IEEE Transactions on Automatic Control,Issue Date:08.June.2021

Social Media and Misleading Information in a Democracy: A Mechanism Design Approach

Aditya Dave, *Student Member, IEEE*, Ioannis Vasileios Chremos, *Student Member, IEEE*, Andreas A. Malikopoulos, *Senior Member, IEEE*

***Abstract*—In this paper, we present a resource allocation mechanism for the problem of incentivizing filtering among a finite number of strategic social media platforms. We consider the presence of a strategic government and private knowledge of how misinformation affects the users of the social media platforms. Our proposed mechanism incentivizes social media platforms to filter misleading information efficiently, and thus indirectly prevents the spread of fake news. In particular, we design an economically inspired mechanism that strongly implements all generalized Nash equilibria for efficient filtering of misleading information in the induced game. We show that our mechanism is individually rational, budget balanced, while it has at least one equilibrium. Finally, we show that for quasi-concave utilities and constraints, our mechanism admits a generalized Nash equilibrium and implements a Pareto efficient solution.**

***Index Terms*—Social media, fake news, mechanism design, Nash-implementation**

I. Introduction

For the last few years, political commentators have been indicating that we live in a *post-truth* era [1], wherein the deluge of information available on the internet has made it extremely difficult to identify facts. As a result, individuals have developed a tendency to form their opinions based on the *believability* of presented information rather than its truthfulness [2]. This phenomenon is exacerbated by the business practices of social media platforms, which often seek to maximize the *engagement* of their users at all costs. In fact, the algorithms developed by platforms for this purpose often promote conspiracy theories among their users [3].

The sensitivity of users of social media platforms to conspiratorial ideas makes them an ideal terrain to conduct political misinformation campaigns [4], [5]. Such campaigns are especially effective tools to disrupt democratic institutions, because the functioning of stable democracies relies on *common knowledge* about the political actors and the processes they can use to gain public support [6]. The trust held by the citizens of a democracy on common knowledge includes: (i) trust that all political actors act in good faith when contesting for power, (ii) trust that elections lead to a free and fair transfer of power between the political actors, and (iii) trust that democratic institutions ensure that elected officials wield their power in the best interest of the citizens. In contrast, citizens of democracies often have a *contested knowledge* regarding who should hold power and how they should use it [6]. The introduction of *alternative facts* can reduce the trust on common knowledge about democracy, especially if they become accepted beliefs among the citizens. Such disruptions on the trust on common knowledge can be found in the 2016 U.S. elections [7] and Brexit Campaign in 2016 [8], where the spread of misinformation through social media platforms resulted in a large number of citizens mistrusting the results of voting.

To tackle this growing phenomenon of misinformation, in this paper, we consider a finite group of social media platforms, whose users represent the citizens in a democracy, and a democratic government. Every post in the platforms is associated with a parameter that captures its informativeness, which can take values between two extremes: (i) completely factual and (ii) complete misinformation. In our framework, posts that exhibit misinformation can lead to a decrease in trust on common knowledge among the users [9]–[12]. In addition, social media platforms are considered to have the technologies to *filter*, or label, posts that intend to sacrifice trust on common knowledge. Thus, the government seeks to incentivize the social media platforms to use these technologies and filter any misinformation included in the posts.

Motivated by capitalistic values, we induce a *misinformation filtering game* to describe the interactions between the social media platforms and the government. In this game, each platform acts as strategic player seeking to maximize their advertisement revenue from the engagement of their users [7], [13]. User engagement is a metric that can be used to quantify the interaction of users with a platform, and subsequently, how much time they spend on the platform. Recent efforts reported in the literature on misinformation in social media platforms have indicated that increasing filtering of misinformation leads to decreasing of user engagement [14]. There are many possible reasons for this phenomenon. First, filtering reduces the total number of posts propagating across the social network. Second, the users whose opinions are filtered may perceive this action as dictatorial censorship [15], and as a result, they may chose to express their opinions in other platforms. Finally, misinformation tends to elicit stronger reactions, e.g., surprise, joy, sadness, as compared to factual posts [16], which may increase user engagement. Thus, each platform is reluctant to filter misinformation.

In our framework, we consider that the government is also a strategic player, whose utility increases as the trust of the users of social media platforms on common knowledge increases. Consequently, increasing filtering of misinformation by the

This research was supported by the Sociotechnical Systems Center (SSC) at the University of Delaware.

The authors are with the Department of Mechanical Engineering, University of Delaware, Newark, DE, 19716, USA (emails: adidave@udel.edu; ichremos@udel.edu; andreas@udel.edu).

social media platforms increases the utility of the government. Thus the government is willing to make an investment to incentivize the social media platforms to filter misinformation. In our approach, we use mechanism design to distribute this investment among the platforms optimally, and in return, implement an optimal level of filtering.

Mechanism design was developed for the implementation of system-wide optimal solutions to problems involving multiple rational players with conflicting interests, each with private information about preferences [17]. Note that this approach is different from traditional approaches to decentralized control with private information [18]–[21] because the players are not a part of the same time, but in fact, have private and competitive utilities. The fact that Mechanism design optimizes the behaviour of competing players has led to broad applications spanning different fields including economics, politics, wireless networks, social networks, internet advertising, spectrum and bandwidth trading, logistics, supply chain, management, grid computing, and resource allocation problems in decentralized systems [22]–[28].

The contribution of this paper is as follows. We present an indirect mechanism to incentivize social media platforms to filter misleading information. We show that our proposed mechanism is (i) feasible, (ii) budget balanced, (iii) individual rational, and (iv) strongly implementable at the equilibria of the induced game. We prove the existence of at least one generalized Nash equilibrium and show that our mechanism induces a Pareto efficient equilibrium.

The rest of the paper is organized as follows. In Section II, we provide the modeling framework and problem formulation. In Section III, we present our mechanism, and in Section IV, we prove the associated properties of the mechanism. In Section V, we interpret the mechanism and present a descriptive example. Finally, in Section VI we conclude and present some directions for future research.

II. Problem Formulation

We consider a democratic society consisting of a finite and nonempty set of social media platforms $\mathcal{I} = \{1, \dots, I\}$, $I \in \mathbb{N}$, and a government. We refer to the social media platforms and the government collectively as the *players*, and denote the set of all players by $\mathcal{J} = \mathcal{I} \cup \{0\}$, where the index 0 corresponds to the government. The players strategically take actions in a *misinformation filtering game* that is described in this section.

A. Misinformation Filtering Game for Platforms

Let the informativeness of a post on platform $i \in \mathcal{I}$ be denoted by $x_i \in [0, 1]$, where $x_i = 0$ indicates that the post contains complete misinformation and $x_i = 1$ indicates that the post contains completely factual information. Our hypothesis, inspired by [6], [9]–[12], states that the emergence of posts with many falsehoods and a low informativeness, i.e., $x_i \to 0$, leads to a decrease of trust of the users on common knowledge about democracy. Recall that common knowledge about democracy refers to knowledge of political actors in a democratic society and the process they use to gain public support.

Each social media platform $i \in \mathcal{I}$ has the technological means to detect and filter misinformation. In the misinformation filtering game that we impose in our framework, the action a_i of platform i represents the level of filtering imposed by i and takes values in a feasible set of actions $\mathcal{A} = [0, 1]$. Each action a_i minimizes the spread of a post that has informativeness $x_i < a_i$, while posts with informativeness $x_i \geq a_i$ are unaffected. In practice, filtering of misinformation can be implemented in many ways. The social media platform can place warnings on each post with $x_i < a_i$ to inform the users of their falsehood, or they can modify their algorithms to limit the propagation of such posts among users. Thus, the action a_i represents the lower threshold on informativeness that is acceptable by platform i. To this end, we refer to the action a_i as the filter of platform i.

Each platform $i \in \mathcal{I}$ generates revenue by monetizing the *engagement* of their users through advertisements [13]. By increasing filtering of misinformation there is a decrease in user engagement [14]. This may be due to a perception of censorship among users [15], and as a result, they may chose to express their opinions in other platforms. Consider, for example, platform $l \in \mathcal{I}$ with a filter $a_l > a_i$. Some of the users of l, whose posts have been marked up by the filter, may migrate to platform i which will lead to an increase in the engagement of platform i. This phenomenon motivates us to define a set of *competing platforms*.

Definition 1. For each platform $i \in \mathcal{I}$, the set $\mathcal{C}_i \subset \mathcal{I}$, with $i \in \mathcal{C}_i$, is the set of *competing platforms* whose choice of filters has an impact on the engagement of platform i.

To simplify the presentation of our results, we consider that for any two platforms $i, k \in \mathcal{I}$, if $i \in \mathcal{C}_k$, then $k \in \mathcal{C}_i$. However, our mechanism can easily be extended to the case of asymmetric competition among social media. Given the set of competing platforms $\mathcal{C}_i$, we can define a *valuation function* of platform i.

Definition 2. The *valuation function* of a social media platform $i \in \mathcal{I}$ is $v_i\big(a_k : k \in \mathcal{C}_i\big) : \mathcal{A}^{|\mathcal{C}_i|} \to \mathbb{R}_{\geq 0}$. It is a decreasing function with respect to a_i and strictly increasing with respect to a_l for all $l \in \mathcal{C}_{-i}$, where $\mathcal{C}_{-i} = \mathcal{C}_i \setminus \{i\}$.

The valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big)$ corresponds to the revenue generated by platform i given the user engagement after all platforms have implemented their filters. A higher value of a_i will result in decreasing user engagement in platform i, and thus their revenue. On the other hand, a higher value of a_l of another competing platform $l \in \mathcal{C}_{-i}$ will result in increasing user engagement, and thus revenue, in platform i.

Next, recall from the discussion in the previous section that filtering of misinformation in a social media platform increases the trust of the users of this platform on common knowledge about democracy. Next, for each platform $i \in \mathcal{I}$, we define the average trust function on common knowledge.

Definition 3. The *average trust function* of the users of

platform $i \in \mathcal{I}$ on common knowledge is $h_i(a_i) : \mathcal{A} \to [0,1]$, and it is a strictly increasing function with respect to a_i.

The average trust function $h_i(a_i)$ captures the impact of filter a_i on the trust on common knowledge across the users of platform i. A low value of $h_i(a_i)$ implies that a_i leads to low trust on common knowledge for the users of platform i, and vice versa. In practice, platform i can measure the opinions expressed by their users [29] through surveys, and over time, use these measurements to estimate the impact of filter a_i using the average trust function $h_i(a_i)$.

B. Misinformation Filtering Game for the Government

Recall that, in our framework, the government is considered the strategic player $0 \in \mathcal{J}$. The government's objective is to maximize the trust of the users of all social media platforms on common knowledge. Therefore, the government selects an action $a_0 \in \mathcal{A} = [0,1]$ that designates a lower bound which must be satisfied by the aggregate average trust of all social media platforms in $\mathcal{I}$. To this end, we refer to the action a_0 as the government's lower bound on trust on common knowledge.

Let $N_i \in \mathbb{N}$ be the total number of users of the social media platform $i \in \mathcal{I}$. Then, the fraction of the number of users of i with respect to the total number of users of all platforms is

$$n_i = \frac{N_i}{\sum_{l \in \mathcal{I}} N_l}. \tag{1}$$

The fraction n_i represents the contribution of users in platform i on the average trust on common knowledge about democracy. Since $\sum_{i \in \mathcal{I}} n_i = 1$, the aggregate average trust on common knowledge is $\sum_{i \in \mathcal{I}} n_i \cdot h_i(a_i)$. In our framework, the government's role is to select the lower-bound a_0 for the aggregate average trust. After the government decides on a_0, each platform $i \in \mathcal{I}$ who decides to participate in the game must select a filter a_i that satisfies the following constraint:

$$a_0 - \sum_{i \in \mathcal{I}} n_i \cdot h_i(a_i) \leq 0. \tag{2}$$

Next, we define the government's valuation as a function of the lower bound on trust a_0.

Definition 4. The *valuation function* of the government is $v_0(a_0) : [0,1] \to \mathbb{R}_{\geq 0}$, and it is an increasing function with respect to the lower bound a_0.

The government's valuation function $v_0(a_0)$ assigns a monetary value on the lower bound a_0. Recall that the government seeks to increase the trust on common knowledge among the users of all social media platforms. Thus, the government's valuation increases as the lower bound on aggregate average trust increases. We also consider that the government might have limited resources available to invest in this problem, i.e., there exists a finite monetary budget $b_0 \in \mathbb{R}_{\geq 0}$ representing the maximum possible expenditure of the government for this problem.

C. Information Structure

In this subsection, we specify the private and public information structure corresponding to each player in the imposed game.

1) Public information: The set of competing platforms $\mathcal{C}_i$ and fraction of users n_i of each platform $i \in \mathcal{I}$ are known to all players in set $\mathcal{J}$. Moreover, the set of feasible actions $\mathcal{A}$ is known to all players in the set $\mathcal{J}$.

2) Valuation functions: The valuation function $v_i(\cdot)$ of each social media platform $i \in \mathcal{I}$ is considered private information, and thus, it is known only to platform i. Similarly, the valuation function $v_0(\cdot)$ and the budget b_0 of the government are private information of the government.

3) Average trust functions: The average trust function $h_i(\cdot)$ of social media platform $i \in \mathcal{I}$ is considered private information, and thus, it is known only to platform i (it is not known to the government).

D. Assumptions

In the modeling framework presented above, we impose the following assumptions:

Assumption 1. For each platform $i \in \mathcal{I}$, $|\mathcal{C}_i| \geq 3$.

We impose this assumption to simplify the exposition of our mechanism. Assumption 1 implies that each user subscribes in multiple social media platforms. It has been shown in the literature that each user, on an average, subscribes to 8 social media platforms [30]. Nevertheless, we present an extension of our mechanism for $|\mathcal{C}_i| \geq 2$ in Appendix A.

Assumption 2. The valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big) : \mathcal{A}^{|\mathcal{C}_i|} \to \mathbb{R}_{\geq 0}$ of each social media platform $i \in \mathcal{I}$ is a concave and differentiable function with respect to a_k.

The concavity of $v_i\big(a_k : k \in \mathcal{C}_i\big)$ captures the diminishing marginal change in engagement due to additional filtering. Practically, the higher the value of a_i, the more users of platform i will perceive the filter as censorship of their opinions. Thus, for platform i, increasing a low-value filter may lead to a lesser loss in engagement as compared to increasing a filter whose value is already high. Nevertheless, to ensure the robustness of our proposed mechanism, we also present an analysis of our system by relaxing Assumption 2 in Section IV-A.

Assumption 3. The average trust function $h_i(a_i) : \mathcal{A} \to [0,1]$ of each social media platform $i \in \mathcal{I}$ is a concave and differentiable function with respect to a_i.

The concavity of $h_i(a_i)$ implies that, for large values of a_i, a small incremental change of a_i would not have a significant impact on the average trust of users on common knowledge. Practically, this implies low values of a_i will have a major impact on the average trust function. Nevertheless, to ensure the robustness of our mechanism, we also present an analysis of our system by relaxing Assumption 3 in Section IV-A.

Assumption 4. The valuation function of the government $v_0(a_0) : [0,1] \to \mathbb{R}_{\geq 0}$ is a concave and differentiable function with respect to the lower-bound a_0.

Practically, for high values of a_0, the government might not be interested in investing additional resources to increase a_0 even more, as the impact on improving common knowledge would not be significant. Nevertheless, we also present an analysis of our system by relaxing Assumption 4 in Section IV-A.

Assumption 5. The output of the function $h_i(a_i)$ can be monitored by any competing platform $l \in \mathcal{C}_{-i}$, and a violation of the condition (2) can be detected by the government.

Assumption 5 helps us enforce the mechanism, which we present in Section III, in a static environment. In the mechanism, each platform $i \in \mathcal{I}$ commits to a minimum value of the average trust function among their users which can be achieved by choosing an appropriate value for a_i. Consider that a platform i selects a value a_i that fails to satisfy this commitment. Practically, the government can detect a violation of (2) by gauging public opinion on the internet and through surveys. However, the government does not know the function $h_i(\cdot)$ of platform i, and thus, would penalize each platform in $\mathcal{I}$ equally for the violation of (2). To avoid the penalty for the failure of platform i, a competing platform $l \in \mathcal{C}_{-i}$ can report the violation of i. Thus, it is reasonable to consider that each platform $i \in \mathcal{I}$ monitors the output $h_l(a_l)$ of each competing platform $l \in \mathcal{C}_{-i}$ to maximize their own utility. We believe that using a dynamic mechanism, we could potentially relax Assumption 5 [26]. This would be a potential direction for future research.

Assumption 6. The government ensures that any social media platform $i \in \mathcal{I}$ that does not participate in the mechanism receives no benefits from the filters of participating social media.

In static mechanisms, the ability to exclude a player from receiving benefits of some common resource is a necessary condition for voluntary participation of players without any monetary investment [31]. This condition is often assumed implicitly in the literature [22]–[25]. In our mechanism, the government can make an investment up to the budget b_0. Thus, we assume *partial excludability* in Assumption 6, where a non-participating platform i still receives the maximum valuation for selecting filter $a_i = 0$, but cannot receive benefits from the filters of any participating platforms. In practice, the government can publicize that platform i has chosen not to contribute in a collective endeavor to filter misinformation. The resulting loss in credibility among the users of the platforms that participate will minimize their migration to platform i. This assumption may be relaxed using a dynamic mechanism, which could be another direction for future research [32].

E. Problem Statement

Since there is a conflict of interest between the government and the social media platforms, we consider that the government hires a social planner to design a mechanism to impose the misinformation filtering game. The mechanism must serve two purposes: (i) incentivize all platforms to voluntarily participate in the game, and (ii) induce a selection of filters that maximizes the *social welfare* of the system. The social welfare of the system is the sum of utilities of all players, formally defined next. To meet these objectives, the social planner asks each player $i \in \mathcal{J}$ to send a message m_i from a set of feasible messages $\mathcal{M}_i$. Based on the message profile $m = (m_0, m_1, \ldots, m_{|\mathcal{I}|})$, the social planner assigns a tax $\tau_i(m) \in \mathbb{R}$ for each social media platform $i \in \mathcal{I}$, and an investment $\tau_0(m) \in \mathbb{R}_{\geq 0}$ for the government. The message and tax of each player is formally defined in Section III-B. By convention, a tax $\tau_i(m) > 0$ is a payment made by player $i \in \mathcal{J}$, and a tax $\tau_i(m) < 0$ is a subsidy given to player i. Thus, the taxes of the platforms can be either monetary payments or subsidies, whereas, the government may never collect a monetary subsidy from any platform. Note that the social planner must not receive any profit, nor incur any losses, for designing and implementing the mechanism, which implies that the mechanism should be budget balanced, i.e., $\sum_{i\in\mathcal{J}} \tau_i(m) = 0$.

Next, we define the utilities of the players.

Definition 5. The *utility* of each platform $i \in \mathcal{I}$ is given by $u_i\big(m, a_k : k \in \mathcal{C}_i\big) := v_i\big(a_k : k \in \mathcal{C}_i\big) - \tau_i(m)$, while the utility of the government is given by $u_0(m, a_0) := v_0(a_0) - \tau_0(m)$.

The social welfare is $u_0(m, a_0) + \sum_{i\in\mathcal{I}} u_i(m, a_k : k \in \mathcal{C}_i)$. The optimization problem for the social planner is to maximize the social welfare, and it is formulated as follows.

Problem 1.

$$\max_{a,\tau(m)} \left(v_0(a_0) - \tau_0(m) + \sum_{i\in\mathcal{I}} \Big(v_i\big(a_k : k \in \mathcal{C}_i\big) - \tau_i(m) \Big) \right), \tag{3}$$

$$\text{subject to: } 0 \leq a_i \leq 1, \quad \forall i \in \mathcal{J}, \tag{4}$$

$$a_0 - \sum_{i\in\mathcal{I}} n_i \cdot h_i(a_i) \leq 0, \tag{5}$$

$$0 \leq \tau_0(m) \leq b_0, \tag{6}$$

$$\sum_{i\in\mathcal{J}} \tau_i(m) = 0, \tag{7}$$

where $a = \big(a_0, a_1 \ldots, a_I\big)$ and $\tau(m) = \big(\tau_0(m), \tau_1(m), \ldots, \tau_{|\mathcal{I}|}(m)\big)$ denote the action and tax profiles of all players, respectively.

In Problem 1, (5) ensures that the aggregate average trust of all users satisfies the government's lower bound a_0, (6) restricts the government's investment $\tau_0(m)$ to be within the available budget, and (7) ensures that the mechanism is budget balanced.

Note that, in Problem 1, the social planner does not have knowledge about the functional form of either the valuation function $v_i(\cdot)$ of any player $i \in \mathcal{J}$, or the average trust function $h_i(\cdot)$ of any platform $i \in \mathcal{I}$. If the social planner knew these functions, then she could solve Problem 1 using standard optimization methods to allocate the optimal filter a_i and tax $\tau_i(m)$ to each platform $i \in \mathcal{I}$, and the optimal lower bound a_0 and investment $\tau_0(m)$ to the government. The objective function of Problem 1 is differentiable and concave,

and the set of feasible solutions is non-empty, convex, and compact. Thus, Problem 1 is a convex optimization problem with a unique optimal solution [33]. However, this solution cannot be computed directly by the social planner because of the private information of the players. Note that if the social planner simply asks the players to report their private information, then the players may not be truthful. Thus, the social planner seeks to design the taxes $\tau_i(m)$ for each player $i \in \mathcal{J}$ to incentivize the players to be truthful while, at the same time, maximizing the social welfare.

Remark 1. The government has a no compelling reason to misreport to the social planner their budget b_0. Thus, we consider that the social planner has knowledge of b_0.

Remark 2. By maximizing the social welfare $u_0(m, a_0) + \sum_{i\in\mathcal{I}} u_i(m, a_k : k \in \mathcal{C}_i)$ in Problem 1, the utility of each player is maximized. Hence, participation of the players in the mechanism is incentivized. Note that the government is not in the position to design the mechanism because they would seek to optimize only their own utility $u_0(m, a_0)$. Thus, the government hires the social planner to design and implement the mechanism described next.

III. Mechanism Design Approach

In this section, we present a two-step mechanism to incentivize filtering of misinformation among social media platforms. The objective of the first step is to ensure that the social media platforms voluntarily agree to participate in the mechanism. The objectives of the second step are to: (i) extract truthful information from the participating platforms, (ii) derive the optimal level of investment for the government, and (iii) design appropriate taxes for the platforms to maximize the social welfare of the system.

A. Step One - The Participation Step

In step one of the mechanism, each social media platform $i \in \mathcal{I}$ must decide whether to participate in the mechanism, with complete knowledge of the rules of the second step of the mechanism described in the next subsection. Consider a platform $i \in \mathcal{I}$ that chooses not to participate in the mechanism. Thus, this platform neither pays taxes nor receives any subsidies from the government, i.e., $\tau_i(m) = 0$. Furthermore, platform i is free to select the lowest value of $a_i = 0$ that maximizes the valuation $v_i\big(a_k : k \in \mathcal{C}_i\big)$. Meanwhile, another competing platform $l \in \mathcal{C}_{-i}$ may decide to participate in the mechanism and subsequently implement a non-zero filter a_l. From Assumption 6, the government ensures that platform l receives no utility as a result of filter a_l. Thus, the utility of the non-participating platform i is given by $v_i\big(a_k = 0 : k \in \mathcal{C}_i\big)$. We will use this utility for a non-participating platform in Theorem 4 of Section IV to establish that all platforms decide to voluntarily participate in step one of the mechanism.

B. Step Two - The Bargaining Step

In step two, the social planner asks each player $i \in \mathcal{J}$ to broadcast a message m_i from a set of feasible messages $\mathcal{M}_i$. For each platform $i \in \mathcal{I}$, let $\mathcal{D}_i = \mathcal{C}_i \cup \{0\}$, and $\mathcal{D}_{-i} = \mathcal{D}_i \backslash \{i\}$. The message of platform i is defined as

$$m_i := (\tilde{h}_i, \tilde{p}_i, \tilde{a}_i), \tag{8}$$

where $\tilde{h}_i \in \mathbb{R}_{\geq 0}$ is the minimum average trust that platform i proposes to achieve through filtering; $\tilde{p}_i \in \mathbb{R}_{\geq 0}^{|\mathcal{D}_{-i}|}$ is the collection of prices that platform i is willing to pay or receive per unit changes in the filters of other competing platforms (except i) and the government's lower bound, given by

$$\tilde{p}_i := (\tilde{p_l}^i : l \in \mathcal{D}_{-i}); \tag{9}$$

and $\tilde{a}_i = (\tilde{a}_k^i : k \in \mathcal{D}_i)$, $\tilde{a}_i \in \mathbb{R}^{|\mathcal{D}_i|}$, is the profile of filters for all competing platforms (including i) and government's lower bound proposed by platform i.

Remark 3. Note that each platform proposes a filter for themselves, denoted by $\tilde{a}_i^i$, in their message m_i. However, as it can be seen in (9), platform i does not propose a price corresponding to $\tilde{a}_i^i$. This is because we want to give every platform the ability to influence their filter, but not the ability to influence the price associated with their own filter.

The message of the government is $m_0 := (\tilde{p}_0, \tilde{a}_0^0)$, where $\tilde{p}_0 \in \mathbb{R}_{\geq 0}$ is the price that the government is willing to pay or receive per unit change of the average trust, and $\tilde{a}_0^0 \in \mathbb{R}$ is the lower bound proposed by the government. Note that our mechanism respects the privacy of each platform $i \in \mathcal{I}$ since she does not request either their valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big)$ or their average trust function $h_i(a_i)$. Similarly, the government is not forced to publicly reveal the functional form of their valuation function $v_0(a_0)$. Also each platform i is free to select any feasible values for the components of the message m_i.

Based on the message profile $m := (m_0, m_1, \ldots, m_{|\mathcal{I}|})$ that the social planner receives, she allocates the following parameters to the players:

1) The social planner allocates a filter to each platform $i \in \mathcal{I}$ and a lower bound to the government such that the constraints of Problem 1 are satisfied. The filter allocated by the social planner to platform i is $\alpha_i(m) := \sum_{k\in\mathcal{C}_i} \frac{\tilde{a}_i^k}{|\mathcal{C}_i|}$, i.e., the average of the filters proposed by all competing platforms including i. The lower bound allocated by the social planner to the government is $\alpha_0(m) = \sum_{k\in\mathcal{J}} \frac{\tilde{a}_0^k}{|\mathcal{J}|}$, i.e., the average of the lower bounds proposed by all platforms and the government.

2) The social planner allocates a minimum average trust $\eta_i(m) \in [0, 1]$ to each platform $i \in \mathcal{I}$, given by

$$\eta_i(m) = \min\left\{ \frac{n_i \cdot \tilde{h}_i}{\sum_{k\in\mathcal{I}} n_k \cdot \tilde{h}_k} \cdot \alpha_0(m),\ 1 \right\}, \tag{10}$$

where the social planner will not accept a message m_i from a platform i that might lead to a situation where $\sum_{k\in\mathcal{I}} n_k \cdot \tilde{h}_k = 0$. The allocated minimum average trust, $\eta_i(m)$, is a lower bound on average trust that must be achieved by platform i. Let the filter implemented by platform i be a_i. Then, platform i must ensure that $n_i \cdot h_i(a_i) \geq \eta_i(m)$. Recall from the information structure that a potential violation of this condition cannot be detected by the social planner since she does not have explicit knowledge of the function $h_i(\cdot)$. However, by

Assumption 5, the output of $h_i(a_i)$ can be monitored by any other competing platform $l \in \mathcal{C}_{-i}$. Any violation of $n_i \cdot h_i(a_i) \geq \eta_i(m)$ will be reported by platform l to the social planner, in order to ensure that platform i implements the largest filter a_i, and maximizes the utility $u_l(m, a_k : k \in \mathcal{C}_l)$. This prevents platforms from violating the constraint imposed by the allocated minimum average trust $\eta_i(m)$.

3) The social planner allocates a price ${\pi_l}^i := \sum_{k \in \mathcal{C}_{-l}: k \neq i} \frac{{\tilde{p}_l}^k}{|\mathcal{C}_l|-2}$, $\pi_l^i \in \mathbb{R}_{\geq 0}$, to each platform $i \in \mathcal{I}$, corresponding to the allocated filter $\alpha_l(m)$ of every other competing platform $l \in \mathcal{C}_{-i}$. This price is derived as the average of prices proposed for the allocated filter $\alpha_l(m)$ by all competing platforms in $\mathcal{C}_{-l}$ except i. Thus, the allocated price ${\pi_l}^i$ is independent of the prices proposed by both platforms i and l. Similarly, the social planner allocates the price $\pi_0 = \sum_{i \in \mathcal{I}} \frac{\tilde{p}_0^i}{|\mathcal{I}|}$ to the government. Note that even though the prices allocated to each player depend on the message profile m, we do not present them with the argument of m to simplify our notation and improve the readability of the subsequent equations.

4) The social planner allocates the following tax to each social media platform $i \in \mathcal{I}$,

$$\tau_i(m) := -\tilde{p}_0 \cdot \eta_i(m) - \sum_{l \in \mathcal{C}_{-i}} {\pi_i}^l \cdot \alpha_i(m) + \sum_{l \in \mathcal{C}_{-i}} {\pi_l}^i \cdot \alpha_l(m) + \sum_{l \in \mathcal{C}_{-i} \cup \{0\}} {\tilde{p}_l}^i \cdot (\tilde{a}_l^i - \tilde{a}_l^{-i})^2, \quad (11)$$

where $\tilde{a}_l^{-i} = \sum_{k \in \mathcal{C}_l : k \neq i} \frac{\tilde{a}_l^k}{|\mathcal{C}_l|-1}$, for each $l \in \mathcal{C}_{-i}$, is the average of the proposed filters for l by all competing platforms except $i \in \mathcal{I}$, and $\tilde{a}_0^{-i} = \sum_{k \in \mathcal{J}_{-i}} \frac{\tilde{a}_0^k}{|\mathcal{J}|-1}$ is the average of lower bounds proposed by all players except i. The tax $\tau_i(m)$ of platform i in (11) can be interpreted as follows: (i) the first term in (11) represents a subsidy given by the government to platform i for the increase in average trust among the users of platform i; (ii) the second term in (11) is a collection of subsidies given by each competing platform $l \in \mathcal{C}_{-i}$ to platform i for the increase in valuation $v_l\big(a_k : k \in \mathcal{C}_l\big)$ due to the allocated filter α_i; (iii) the third term in (11) is a payment by platform i for the increase in valuation $v_i\big(a_k : k \in \mathcal{C}_i\big)$ due to the allocated filter α_l of each competing platform $l \in \mathcal{C}_{-i}$; and (iv) the fourth term in (11) is a collection of penalties to platform i if either the filter proposed in message m_i for any competing platform $l \in \mathcal{C}_{-i}$ is inconsistent with the filters proposed by other platforms, or if the lower bound proposed in m_i is inconsistent with the lower bound proposed by other players. Note that the fourth term also penalizes platform i for higher values of proposed prices ${\tilde{p}_l}^i$ and thus, ensures that the platform i proposes lower prices for the actions of other players.

Finally, the social planner allocates the following investment to the government:

$$\tau_0(m) = \pi_0 \cdot \alpha_0(m) + (\tilde{p}_0 - \pi_0)^2, \quad (12)$$

where the first term is the total investment made by the government for the allocated low bound $\alpha_0(m)$, and the second term is a penalty when the price proposed by the government deviates from the price allocated to the government.

Remark 4. Note that in (11), for some filter $a_i > 0$ of platform i, the social planner takes a payment from each competing platform $l \in \mathcal{C}_{-i}$ and allocates an equal subsidy to platform i. This subsidy serves a dual purpose: (i) it incentivizes platform i to implement the filter a_i, and (ii) it eventually leads to a fair distribution of the government's investment among all platforms.

Remark 5. We presented the step two of the mechanism under the implicit assumption that all social media platforms participate in the mechanism. This does not cause any implications, however, since, as we prove in Theorem 4 next, all platforms eventually, indeed, participate in the mechanism in step one.

The step two of the mechanism is characterized by the tuple $\langle \mathcal{M}, g(\cdot) \rangle$, where $\mathcal{M} = \mathcal{M}_0 \times \mathcal{M}_1 \times \cdots \times \mathcal{M}_{|\mathcal{I}|}$ is the complete message space of all players, and $g(\cdot) : \mathcal{M} \to \mathcal{O}$ is the outcome function that maps each message profile to a set of outcomes $\mathcal{O}$. The set of outcomes is in the form

$$\mathcal{O} := \Big\{ \big(\alpha_0(m), \alpha_1(m), \ldots, \alpha_{|\mathcal{I}|}(m)\big), \big(\tau_0(m), \tau_1(m), \ldots, \tau_{|\mathcal{I}|}(m)\big) : \alpha_i(m) \in \mathcal{A},\ \tau_i(m) \in \mathbb{R},\ i \in \mathcal{J} \Big\}, \quad (13)$$

and the outcome function $g(m)$ determines the outcome of any given message profile $m = (m_0, m_1, \ldots, m_I) \in \mathcal{M}$.

C. Generalized Nash Equilibrium and the Induced Game

Formally, a mechanism $\langle \mathcal{M}, g(\cdot) \rangle$ together with the utility functions $(u_i)_{i \in \mathcal{I}}$ induces a game in which the social planner allocates the filters $(\alpha_1(m), \ldots, \alpha_i(m))$ to the platforms and the lower bound $\alpha_0(m)$ to the government. Each platform $i \in \mathcal{I}$ that participates in the mechanism must implement the filter $a_i = \alpha_i(m)$, and the government must select the lower bound $a_0 = \alpha_0(m)$. Note that platform i can influence their allocated filter $\alpha_i(m)$ with their message m_i. Thus, the strategy of platform i in the induced game is given by the message $m_i \in \mathcal{M}_i$ [17], with a constraint that $\alpha_i(m) \in \mathcal{S}_i(m)$, where

$$\mathcal{S}_i(m) = \{a_i \in \mathcal{A} : n_i \cdot h_i(a_i) \geq \eta_i(m)\}. \quad (14)$$

Thus, the set of feasible allocations $\mathcal{S}_i(m)$ for $i \in \mathcal{I}$ is a function of the messages of all social media in $\mathcal{I}$ and the government. The strategy of the government is denoted by the message m_0 and the set of feasible strategies is given by $\mathcal{M}_0$. For such a game, we select the solution concept of the generalized Nash equilibrium (GNE) [34]. Let $m_{-i} = (m_0, \ldots, m_{i-1}, m_{i+1}, \ldots, m_I)$. A message profile $m^* = (m_i^* : i \in \mathcal{J})$ is the GNE of the induced game, if (i) for each $i \in \mathcal{I}$,

$$u_i\big((m_i^*, m_{-i}^*), \alpha_k(m_i^*, m_{-i}^*) : k \in \mathcal{C}_i\big) \geq u_i\big((m_i, m_{-i}^*), \alpha_k(m_i, m_{-i}^*) : k \in \mathcal{C}_i\big), \quad (15)$$

for all $m_i \in \mathcal{M}_i$ and $\alpha_i \in \mathcal{S}_i(m)$; and (ii) the message m_0^* of the government is such that $u_0\big((m_0^*, m_{-0}^*), \alpha_0(m_0^*, m_{-0}^*)\big) \geq u_0\big((m_0, m_{-0}^*), \alpha_0(m_0, m_{-0}^*)\big)$, for all $m_0 \in \mathcal{M}_0$. To simplify the notation, in the remaining of the paper, we denote the utility of platform $i \in \mathcal{I}$ by $u_i(m_i, m_{-i})$ and the utility of the government by $u_0(m_0, m_{-0})$.

TABLE I
A SUMMARY OF THE KEY VARIABLES

Symbol	Explanation
m_i	The message broadcast by player $i \in \mathcal{J}$
a_i	The filter of platform $i \in \mathcal{I}$
$\tilde{a}_k^i$	The filter proposed by platform $i \in \mathcal{I}$ for platform $k \in \mathcal{C}_i$
$\alpha_i(m)$	The filter allocated to platform $i \in \mathcal{I}$
a_0	The government's lower bound on trust
$\tilde{a}_0$	The lower bound proposed by the government
$\tilde{a}_0^i$	The lower bound proposed by platform $i \in \mathcal{I}$ for the government
$\alpha_0(m)$	The lower bound allocated to the government
$v_i(\cdot)$	The valuation function of player $i \in \mathcal{J}$
$h_i(\cdot)$	The average trust function of platform $i \in \mathcal{I}$
$\tilde{h}_i$	The proposed minimum average trust of platform $i \in \mathcal{I}$
$\eta_i(m)$	The allocated minimum average trust for platform $i \in \mathcal{I}$
$\tilde{p}_l^i$	The price proposed by platform $i \in \mathcal{I}$ corresponding to player $l \in \mathcal{D}_{-i}$
π_l^i	The price allocated to platform $i \in \mathcal{I}$ corresponding to player $l \in \mathcal{D}_{-i}$
$\tilde{p}_0$	The price proposed by the government
π_0	The price allocated to the government
$\tau_i(m)$	The tax allocated to player $i \in \mathcal{J}$

Remark 6. In general, the GNE solution concept is defined for a game with complete information. However, we adopt this solution in our induced game despite the fact that the valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big)$ and the average trust function $h_i(a_i)$ are the private information of platform i. We resolve this discrepancy by considering that the induced game is played repeatedly over multiple iterations, and thus, the social media platforms can utilize an iterative learning process to find a GNE. This interpretation of a GNE is consistent with the theory of mechanism design [35].

D. Summary of the Notation

We summarize the variables introduced in Sections II and III in Table I. As a general guideline, we use lowercase letters of the English alphabet to denote variables and functions, lowercase letters with tilde to denote variables in a message, and lowercase letters of the Greek alphabet to indicate variables allocated to the players by the social planner. We use scripted letters to denote sets. Furthermore, we use $\mathbb{R}$ to denote the set of real numbers, $\mathbb{R}_{\geq 0}$ to denote the set of non-negative real numbers, and $\mathbb{N}$ to denote the set of natural numbers.

IV. PROPERTIES OF THE MECHANISM

In this section, we show that our proposed mechanism has the following desirable properties: (i) budget balance at GNE, (ii) feasibility at GNE, (iii) strong implementation, (iv) existence of at least one GNE, and (v) individual rationality.

Recall that each social media platform $i \in \mathcal{I}$ is a strategic player who seeks to maximize their utility $u_i(m_i, m_{-i})$ through the choice of message $m_i \in \mathcal{M}_i$. Thus, we can define the following optimization problem from the perspective of platform $i \in \mathcal{I}$ in the induced game.

Problem 2. The optimization problem for social media platform $i \in \mathcal{I}$ in the induced game is

$$\max_{m_i \in \mathcal{M}_i} \; v_i\big(\alpha_k(m) : k \in \mathcal{C}_{-i}\big) - \tau_i(m), \tag{16}$$

$$\text{subject to: } 0 \leq \alpha_i(m) \leq 1, \tag{17}$$

$$\eta_i(m) - n_i \cdot h_i\big(\alpha_i(m)\big) \leq 0, \tag{18}$$

where the objective function in (16) is the utility $u_i(m_i, m_{-i})$ of platform i, (17) ensures that the allocated filter of platform i is feasible, and (18) ensures that the fraction of average trust among users of platform i is greater than the minimum average trust allocated by the social planner.

Note that the social planner can ensure that (17) and (18) are hard constraints by imposing a tax $\tau_i(m) \to \infty$ when they are violated. Next, recall that the government is also a strategic player in the induced game who seeks to maximize their utility $u_0(m_0, m_{-0})$ through the choice of message $m_0 \in \mathcal{M}_0$.

Problem 3. The optimization problem for the government is

$$\max_{m_0 \in \mathcal{M}_0} \; v_0\big(\alpha_0(m)\big) - \tau_0(m), \tag{19}$$

$$\text{subject to: } 0 \leq \alpha_0(m) \leq 1, \tag{20}$$

$$\pi_0 \cdot \alpha_0(m) - b_0 \leq 0, \tag{21}$$

where the objective in (19) is the utility $u_0(m_0, m_{-0})$ of the government, (20) ensures that the government's lower bound a_0 is feasible, and (21) ensures that the total government's investment is less than their budget b_0.

Remark 7. Consider an optimal solution $m_i^* \in \mathcal{M}_i$ of Problem 2 for each platform $i \in \mathcal{I}$, and an optimal solution $m_0^* \in \mathcal{M}_0$ of Problem 3 for the government. The message profile $m^* = \big(m_0^*, m_1^*, \ldots, m_{|\mathcal{I}|}^*\big) \in \mathcal{M}$ satisfies (15), and thus, forms a GNE of the induced game.

Next, we establish some basic properties of the mechanism in Lemmas 1 and 2 at any GNE, if one exists. In Lemma 1, we refer to Problem 3 to show that the government's proposed price at any GNE of the induced game is equal to the average price proposed by all social media.

Lemma 1. *Let the message profile $m^* \in \mathcal{M}$ be a GNE of the induced game. Then, ${\tilde{p}_0}^* = \pi_0^*$ for the government.*

Proof. Since the objective function in Problem 3 is concave with respect to the price $\tilde{p}_0$, the price ${\tilde{p}_0}^*$ at GNE can be using the equation $\frac{\partial u_0}{\partial \tilde{p}_0}\big|_{{\tilde{p}_0}^*} = 2 \cdot ({\tilde{p}_0}^* - \pi_0^*) = 0$, which yields ${\tilde{p}_0}^* = \pi_{-0}^*$. □

Similarly, in the next result (Lemma 2), we refer to Problem 2 to establish that, at any GNE, the filters proposed by all social media platforms in $\mathcal{C}_i$ for platform i are the equal, unless the corresponding price proposal is 0. Furthermore, at every GNE, if one exists, the lower bound proposed by all platforms is the same, unless the corresponding price proposal is 0.

Lemma 2. *Let the message profile $m^* \in \mathcal{M}$ be a GNE of the induced game. Then, for ${\tilde{p}_k}^i \neq 0$, we have $\tilde{a}_k^{i*} = \tilde{a}_k^{-i*}$ for every social media platform $i \in \mathcal{I}$, for every $k \in \mathcal{D}_{-i}$.*

Proof. The proof is similar to the proof of Lemma 1, and thus, due to space limitations, it is omitted. □

Next, we use the properties established in Lemmas 1 and 2 to show that our proposed mechanism is budget balanced at any GNE, if one exists, i.e., the social planner redistributes all the payments it collects from the players as subsidies to the players.

Theorem 1 (**Budget Balance**). *Consider any GNE $m^* \in \mathcal{M}$ of the induced game. Then, the proposed mechanism is budget balanced, i.e., $\sum_{i\in\mathcal{J}} \tau_i(m^*) = 0$.*

Proof. From Lemmas 1 and 2, the tax $\tau_i^* = \tau_i(m^*)$ for social media platform i at GNE is $\tau_i^* = -\tilde{p}_0{}^* \cdot \eta_i(m^*) - \sum_{l\in\mathcal{C}_{-i}} \pi_i{}^l \cdot \alpha_i(m^*) + \sum_{l\in\mathcal{C}_{-i}} \pi_l{}^i \cdot \alpha_l(m^*)$. The tax τ_0^* for the government at GNE is $\tau_0^* = \tilde{p}_0{}^* \cdot \alpha_0(m^*)$, where $\tilde{p}_0{}^*$ is the price per unit change on average trust at GNE. Since $\sum_{i\in\mathcal{I}} \eta_i(m) = \alpha_0(m)$, for all $m \in \mathcal{M}$, then at GNE we have $\sum_{i\in\mathcal{J}} \tau_i^* = \sum_{i\in\mathcal{I}} \Big[- \sum_{l\in\mathcal{C}_{-i}} \pi_i{}^l \cdot \alpha_i(m^*) + \sum_{l\in\mathcal{C}_{-i}} \pi_l{}^i \cdot \alpha_l(m^*) \Big] = 0$. □

In the next result (Lemma 3), we establish that every GNE, $m^* \in \mathcal{M}$, if one exists, of the induced game leads to an allocation of filters for the platforms and a lower bound for the government that forms a feasible solution of Problem 1. In other words, every GNE of the induced game ensures that all constraints of Problem 1 are satisfied.

Lemma 3 (**Feasibility**). *Every GNE message profile $m^* \in \mathcal{M}$ leads to a filter profile $\big(\alpha_1(m^*), \ldots, \alpha_{|\mathcal{I}|}(m^*)\big)$ and lower bound $\alpha_0(m^*)$, which is a feasible solution of Problem 1.*

Proof. Every GNE message profile m^* satisfies (17) - (18) and (20) - (21). From Theorem 1, $\sum_{i\in\mathcal{J}} \tau_i(m^*) = 0$. For each $i \in \mathcal{I}$, $\eta_i(m) \le n_i \cdot h_i(\alpha_i(m))$, and $\sum_{i\in\mathcal{I}} \eta_i(m) = \alpha_0(m)$. Hence, $\sum_{i\in\mathcal{I}} h_i(\alpha_i(m)) \ge \alpha_0(m)$. □

In the next result (Lemma 4), we establish that every social media platform $i \in \mathcal{I}$ can unilaterally deviate in the message $m_i \in \mathcal{M}_i$, to achieve any desired allocation of filters for every competing platform, including itself. This property of our mechanism ensures that each platform $i \in \mathcal{I}$ can attain any filter $\hat{a}_i \in \mathcal{A}$, irrespective of the filters proposed by the competing platforms.

Lemma 4. *Given the message profile $m_{-i} \in \mathcal{M}_{-i}$, the social media platform $i \in \mathcal{I}$ can unilaterally deviate in their message $m_i \in \mathcal{M}_i$ to attain any filter $\hat{a}_k \in \mathcal{A}$ as the allocated filter $\alpha_k(m) \in \mathcal{S}_k(m)$, for all $k \in \mathcal{C}_i$.*

Proof. Let $m_{-i} = \big(m_0, \ldots, m_{i-1}, m_{i+1}, \ldots, m_{|\mathcal{I}|}\big)$ be the message profile of all players in $\mathcal{J}_{-i}$. Then, platform i can propose a filter $\tilde{a}_k^i = \hat{a}_k - \sum_{l\in\mathcal{C}_k : l\neq i} \frac{\tilde{a}_k^l}{|\mathcal{C}_k|-1}$, to ensure that $\alpha_k(m) = \hat{a}_k$ for each $k \in \mathcal{C}_i$. Moreover, platform i can propose a lower bound $\tilde{a}_0^i = -\sum_{l\in\mathcal{J}_{-i}} \tilde{a}_0^l$ for the government, to ensure that $\alpha_0(m) = 0$, and subsequently, $\alpha_k(m) = \hat{a}_k \in \mathcal{S}_k(m)$ for all $k \in \mathcal{C}_i$. □

Next, we establish that, at any GNE, if one exists, of the induced game the allocated filters for all platforms and the allocated lower bound for the government result in the optimal solution of Problem 1.

Theorem 2 (**Strong Implementation**). *Consider any GNE $m^* \in \mathcal{M}$ of the induced game. Then, the allocated filter profile $\big(\alpha_1(m^*), \ldots, \alpha_{|\mathcal{I}|}(m^*)\big)$ and the allocated lower bound $\alpha_0(m^*)$ at equilibrium is equal to the optimal solution a^{*o} of Problem 1.*

Proof. Let $\alpha(m^*) = \big(\alpha_1(m^*), \ldots, \alpha_{|\mathcal{I}|}(m^*)\big)$. Then, the GNE message profile m^* satisfies, for platform $i \in \mathcal{I}$, the following Kush-Kahn-Tucker (KKT) conditions for optimality:

$$\frac{\partial v_i}{\partial \alpha_i}\bigg|_{\alpha(m^*)} + \sum_{l\in\mathcal{I}_{-i}} \pi_l^i - \lambda_i^i + \mu_i^i + \nu_i^i \cdot \frac{\partial h_i}{\partial \alpha_i}\bigg|_{\alpha(m^*)} = 0, \tag{22}$$

$$\frac{\partial v_i}{\partial \alpha_l}\bigg|_{\alpha(m^*)} - \pi_l{}^i = 0, \quad \forall l \in \mathcal{C}_{-i}, \tag{23}$$

$$\tilde{p}_0{}^* - \nu_i^i = 0, \tag{24}$$

$$\lambda_i^i \cdot \big(\alpha_i(m^*) - 1\big) = 0, \tag{25}$$

$$\mu_i^i \cdot \alpha_i(m^*) = 0, \tag{26}$$

$$\nu_i^i \cdot \big(\eta_i(m^*) - h_i(\alpha_i(m^*))\big) = 0, \tag{27}$$

$$\lambda_i^i, \mu_i^i, \nu_i^i \ge 0, \tag{28}$$

where (22) - (24) are the derivatives of the Lagrangian of platform i with respect to $\alpha(m)$ and $\eta_i(m)$, for Problem 2, and (25) - (28) are constraints on the Lagrange multipliers $(\lambda_i^i, \mu_i^i, \nu_i^i)$. From (24), $\nu_i^i = \tilde{p}_0{}^*$ for all $i \in \mathcal{I}$. Substituting (23) in (22), we have

$$\sum_{k\in\mathcal{C}_i} \frac{\partial v_k}{\partial \alpha_i}\bigg|_{\alpha(m^*)} - \lambda_i^i + \mu_i^i + \nu_i^i \cdot \frac{\partial h_i}{\partial \alpha_i}\bigg|_{\alpha(m^*)} = 0, \tag{29}$$

for all $i \in \mathcal{I}$. Similarly, the KKT conditions for Problem 3 are:

$$\frac{\partial v_0}{\partial \alpha_0}\bigg|_{\alpha_0(m^*)} - \tilde{p}_0{}^* - \lambda_0^0 + \mu_0^0 + \omega_0^0 \cdot \tilde{p}_0{}^* = 0, \tag{30}$$

$$\lambda_0^0 \cdot \big(\alpha_0(m^*) - 1\big) = 0, \tag{31}$$

$$\mu_0^0 \cdot \alpha_0(m^*) = 0, \tag{32}$$

$$\omega_0^0 \cdot \big(\tilde{p}_0{}^* \cdot \alpha_0(m^*) - b_0\big) = 0, \tag{33}$$

$$\lambda_0^0, \mu_0^0, \omega_0^0 \ge 0, \tag{34}$$

where (30) is the derivative of the Lagrangian, and (31) - (34) are constraints on the Lagrange multipliers $(\lambda_0^0, \mu_0^0, \omega_0^0)$.

The optimal solution $a^{*o} = \big(a_0^{*o}, a_1^{*o}, \ldots, a_{|\mathcal{I}|}^{*o}\big)$ of Problem 1 satisfies the following KKT conditions:

$$\sum_{k\in\mathcal{C}_i} \frac{\partial v_k}{\partial a_i}\bigg|_{a_i^{*o}} - \lambda_i + \mu_i + \nu \cdot \frac{\partial h_i}{\partial a_i}\bigg|_{a_i^{*o}} = 0, \quad \forall i \in \mathcal{I}, \tag{35}$$

$$\frac{\partial v_0}{\partial a_0}\bigg|_{a_0^{*o}} - \lambda_0 + \mu_0 - \nu - \omega \cdot \pi_0 = 0, \tag{36}$$

$$\lambda_i \cdot (a_i^{*o} - 1) = 0, \qquad \forall i \in \mathcal{J}, \tag{37}$$

$$\mu_i \cdot a_i^{*o} = 0, \qquad \forall i \in \mathcal{J}, \tag{38}$$

$$\nu \cdot \big(a_0^{*o} - h_i(a_i^{*o})\big) = 0, \tag{39}$$

$$\omega \cdot (\pi_0 \cdot a_0^{*o} - b_0) = 0, \tag{40}$$

$$\lambda_i, \mu_i, \omega, \nu \ge 0, \qquad \forall i \in \mathcal{J}, \tag{41}$$

where (35) - (36) are the derivatives of the Lagrangian, and (37) - (38) are constraints on the Lagrange multipliers $(\lambda_i, \mu_i, \omega, \nu : i \in \mathcal{J})$. By setting $\pi_0 = \tilde{p}_0{}^*$, $\lambda_i = \lambda_i^i$, $\mu_i = \mu_i^i$, $\nu = \tilde{p}_0{}^*$, $\omega = \omega_0^0$, $a_i^{*o} = \alpha_i(m^*)$, which implies that the efficient allocation of filters for all platforms and lower bound for the government is implemented by all GNE of the induced game. □

Next, we show that our mechanism guarantees the existence of at least one GNE for the induced game. This ensures that the results of Lemmas 1 - 3 and Theorems 1 - 2 are always valid for the induced game.

Theorem 3 (GNE existence). *Let $a^{*o} = \big(a_0^{*o}, a_1^{*o}, \ldots, a_{|\mathcal{I}|}^{*o}\big)$ be the unique optimal solution of Problem 1. Then, there is a GNE message profile $m^* \in \mathcal{M}$ of the induced game that guarantees that the filter profile $\big(\alpha_1(m^*), \ldots, \alpha_{|\mathcal{I}|}(m^*)\big)$ and lower bound $\alpha_0(m^*)$ at GNE satisfy $\alpha_i(m^*) = a_i^{*o}$, for all $i \in \mathcal{J}$.*

Proof. Consider that the optimal solution a^{*o} which satisfies the KKT conditions for Problem 1 with the corresponding Lagrange multipliers $(\lambda_i, \mu_i, \nu, \omega : i \in \mathcal{J})$. Taking similar steps to the proof of Theorem 2, we can show that for $\tilde{p}_0 = \pi_0 = \nu$, the Lagrange multipliers of Problems 2 and 3 are $\lambda_i^i = \lambda_i$, $\mu_i^i = \mu_i$, $\nu_i^i = \nu$, $\omega_0^0 = \omega$, $i \in \mathcal{J}$, and the allocated prices are $\pi_l{}^i = \frac{\partial v_i}{\partial \alpha_l}\big|_{a^{*o}}$, for all $l \in \mathcal{C}_{-i}$. This implies that the allocated filters at GNE are $\alpha_i(m^*) = a_i^{*o}$ for all platforms $i \in \mathcal{I}$, and the allocated lower bound of the government is $\alpha_0(m^*) = a_0^{*o}$. □

Next, we consider the step one (the participation step) of our mechanism from Section III-A. We first note that the government always participates in the mechanism for the opportunity to incentivize misinformation filtering among the platforms. In the following result (Theorem 4), we invoke Assumption 6 and the properties of our mechanism, to show that in step one, every social media platform voluntarily decides to participate in the mechanism. This property is also called individual rationality of the mechanism as it ensures voluntary participation of rational players without dictatorship.

Theorem 4 (Individually Rational). *The proposed mechanism is individually rational, i.e., each platform $i \in \mathcal{I}$ prefers the outcome of every GNE of the induced game to the outcome of not participating.*

Proof. Consider any GNE message profile m^*. By Lemma 4, given profile m_{-i}^*, there exists a message $m_i \in \mathcal{M}_i$ for platform i such that $\alpha_0(m_i, m_{-i}^*) = 0$. Furthermore, platform i can unilaterally deviate in their message m_i to ensure that for every platform $k \in \mathcal{C}_i$, the allocated filter is given by $\alpha_k(m_i, m_{-i}^*) = 0$. Assumption 6 implies that the utility of a non-participating platform $i \in \mathcal{I}$ is given by $v_i(a_k = 0 : k \in \mathcal{C}_i)$. Consider the message $m_i = (\tilde{h}_i, \tilde{p}_i, \tilde{a}_i)$ defined in (8) with $\tilde{p}_l{}^i = 0$, for all $l \in \mathcal{C}_{-i} \cup \{0\}$, $\tilde{a}_k^i = -\sum_{l \in \mathcal{C}_{-i}} \tilde{a}_k^l$, for all $k \in \mathcal{C}_{-i}$, and $\tilde{a}_0^i = -\sum_{l \in \mathcal{J}_{-i}} \tilde{a}_0^i$. Then, the allocation $\alpha_k(m_i, m_{-i}^*) = 0$ is feasible for every platform $k \in \mathcal{C}_i$ and the corresponding tax for social media platform i is given by $\tau_i = 0$. The utility $u_i(m_i, m_{-i}^*)$ of social media platform i is given by $u_i(m_i, m_{-i}^*) = v_i(0, \ldots, 0) - 0$. From the definition of the GNE in (15), we have $u_i(m^*) \geq u_i(m_i, m_{-i}^*)$. Hence, $u_i(m^*) \geq v_i(0, \ldots, 0)$. We observe that the utility $u_i(m^*)$ at any GNE $m^* \in \mathcal{M}$ of a platform $i \in \mathcal{I}$, that decides to participate in the mechanism, is equal to or greater than their utility when not participating in the mechanism. Thus, in step one of the mechanism, the weakly dominant action of every social media platform $i \in \mathcal{I}$ is to participate in the mechanism. □

A. Extension to Quasi-Concave Valuations

In this subsection, we relax Assumptions 2 - 4, and replace them with the following more general assumptions: (i) The valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big) : \mathcal{A}^{|\mathcal{C}_i|} \to \mathbb{R}_{\geq 0}$ of every platform $i \in \mathcal{I}$ is quasi-concave, differentiable, and have the same monotonic properties as before. (ii) The valuation function $v_0(a_0) : \mathcal{A} \to \mathbb{R}_{\geq 0}$ of the government is quasi-concave, differentiable and increasing with respect to a_0. (iii) The average trust function $h_i(a_i) : \mathcal{A} \to [0, 1]$ of any social media platform $i \in \mathcal{I}$ is a differentiable and increasing with respect to a_i. We cannot use the KKT conditions to prove the existence of a GNE and strong implementation under these relaxed assumptions. However, note that at any GNE, if one exists, the proposed mechanism is still budget balanced, feasible and individually rational. In addition, Lemmas 1, 2, and 4 also hold as they do not depend on the concavity of the valuation.

Next, we prove that for the relaxed assumptions, there exists a GNE and that it induces a Pareto efficient equilibrium in the game. Pareto efficiency refers to the condition where we cannot improve the utility of any player without decreasing the utility of another player in the induced game [17]. Pareto efficiency is a weaker property in comparison to the strong implementation achieved by our mechanism for concave valuation functions.

Theorem 5. *Let the valuation function $v_i(a_k : k \in \mathcal{C}_i)$ be quasi-concave and differentiable for all players $i \in \mathcal{J}$ and consider the game $\langle \mathcal{M}, g(\cdot), (u_i)_{i \in \mathcal{I}} \rangle$. Then, (i) there exists a GNE for the induced game, and (ii) every GNE of the induced game is Pareto efficient.*

Proof. 1) Existence: Consider the social media platform $i \in \mathcal{I}$. Lemma 2 implies that at GNE, the message m_i must lie in the set $\mathcal{M}_i' := \big\{m_i \in \mathcal{M}_i : \tilde{p}_l^i \cdot (\tilde{a}_l^i - a_l^{-i}) = 0,\ \forall l \in \mathcal{D}_{-i}\big\}$. For all $m_i \in \mathcal{M}_i'$, we can write the utility $u_i(m)$ as

$$u_i(m) = v_i(\alpha_k(m) : k \in \mathcal{C}_i) + \tilde{p}_0 \cdot \eta_i(m) + \sum_{l \in \mathcal{C}_{-i}} \pi_i{}^l \cdot \alpha_i(m) - \sum_{l \in \mathcal{C}_{-i}} \pi_l{}^i \cdot \alpha_l(m), \quad (42)$$

where the prices $\tilde{p}_0$, $\pi_i{}^l$, and $\pi_l{}^i$ for any $l \in \mathcal{C}_{-i}$ are independent of message m_i. We observe that $u_i(m) = u_i(\eta_i, \alpha_k : \alpha_k \in \mathcal{D}_i)$. Lemma 4 implies that given a message profile m_{-i} of all platforms and the government in $\mathcal{J}_{-i}$, platform i can unilaterally deviate in their message $m_i \in \mathcal{M}_i$ to receive any allocation $\alpha_k(m) \in \mathcal{A}$, for all $k \in \mathcal{D}_i$. Thus, instead of the message m_i, we equivalently consider that the action of platform i is to select the tuple $\beta_i = \big(\eta_i, \alpha_k : k \in \mathcal{D}_i\big)$, that

takes values in the set $\mathcal{B}_i = \big\{[0,1] \times \mathcal{A}^{|\mathcal{D}_i|} : n_i \cdot h_i(\alpha_i) - \eta_i \geq 0\big\}$. For the differentiable function $h_i(a_i)$, the set $\mathcal{B}_i$ is convex, compact, and independent of the message profile m_{-i}. Similarly, the action of the government α_0 takes values in the set $\mathcal{A}$ that is compact, convex, and independent of the message profile m_{-0}.

Let the valuation $v_i\big(a_k : k \in \mathcal{C}_i\big)$ for every platform $i \in \mathcal{I}$ be quasi-concave and differentiable, and let $\beta = \big(\beta_0, \beta_1, \ldots, \beta_{|\mathcal{I}|}\big)$. Then, for every $i \in \mathcal{I}$, the utility $u_i(\beta)$ in (42) is also quasi-concave and differentiable with respect to the action $\beta_i \in \mathcal{B}_i$. A similar argument implies that the government's utility $u_0(\alpha_0)$ is quasi concave and differentiable with respect to their action α_0. Hence, it follows from Glicksberg's theorem that there exists a Nash Equilibrium (NE) for the induced game [36]; since, by definition, any NE is also a GNE, it follows that there exists a GNE for the induced game.

2) Pareto efficiency: It is sufficient in our case to show that the NE can be characterized by a Walrasian equilibrium as all Walrasian equilibria are Pareto efficient [17]. So, as in part 1, consider an arbitrary NE action profile $\beta^* = \big(\alpha_0^*, \beta_1^*, \ldots, \beta_{|\mathcal{I}|}^*\big)$ that takes values in the set $\mathcal{A} \times \mathcal{B}_1 \times \cdots \times \mathcal{B}_{|\mathcal{I}|}$. From the definition of the NE, for every platform $i \in \mathcal{I}$ it holds that

$$u_i(\beta^*) \geq u_i(\beta_i, \beta_{-i}^*), \quad \forall \beta_i \in \mathcal{B}_i. \tag{43}$$

Note that the NE prices $\tilde{p_0}^*$, π_l^{*i}, ${\pi_l}^{*i}$, for all $l \in \mathcal{I}_{-i}$ cannot be influenced by platform i, i.e., every social media platform is a price taker. Then, using the definition of the NE in (43) with the utility $u_i(m)$ in (42), we can write for platform i that

$$\beta_i^* = \arg\max_{\beta_i \in \mathcal{B}_i} \Bigg\{ v_i(\alpha_k : k \in \mathcal{C}_i) + \tilde{p_0}^* \cdot \eta_i + \sum_{l \in \mathcal{I}_{-i}} {\pi_i}^{*l} \cdot \alpha_i - \sum_{l \in \mathcal{I}_{-i}} {\pi_l}^{*i} \cdot \alpha_l \Bigg\}. \tag{44}$$

Similarly, the government also behaves as a price taker because it cannot influence the NE price π_0^*. For the government at NE, we can write that

$$\alpha_0^* = \arg\max_{\alpha_0 \in \mathcal{A}} \{v_0(\alpha_0) - \pi_0^* \cdot \alpha_0\}. \tag{45}$$

It follows immediately that the NE action profile β^* constitutes a Walrasian equilibrium and thus, the NE for the induced game forms a Pareto efficient equilibrium [17]. Since any NE is also a GNE by definition, it follows that every GNE of the induced game is Pareto efficient. □

Remark 8. The GNE induced by our mechanism may not lead to allocated filters for platforms and lower bound for the government that form an optimal solution of Problem 1 using quasi-concave valuations. However, Theorem 5 establishes that a GNE still exists for such a system, and that it leads to a Pareto efficient allocation, where no player's utility can be improved without decreasing the utility of another player. Thus, from Theorem 4, we can conclude that for quasi-concave valuation functions, our mechanism incentivizes some misinformation filtering but may lead to suboptimal social welfare.

V. Discussion

A. Interpretation of the Results

In this subsection, we present an explanation of the mechanism presented in Section III and the main results derived in Section IV. The social planner seeks to design an efficient mechanism with the following two properties: (i) it should induce voluntary participation among all social media platforms, and (ii) it should maximize the social welfare, i.e., maximize the sum of utilities of all players. Note that the social welfare increases as the valuation function $v_0(a_0)$ increases, which, in turn, increases with respect to the lower bound on aggregate average trust, a_0. A sufficiently high lower bound a_0 indirectly ensures that some platforms implement non-zero filters to raise the average trust of their users. Thus, a mechanism that satisfies properties (i) and (ii) also incentivizes platforms to implement filtering, conditional on the government's valuation $v_0(a_0)$ and budget b_0 being sufficiently large. The challenge faced by the social planner is to achieve these properties without knowledge of the valuation function $v_0(a_0)$ of the government, the valuation function $v_i\big(a_k : k \in \mathcal{C}_i\big)$ of any platform $i \in \mathcal{I}$, and the average trust function $h_i(a_i)$ of any social media platform $i \in \mathcal{I}$.

To meet this challenge, we present a two-step mechanism in Section III. In the step one (the participation step) of the mechanism, the social planner asks each social media platform to decide whether they wish to participate in the mechanism. This is an essential question because the government is not dictatorial, i.e., it cannot force platforms to participate in the mechanism. By refusing to participate in the mechanism, platform i can select no filter and pay no tax. However, platform i also receives no subsidy from the government, nor benefits from the filters of platforms that do participate. We prove in Theorem 4 of Section IV that the utility of any platform $i \in \mathcal{I}$ after participating in the mechanism is greater than or equal to their utility when they do not participate. Thus, the weakly dominant action of every platform in step one is to participate in the mechanism, establishing property (i).

In the step two (the bargaining step) of the mechanism, the social planner asks each player $i \in \mathcal{J}$ to broadcast a message $m_i \in \mathcal{M}_i$. Based on the message profile $m = \big(m_0, m_1, \ldots, m_{|\mathcal{I}|}\big)$, the social planner allocates a minimum average trust $\eta_i(m)$, a filter $\alpha_i(m)$, and a tax $\tau_i(m)$ to each platform $i \in \mathcal{I}$. Similarly, she allocates a lower bound $\alpha_0(m)$ and tax $\tau_0(m)$ to the government. By participating in the mechanism in step one, each player $i \in \mathcal{J}$ agrees to implement the allocated filters, and either pay or receive the allocated tax. The rules defined by the social planner induce a game among the players whose equilibrium is defined as a GNE. The structure of the messages, and various parameters allocated by the social planner lead to the properties of the mechanism in Section IV.

We derive most of the properties of the mechanism in Section IV for a state where the platforms and the government are at a GNE. Lemmas 1 and 2 establish preliminary properties of the tax functions $\tau_i(m)$ of each player $i \in \mathcal{I}$. They show that at the GNE, each player i has to be consistent in their

message m_i with respect to the messages of other players. This consistency check ensures that no player can benefit from a manipulation of the mechanism by proposing arbitrary prices, filters, or lower bounds. Then, we use the results of Lemmas 1 and 2 to derive Theorem 1, which proves that at any GNE the mechanism is budget balanced, i.e., the sum of all taxes is 0. This is a desirable property for the mechanism because the social planner is now guaranteed to simply take the investment of the government $\tau_0(m)$ and redistribute it among the social media platforms, without worrying about leftover funds or insufficient funds. Next, we show in Lemma 3 that every GNE of the induced game is a feasible solution to the problem of maximizing social welfare. Lemma 4 proves that any social media platform $i \in \mathcal{I}$ can always achieve any desired filter in $\mathcal{A}$, including 0, by selecting an appropriate message $m_i \in \mathcal{M}_i$. This property holds irrespective of the messages selected by the other players in $\mathcal{J}_{-i}$, and establishes that a participating platform has a free choice to control their allocated filter.

All preceding results allow us to prove in Theorem 2 that every GNE of the induced game maximizes the social welfare of the system. In Theorem 3 we prove that the induced game is guaranteed to have at least one GNE. Theorem 2 and Theorem 3, together, imply that the mechanism maximizes the social welfare of the system, establishing property (ii). Thus, we have shown that our mechanism does, indeed, incentivize platforms to filter misinformation.

Finally, in Section IV-A we consider quasi-concave valuation functions for all players to relax some of our assumptions. In Theorem 5, we establish that the induced game is still guaranteed to have a GNE, and that it is Pareto efficient. Thus, we observe that our mechanism still incentivizes some amount of filtering, but may lead to suboptimal social welfare.

B. An Example

In this subsection, we present a descriptive example of how our proposed mechanism may play out in a realistic setting. Consider three major social media platforms: Facebook, Twitter, and Reddit. These platforms allow users from different socioeconomic and political backgrounds to obtain the latest news. Typically, users access either Facebook, Twitter, or Reddit via their smartphone app and engage with them by scrolling down, liking, or sharing posts that feature news and personal opinions. The amount of time spent by all users on the platform and the number of actions taken by them collectively define the engagement generated by the platform [7], [13].

As user engagement is a primary driver of advertisement revenue, Facebook, Twitter, and Reddit regularly optimize their post recommendation algorithms to maximizing user engagement. Over time, these algorithms have evolved to promote posts with a high chance of generating engagement among users, without accounting for their impact on the opinions of the users [3]. This has led to the formation of echo chambers, or opinion bubbles among many users, where they repeatedly interact only with posts that align with their own biases on any topic. For many users, their prior biases lead to a repeated exposure to misinformation and conspiracy theories [37]. This causes uncertainty among them regarding the integrity democratic institutions [9]–[12]. For example, misinformation during elections reduces people's faith in the fairness of the election results [6], and misinformation about precautions during a pandemic reduces people's trust in public health experts [38].

A democratic government can observe the trust of the country's citizens from the opinions expressed by them on various social media platforms. When the government realizes the impact of misinformation on the trust of the citizens, they seek to implement policies to minimize the spread of misinformation. In practice, each social media platform can filter misinformation by either flagging posts with inaccurate information, following them up with truthful posts, or simply not recommending them to users. However, filtering misinformation is an expensive undertaking for platforms because of (i) the high investment required to identify inaccurate information [39], and (ii) potential decrease in engagement of users who are censored [14]. Thus, the government decides to allocate a fixed budget for the problem, and appoints an independent agency to design appropriate incentives for Facebook, Twitter, and Reddit, while staying within the budget.

The agency presents the rules of our mechanism to the government, and confirms the government's participation. Then, the agency reveals the rules of the mechanism to the platforms, and announces that platforms who choose not to participate in this collaborative effort will be labelled as non-cooperative. Furthermore, the agency assures the three platforms that they need not reveal private information and that they can choose to avoid filtering misinformation even after participating in the mechanism (Lemma 4). These factors ensure that each platform participates voluntarily in the mechanism (Theorem 4). Then, the agency asks each of Facebook, Twitter, and Reddit to propose a minimum threshold to which they will raise the trust of their users in democratic institutions. The tax incentives given to each platform will be proportional to this threshold. Simultaneously, the agency asks the government to propose a minimum acceptable level for the average of all platforms' thresholds. The government's investment will be proportional to this minimum average. The agency also asks each platform to propose various filters and prices they are willing to pay or receive for the proposed filters. Similarly, the government proposes a price for their proposed minimum average.

The agency then publicly reveals all proposals and transparently uses the rules of the mechanism to assign a potential subsidy/payment, and potential filter to each platform. Similarly, she assigns a potential amount of investment and minimum average to the government. These assignments become binding only if all stakeholders, Facebook, Twitter, Reddit, and the government, accept the assignments. If any stakeholder is dissatisfied, the agency asks all of them to change their proposals and resubmit. This process is repeated until all the stakeholders reach a consensus. The mechanism ensures that such a consensus exists (Theorem 3) and that it is the best possible result for all stakeholders (Theorem 2). As long as the government is sufficiently committed to addressing the problem of misinformation, the mechanism ensures that at the consensus, the platforms will agree to implement misinformation

filters. The allocations become binding on all stakeholders, and the independent agency collects the government's investment. This investment is paid out to each of Facebook, Twitter, and Reddit as a subsidy, only after they achieve the binding level of filtering.

VI. Conclusions and Future Work

Our primary goal in this paper was to design a mechanism to induce a GNE solution in the misinformation filtering game, where (i) each platform agrees to participate voluntarily, and (ii) the collective utility of the government and the platforms is maximized. We designed a mechanism and proved that it satisfies these properties along with budget balance. We also presented an extension of the mechanism with weaker technical assumptions.

Ongoing work focuses on improving the valuation and average trust functions of the social media platforms based on data. We also consider incorporating uncertainty in a platform's estimates of the impact of their filter. These refinements of the modeling framework will allow us to make our mechanism more practical for use in the real world.

Future research should include extending the results of this paper to a dynamic setting in which the social media platforms react in real-time to the proposed taxes/subsidies. In particular, someone could develop an algorithm that the players can use to iteratively arrive at the Nash equilibrium. In such an algorithm, the social planner can receive additional information from the players while they iteratively learn the GNE. Then, she can use this information to change her allocations dynamically, allowing us to relax either Assumption 5 on monitoring of average trust, or Assumption 6 on the excludability of the platforms.

References

[1] W. Davies, "The age of post-truth politics," *The New York Times*, vol. 24, p. 2016, 2016.

[2] J. Cone, K. Flaharty, and M. J. Ferguson, "Believability of evidence matters for correcting social impressions," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 9802–9807, 2019.

[3] Z. Tufekci, "Youtube, the great radicalizer," *The New York Times*, vol. 10, 2018.

[4] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[5] J. Weedon, W. Nuland, and A. Stamos, "Information operations and facebook," *Retrieved from: https://fbnewsroomus. files. wordpress. com/2017/04/facebook-and-information-operations-v1. pdf*, 2017.

[6] H. Farrell and B. Schneier, "Common-knowledge attacks on democracy," *Berkman Klein Center Research Publication*, no. 2018-7, 2018.

[7] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[8] O. Analytica, "Russia will deny cyberattacks despite more us evidence," *Emerald Expert Briefings*, no. oxan-db, 2018.

[9] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Science vs conspiracy: Collective narratives in the age of misinformation," *PloS one*, vol. 10, no. 2, p. e0118093, 2015.

[10] E. Brown, "Propaganda, misinformation, and the epistemic value of democracy," *Critical Review*, vol. 30(3–4), pp. 194–218, 2018.

[11] J. A. Tucker, Y. Theocharis, M. E. Roberts, and P. Barberá, "From liberation to turmoil: Social media and democracy," *Journal of Democracy*, vol. 28(4), pp. 46–59, 2017.

[12] A. Sternisko, A. Cichocka, and J. J. Van Bavel, "The dark side of social movements: Social identity, non-conformity, and the lure of conspiracy theories," *Current opinion in psychology*, vol. 35, pp. 1–6, 2020.

[13] R. Jaakonmäki, O. Müller, and J. Vom Brocke, "The impact of content, context, and creator on user engagement in social media marketing," *Proceedings of the 50th Hawaii international conference on system sciences*, 2017.

[14] O. Candogan and K. Drakopoulos, "Optimal signaling of content accuracy: Engagement vs. misinformation," *Operations Research*, vol. 68, no. 2, pp. 497–515, 2020.

[15] E. A. Vogels, A. Perrin, and M. Anderson. (2020) Most americans think social media sites censor political viewpoints. [Online]. Available: https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-si

[16] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[17] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic theory*. Oxford University Press, 1995.

[18] A. Dave and A. Malikopoulos, "The prescription approach to decentralized stochastic control with word-of-mouth communication," *arXiv e-prints*, p. arXiv:1907.12125, Sep 2019.

[19] A. Mahajan, N. C. Martins, M. C. Rotkowitz, and S. Yüksel, "Information structures in optimal decentralized control," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE, 2012, pp. 1291–1306.

[20] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," *IEEE Transactions on Automatic Control*, vol. 58, no. 7, pp. 1644–1658, 2013.

[21] A. A. Malikopoulos, C. G. Cassandras, and Y. J. Zhang, "A decentralized energy-optimal control framework for connected automated vehicles at signal-free intersections," *Automatica*, vol. 93, no. April, pp. 244–256, 2018.

[22] S. Sharma and D. Teneketzis, "Local public good provisioning in networks: A Nash implementation mechanism," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 11, pp. 2105–2116, 2012.

[23] A. Sinha and A. Anastasopoulos, "Generalized proportional allocation mechanism design for multi-rate multicast service on the internet," *51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 146–153, 2013.

[24] A. Kakhbod and D. Teneketzis, "An efficient game form for unicast service provisioning," *IEEE Transactions on Automatic Control*, vol. 57, no. 2, pp. 392–404, 2011.

[25] R. Jain and J. Walrand, "An efficient nash-implementation mechanism for network resource allocation," *Automatica*, vol. 46(8), pp. 1276–1283, 2010.

[26] M. Zhang and J. Huang, "Efficient network sharing with asymmetric constraint information," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1898–1910, 2019.

[27] I. V. Chremos and A. A. Malikopoulos, "A Socially-Efficient Emerging Mobility Market," *arXiv preprint arXiv:2011.14399*, 2020.

[28] I. V. Chremos and A. Malikopoulos, "Social resource allocation in a mobility system with connected and automated vehicles: A mechanism design problem," *arXiv preprint arXiv:1909.13122*, 2019.

[29] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France," *New media & society*, vol. 16, no. 2, pp. 340–358, 2014.

[30] Datareportal. (2020) Social media users by platforms. [Online]. Available: https://datareportal.com/social-media-users

[31] T. Saijo and T. Yamato, "Fundamental impossibility theorems on voluntary participation in the provision of non-excludable public goods," *Review of Economic Design*, vol. 14, no. 1-2, pp. 51–73, 2010.

[32] F. Farhadi, H. Tavafoghi, D. Teneketzis, and S. J. Golestani, "An efficient dynamic allocation mechanism for security in networks of interdependent strategic agents," *Dynamic Games and Applications*, vol. 9(4), pp. 914–941, 2019.

[33] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[34] F. Facchinei and C. Kanzow, "Generalized nash equilibrium problems," *Annals of Operations Research*, vol. 175, no. 1, pp. 177–211, 2010.

[35] T. Groves and J. O. Ledyard, "Optimal allocation of public goods: A solution to the "free rider" problem," *Econometrica: Journal of the Econometric Society*, pp. 783–809, 1977.

[36] D. Fudenberg and J. Tirole, *Game theory*. Cambridge, MA: MIT Press, 1991, 1991.

[37] H. Margetts, "Rethinking democracy with social media," *Political Quarterly*, vol. 90, 2018.

[38] M. Motta, D. Stecula, and C. Farhart, "How right-leaning media coverage of covid-19 facilitated the spread of misinformation in the early stages of the pandemic in the us," *Canadian Journal of Political Science/Revue canadienne de science politique*, pp. 1–8, 2020.

[39] D. Graves, "Understanding the promise and limits of automated fact-checking," *Technical Report from Reuters Institute for the Study of Journalism*, 2018.

APPENDIX A

In this appendix, we present an extension of by relaxing Assumption 1 to a more general assumption that no platform has a monopoly on its users. The mechanism presented in this assumes that for any platform $i \in \mathcal{I}$ with the set of competing platforms $\mathcal{C}_i$, it holds that $|\mathcal{C}_i| \geq 2$.

We consider the same step one (the participation step) for the mechanism as before. Then in step two (the bargaining step), the message of platform i is defined as

$$m_i := (\tilde{h}_i, \tilde{p}_i, \tilde{a}_i), \tag{46}$$

where $\tilde{h}_i \in \mathbb{R}_{\geq 0}$ is the minimum average trust that platform i proposes to achieve through filtering; $\tilde{p}_i$, is the collection of prices that platform i is willing to pay or receive per unit changes in the filters of other competing platforms (except i) and the government's lower bound, given by

$$\tilde{p}_i := \begin{cases} (\tilde{p}_l{}^i : l \in \mathcal{D}_i), & \text{if } |\mathcal{C}_i| = 2, \\ (\tilde{p}_l{}^i : l \in \mathcal{D}_{-i}), & \text{if } |\mathcal{C}_i| \geq 3, \end{cases} \tag{47}$$

where $\tilde{p}_l{}^i \in \mathbb{R}_{\geq 0}$ for all $i, l \in \mathcal{J}$; and $\tilde{a}_i := (\tilde{a}_k^i : k \in \mathcal{D}_i)$, with $\tilde{a}_i \in \mathbb{R}^{|\mathcal{D}_i|}$ is the profile of filters for all competing platforms (including i) and government's lower bound proposed by platform i.

The message of the government is $m_0 := (\tilde{p}_0, \tilde{a}_0^0)$, where $\tilde{p}_0 \in \mathbb{R}_{\geq 0}$ is the price that the government is willing to pay or receive per unit change of the average trust, and $\tilde{a}_0^0 \in \mathbb{R}$ is the lower bound proposed by the government.

Based on the message profile $m := (m_0, m_1, \ldots, m_{|\mathcal{I}|})$ that the social planner receives, she allocates the following parameters to the players:

1) The social planner allocates a filter to each platform $i \in \mathcal{I}$ and a lower bound to the government such that the constraints of Problem 1 are satisfied. The filter allocated by the social planner to platform i is $\alpha_i(m) := \sum_{k \in \mathcal{C}_i} \frac{\tilde{a}_i^k}{|\mathcal{C}_i|}$. The lower bound allocated by the social planner to the government is $\alpha_0(m) := \sum_{k \in \mathcal{J}} \frac{\tilde{a}_0^k}{|\mathcal{J}|}$.

2) The social planner allocates a minimum average trust $\eta_i(m) \in [0, 1]$ to each platform $i \in \mathcal{I}$, given by

$$\eta_i(m) := \min \left\{ \frac{n_i \cdot \tilde{h}_i}{\sum_{k \in \mathcal{I}} n_k \cdot \tilde{h}_k} \cdot \alpha_0(m),\ 1 \right\}, \tag{48}$$

where the social planner will not accept a message m_i from a platform i that might lead to a situation where $\sum_{k \in \mathcal{I}} n_k \cdot \tilde{h}_k = 0$. The allocated minimum average trust, $\eta_i(m)$, is a lower bound on average trust that must be achieved by platform i. Let the filter implemented by platform i be a_i. Then, platform i must ensure that $n_i \cdot h_i(a_i) \geq \eta_i(m)$. Recall from Section III-B that, as a result of Assumption 5, the social planner can prevent the platforms from violating the constraint imposed by $\eta_i(m)$.

3) The social planner also allocates a payment price

$$\pi_l{}^i := \begin{cases} \tilde{p}_l^l, & \text{if } |\mathcal{C}_l| = 2, \\ \sum_{k \in \mathcal{C}_{-l} : k \neq i} \dfrac{\tilde{p}_l{}^k}{|\mathcal{C}_l| - 2}, & \text{if } |\mathcal{C}_l| \geq 3, \end{cases} \tag{49}$$

where $\pi_l^i \in \mathbb{R}_{\geq 0}$, to be paid by platform $i \in \mathcal{I}$ for a unit change in allocated filter $\alpha_l(m)$ of every other competing platform $l \in \mathcal{C}_{-i}$. Furthermore, the social planner allocates a subsidy price

$$\sigma_l^i := \begin{cases} \tilde{p}_i^l, & \text{if } |\mathcal{C}_l| = 2, \\ \sum_{k \in \mathcal{C}_{-l} : k \neq i} \dfrac{\tilde{p}_l{}^k}{|\mathcal{C}_l| - 2}, & \text{if } |\mathcal{C}_l| \geq 3, \end{cases} \tag{50}$$

where $\sigma_l^i \in \mathbb{R}_{\geq 0}$, to be received by platform $i \in \mathcal{I}$ from every other competing platform $l \in \mathcal{C}_{-i}$, for a unit change in allocated filter $\alpha_i(m)$. For the government, the social planner simply allocates a price $\pi_0 := \sum_{i \in \mathcal{I}} \frac{\tilde{p}_0^i}{|\mathcal{I}|}$ to be paid for a unit change in lower bound $\alpha_0(m)$.

Remark 9. Note that when $|\mathcal{C}|_i = 2$, platform $i \in \mathcal{I}$ proposes a price corresponding to their own proposed action $\tilde{a}_i(m)$. In contrast, when $|\mathcal{C}_i| \geq 3$, platform i does not proposes a price corresponding to their own filter. However, we have designed the payment price in (49) and subsidy price (50) so that platform i cannot affect either of these prices with their message m_i. Thus, each still platform behaves as a *price taker* when $|\mathcal{C}_i| = 2$.

4) The social planner allocates the following tax to each social media platform $i \in \mathcal{I}$,

$$\begin{aligned} \tau_i := -\tilde{p}_0 \cdot \eta_i(m) - \sum_{l \in \mathcal{C}_{-i}} \sigma_l^i \cdot \alpha_i(m) + \sum_{l \in \mathcal{C}_{-i}} \pi_l^i \cdot \alpha_l(m) \\ + \sum_{l \in \mathcal{C}_{-i} \cup \{0\}} \tilde{p}_l{}^i \cdot (\tilde{a}_l^i - \tilde{a}_l^{-i})^2 \\ + \sum_{l \in \mathcal{C}_{-i}} \Big(\mathbb{I}(|\mathcal{C}_i| = 2) \cdot (\tilde{p}_i^i - \tilde{p}_i^l)^2 + \mathbb{I}(|\mathcal{C}_l| = 2) \cdot (\tilde{p}_l^i - \tilde{p}_l^l)^2 \Big), \end{aligned} \tag{51}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\tilde{a}_l^{-i} = \sum_{k \in \mathcal{C}_{-l}} \frac{\tilde{a}_l^k}{|\mathcal{C}_l| - 1}$, for each $l \in \mathcal{C}_{-i}$, is the average of the proposed filters for l by all competing platforms except $i \in \mathcal{I}$, and and $\tilde{a}_0^{-i} = \sum_{k \in \mathcal{J}_{-i}} \frac{\tilde{a}_0^k}{|\mathcal{J}| - 1}$ is the average of lower bounds proposed by all players except i. The tax $\tau_i(m)$ of platform i in (51) can be interpreted as follows: (i) the first term in (51) represents a subsidy given by the government to platform i for the increase in average trust among the users of platform i; (ii) the second term in (51) is a collection of subsidies given by each competing platform $l \in \mathcal{C}_{-i}$ to platform i for the increase in valuation $v_l\big(\alpha_k(m) : k \in \mathcal{C}_l\big)$ due to the allocated filter $\alpha_i(m)$; (iii) the third term in (51) is a payment by platform i for the increase in valuation $v_i\big(\alpha_k(m) : k \in \mathcal{C}_i\big)$ due to the allocated filter $\alpha_l(m)$ of each competing platform $l \in \mathcal{C}_{-i}$; (iv) the fourth term in (51) is a collections of penalties to platform i if either the filter proposed in message m_i for any competing platform $l \in \mathcal{C}_{-i}$ is inconsistent the filters proposed by other platforms, or if the lower bound proposed in m_i is inconsistent with the lower bound proposed by other players;

and (v) the fifth term is a collection of penalties to social media i for inconsistency in the proposed price, only applicable if $|\mathcal{C}_i| = 2$, or if $|\mathcal{C}_l| = 2$ for some $l \in \mathcal{C}_{-i}$.

The social planner also proposes the following payment function to the government:

$$\tau_0 := \pi_0 \cdot \alpha_0(m) + (\tilde{p}_0 - \pi_0)^2, \tag{52}$$

where the first term is the total investment made by the government for the allocated lower bound $\alpha_0(m)$, and the second term is a penalty when the price proposed by the government deviates from the price allocated to the government.

Remark 10. Note that the presence of the indicator function $\mathbb{I}(\cdot)$ in (51) does not lead to discontinuities in the utility $u_i(m_i, m_{-i})$ with respect to the message profile m.

Remark 11. The extended mechanism induces a game where the strategy of platform $i \in \mathcal{I}$ is $m_i \in \mathcal{M}_i$, such that $\alpha_i(m) \in \mathcal{S}_i(m)$. The equilibrium for the induced game is given by the GNE, defined in (15).

Then, we note that the results of Lemmas 1 - 2 hold for the extended mechanism. In addition, we prove in Lemma 5 that the equilibrium price received by any platform $i \in \mathcal{I}$ for an allocated filter $\alpha_l(m)$, $l \in \mathcal{C}_{-i}$, is the same as the price paid by the competing platform l.

Lemma 5. *Let message profile $m^* \in \mathcal{M}$ be a GNE of the induced game. Then, for each platform $i \in \mathcal{I}$ and each competing platform $l \in \mathcal{C}_{-i}$, it holds that $\sigma_l^{*i} = {\pi_i}^{*l}$.*

Proof. Consider two social media platforms $i \in \mathcal{I}$ and $l \in \mathcal{C}_{-i}$. The result holds from the definition of σ_l^i and ${\pi_i}^l$ when both $|\mathcal{C}_i| \geq 3$ and $|\mathcal{C}_l| \geq 3$.

Let $|\mathcal{C}_i| = 2$, with $\mathcal{C}_i = \{i, l\}$, and let m_{-i}^* be the message profile at GNE of all players except i. In order to maximize their utility $u_i(m_i, m_{-i}^*)$, platform i must select a price $\tilde{p}_i^{*i}$ that minimizes the tax τ_i in (51). Thus, $\frac{\partial u_i}{\partial \tilde{p}_i^i}\big|_{\tilde{p}_i^{*i}} = 2 \cdot (\tilde{p}_i^{*i} - \tilde{p}_i^{*l}) = 0$, which yields $\tilde{p}_i^{*i} = \tilde{p}_i^{*l}$. Then, the result holds using the definitions of σ_l^i and ${\pi_i}^l$ in (50) and (49), respectively. Through a similar analysis, we can prove the result when $|\mathcal{C}_l| = 2$. □

An additional implication of Lemma 5 is that the fifth term in (51) is 0 at any GNE. Thus, it can be verified that the results of Lemmas 3 - 4 and Theorems 1 - 5 hold for the extended mechanism.