of Proposition 7. Thus, we need to determine whether, and if so how, the cases in propositions 4 and 7 are related.

To do this, we link the inequalities between $r_{\mathcal{D}}$ and $r_{\mathcal{S}}$, and between $r_{\mathcal{D}}^{\delta}$ and $r_{\mathcal{S}}^{\delta}$. Lemma B implies that

$$r_{\mathcal{D}} \leq r_{\mathcal{S}} \iff r_{\mathcal{D}}^{\delta} \leq r_{\mathcal{S}}^{\delta}. \tag{12}$$

Therefore,

- Case (i) of Proposition 4 holds if and only if Case (i) of Proposition 7 holds,
- Case (ii) of Proposition 4 holds if and only if Case (ii) of Proposition 7 holds,

We deal with each of these situations in the following.

Case (i): By Proposition 4.(i) and Proposition 7.(i), the paradox, which is that an article is shared under some censorship policy but not without any censorship, arises if inequality (6) does not hold but inequality (10) holds. This leads to the following condition on r:

$$r_{\mathcal{S}}^{\delta} < r < r_{\mathcal{S}}. \tag{13}$$

By Lemma A, $r_{\mathcal{S}}^{\delta} < r_{\mathcal{S}}$ and there exists some r satisfying condition (13).

Case (ii): By Proposition 4.(ii)a and Proposition 7.(ii)a, the paradox arises if inequality (7) does not hold but inequality (11) holds. This leads to the following condition on r:

$$\frac{1-\delta}{A_{2a}(1-\tau)+1-\delta} < r < \frac{1}{A_{2a}(1-\tau)+1}. \tag{14}$$

By Lemma A, $\frac{1-\delta}{A_{2a}(1-\tau)+1-\delta} < \frac{1}{A_{2a}(1-\tau)+1}$ and there exists some r satisfying condition (14). □

Unbounded Price of Censorship

We have shown that even a low level of censorship can lead to the widespread sharing of misinformation that would otherwise not be shared. Specifically, for any level of censorship, there exist articles that do not spread at all without censorship, i.e., reach only the seed, while they spread throughout the entire network when a censorship policy is applied, i.e., reach the entire population. Therefore, any level of censorship δ results in an unbounded price of anarchy, N.

In the above analysis, we have considered the price of censorship as the worst-case scenario considering that the article containing misinformation has passed the censorship test. However, to reflect the potential harmth of the implied truth effect of censorship, one may consider the actual expected number of shares of any article (r, R) regarding the likelihood of being censored.

We define the *expected price of censorship* as the maximum value of $E[N_{\mathcal{S}}^{\delta}(r)]/N_{\mathcal{S}}(r)$, where $E[N_{\mathcal{S}}^{\delta}(r)]$ is the expected number of agents reached by an article with misinformation of reliability r (that may or not sliped through a censorship policy of level δ), and $N_{\mathcal{S}}(r)$ is the proportion of agents reached by the same article without any censorship.

As the expected number of shares is $(1-\delta)N$, then, by Theorem 8, the expected price of censorship is also unbounded.

IV. Minimalistic example

In this section, we present a minimalistic example that illustrates the price of anarchy.

Consider a system with $N+1$ users all with the same belief $b = 3/4$, and an article with reliability $r > 0$.

We assume $u = c = 1$ and $\kappa = 1/N$ and we assume $\tilde{u} = 1$ and $\tilde{c} = 2/3$ so the payoff of agent i is given by

$$U_i = \begin{cases} 0 & \text{if } a_i = \mathcal{I}, \\ 1/3 - \pi_i & \text{if } a_i = \mathcal{D}, \\ (2\pi_i - 1) + (S_i - D_i)/N & \text{if } a_i = \mathcal{S}. \end{cases}$$

A. Without censorship

In this case, we get the following post-belief π_i for any agent :

$$\pi_i = \frac{3r}{2+r}. \tag{15}$$

Since all agents have the same payoff function, at the most sharing equilibrium, they all choose the same action ($\mathcal{S}$, $\mathcal{I}$ or $\mathcal{D}$). Hence, the payoff of agent i, when all agents share, is

$$U_i(\mathcal{S}) = (2\pi_i - 1) + (N - 0)/N = 2\pi, \tag{16}$$

Let us set $S_i = N$ and $D_i = 0$. Note that under this assumption, $U_i(\mathcal{S}) \geq U_i(\mathcal{D})$ if and only if $\pi_i \geq \frac{1}{9}$, which is equivalent to

$$r \geq \frac{1}{13}. \tag{17}$$

Since $U_i(\mathcal{S}) > U_i(\mathcal{I}) = 0$ for every r, with no censorship, an article is shared among all the population if and only if $r \geq 1/13$ and shared by no agent otherwise.

B. With censorship

Next, consider a censorship removing a proportion $\delta > 0$ of misinformation articles. Because of the implied truth effect r is replaced by $\phi(r)$. Consequently,

$$\pi_i = \frac{3r}{2(1-\delta)(1-r)+3r}.$$

As in the setting without censorship, in the most sharing equilibrium, all agents choose the same action ($\mathcal{S}$, $\mathcal{I}$ or $\mathcal{D}$) since they all have the same payoff function.

Again, under this assumption, $U_i(\mathcal{S}) \geq U_i(\mathcal{I})$ if and only if $\pi_i \geq \frac{1}{9}$, which is now equivalent to

$$r \geq \frac{1-\delta}{13-\delta}. \tag{18}$$

Thus, $U_i(\mathcal{S}) > U_i(\mathcal{I}) = 0$ for any value of r, and an article is shared among the entire population if and only if $r \geq \frac{1-\delta}{13-\delta}$.

Note that $\frac{1-\delta}{13-\delta} < \frac{1}{13}$ for any value of $\delta \in (0,1)$ by Lemma A. Let us then fix a value of r in the interval $[\frac{1-\delta}{13-\delta}, \frac{1}{13})$. Since $r < \frac{1}{13}$, the article is not shared when there is no censorship, by condition (17). Similarly, since $r \geq \frac{1-\delta}{13-\delta}$, the article is shared throughout the entire network as soon as a censorship policy with level δ is implemented, by condition (18).

V. Conclusion

In this paper, we have examined the complex dynamics of misinformation dissemination in the presence of censorship on social networks. Our analysis introduced the concept of the "price of censorship" in this context, quantifying how censorship can inadvertently amplify the spread of false information. We analyzed a formal model in both uncensored and censored environments. Since we were interested in the worst-case scenario of misinformation dissemination, we considered a network structure similar to echo chambers or filter bubbles, where agents with *symmetric prior beliefs* interact. We showed that any level of censorship ($\delta > 0$) allows misinformation to propagate throughout the entire network, resulting in an unbounded price of censorship. Acemoglu et al. [10] have proposed other regulation policies. With very similar calculations, we can show that the price of censorship is unbounded for these policies too.

Future research could explore what happens when beliefs are randomly distributed over a small interval, and investigate whether the paradox still holds for all values of δ. Additionally, it would be valuable to determine the maximum interval size beyond which the paradox no longer occurs. One could also wonder whether the implied truth effect would be as significant in non-Bayesian contexts, such as in the DeGroot model [13], where agents are more focused on the opinions of their neighbors than on the actual true state of the world.

References

[1] G. Pennycook, "The psychology of fake news," *Trends in Cognitive Sciences*, vol. 25, 03 2021.

[2] A. Mian and S. Khan, "Coronavirus: the spread of misinformation," *BMC Medicine*, 03 2020. [Online]. Available: https://doi.org/10.1186/s12916-020-01556-3

[3] K. Jamieson and J. Cappella, *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press, 2008. [Online]. Available: https://books.google.fr/books?id=139Oa4MOsAgC

[4] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, p. e2023301118, 2021. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2023301118

[5] P. Törnberg, "Echo chambers and viral misinformation: Modeling fake news as complex contagion," *PLOS ONE*, vol. 13, p. e0203958, 09 2018.

[6] E. Pariser, *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited, 2011. [Online]. Available: https://books.google.fr/books?id=-FWO0puw3nYC

[7] U. Ecker, J. Hogan, and S. Lewandowsky, "Reminders and repetition of misinformation: Helping or hindering its retraction?" *Journal of Applied Research in Memory and Cognition*, vol. 6, 04 2017.

[8] B. Swire-Thompson, U. Ecker, and S. Lewandowsky, "The role of familiarity in correcting inaccurate information," *Journal of Experimental Psychology Learning Memory and Cognition*, vol. 43, 05 2017.

[9] G. Pennycook, A. Bear, and E. Collins, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management Science*, 08 2019.

[10] D. Acemoglu, A. Ozdaglar, and J. Siderius, "A Model of Online Misinformation," *The Review of Economic Studies*, 12 2023. [Online]. Available: https://doi.org/10.1093/restud/rdad111

[11] T. Roughgarden, *Selfish Routing and the Price of Anarchy*. The MIT Press, 2005.

[12] K. Clayton, S. Blair, J. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, M. Sandhu, R. Sang, R. Scholz-Bright, A. Welch, A. Wolff, A. Zhou, and B. Nyhan, "Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media," *Political Behavior*, vol. 42, 12 2020.

[13] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974. [Online]. Available: http://www.jstor.org/stable/2285509

Appendix

Lemma A. *For all $x > 0$, for all $0 < \delta < 1$, we have*

$$\frac{1-\delta}{x+1-\delta} < \frac{1}{x+1}. \tag{19}$$

Proof. Let us fix $x > 0$ and let $f(\delta) = \frac{1-\delta}{x+1-\delta}$, we have $f'(\delta) = \frac{\delta-1}{(1-\delta+x)^2} < 0$ which leads to the conclusion. □

Lemma B. *For all $A, B > 0$, such that*

$$\frac{1}{A+1} < \frac{1}{B+1}, \tag{20}$$

and for all $0 < \delta < 1$, we have $\frac{1-\delta}{A+1-\delta} < \frac{1-\delta}{B+1-\delta}$.

Proof. Inequality (20) is equivalent to $A > B$, which leads to the conclusion. □