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

Preetesh Verma (2018eeb1171) IIT Ropar, under the guidance of Dr. Shweta Jain

***Abstract*— This report is an attempt by me to provide a summary on the above mentioned Research Paper. The paper presents a resource allocation mechanism for the problem of Government fund distribution among a finite number of social media platforms.The proposed mechanism incentivizes social media platforms to filter misleading information efficiently, and thus indirectly prevents the spread of fake news. The attempt is to design an economically inspired mechanism that strongly implements all generalized Nash equilibria for efficient filtering of misleading information in the induced game and is individually rational, budget balanced, while it has at least one equilibrium.**

I. INTRODUCTION

The deluge of information available on the internet has made it extremely difficult to identify facts.There are several misinformation campaigns being carried out on these social media platforms which tends to disrupt the 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.** The introduction of alternative facts can reduce the trust on common knowledge about democracy.E.g. **2020 American Election and 2016 Brexit.** The government wants to increase the trust of the users in the democratic processes which could only be achieved by reducing the spread of misinformation which itself can be achieved by censoring of fake news(Assuming,social media platforms have the technologies to filter, or label, posts that intend to sacrifice trust on common knowledge.).But it has also been found that censoring of articles by a social media platform generally tends to decrease its user engagement time and number of users. We induce a misinformation filtering game to describe the interactions between the social media platforms and the Government. The social media platforms want to increase the user engagement time on their platforms while for the government who is also a strategic player,the utility increases as the trust of the users of social media platforms on common knowledge increases. Thus, the Government is willing to make an investment to incentivize the social media platforms to filter misinformation. Thus,we need to use the mechanism design **to distribute this investment of the Government among the platforms optimally, and in return,the social media platforms implement an optimal level of filtering.**

II. PROBLEM FORMATION

We consider a democratic society consisting of a finite and nonempty set of social media platforms I = 1,2,3 . . . , I, I E 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 J = I U 0, where the index 0 corresponds to the Government.

A. Misinformation Filtering Game for Platforms

Let the informativeness of a post on platform i E I be denoted by x_i E [0, 1], where x_i = 0 indicates that the post contains complete misinformation and decreases average trust of users and x_i = 1 indicates that the post contains completely factual information. The action a_i of platform i represents the level of filtering imposed by platform i and takes values in a feasible set of actions 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.

Higher the value of a_i, more is the migration of users from platform i to others and reduction in revenue for platform i. This leads to define a set of **competing platforms** as follows:For each platform i E I, the set $C_i \subset$ I, with i E C_i, is the set of competing platforms whose choice of filters has an impact on the user engagement of platform i.

The **valuation function** of a social media platform i E I is $v_i(a_k$: k E $C_i)$: $A^{|C_i|} \rightarrow$ R$\geq$0. It is a decreasing function with respect to a_i and strictly increasing with respect to a_l for all l E C_-i, where C_{-i} = C_i/(i).

The **average trust function** of the users platform i E I on common knowledge is $h_i(a_i)$: A $\rightarrow$ [0, 1], and it is a strictly increasing function with respect to a_i.

B. Misinformation Filtering Game for Government

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 E A = [0, 1] that designates a lower bound which must be satisfied by the aggregate average trust of all social media platforms in I. Let N_i E N be the total number of users of the social media platform i E I. Then, the fraction of the number of users of i with respect to the total number of users of all platforms is

$$n_i = \frac{N_i}{\sum_{lEI} N_l}$$

Since $\sum_{iEI} n_i$=1, the aggregate average trust on common knowledge is $\sum_{iEI} n_i * h_i(a_i)$. After the government decides on a_0, each platform i E I who decides to participate in the game must select a filter a_i that satisfies the following constraint: a_0 - $\sum_{iEI} n_i * h_i(a_i) \leq 0$.

The valuation function of the government is $v_0(a_0)$: [0, 1] $\rightarrow$ R$\geq$0, and it is an increasing function with respect to the lower bound a_0.

C. Information Structure and Assumptions

The private and public information structure corresponding to each player in the imposed game.

1) **Public information:** The set of competing platforms C_i and fraction of users n_i of each platform i E I are known to all players in set J . Moreover, the set of feasible actions A is known to all players in the set J .

2) **Valuation functions:** The valuation function $v_i(\cdot)$ of each social media platform i E 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 E I is considered private information, and thus, it is known only to platform i (it is not known to the government).

Assumption 1 For each platform i E I, $|C_i| \geq 3$. We impose this assumption to simplify the exposition of our mechanism. Assumption 1 implies that each user subscribes in multiple social media platforms.

Assumption 2 The valuation functions of the government, players and the average trust function of the user of platforms is assumed to be concave and differentiable function with respect to a_0, a_k E C_i and a_i respectively.

Assumption 3 The output of the function $h_i(a_i)$ can be monitored by any competing platform l E C_i, and a violation of the average trust condition can be detected by the government. This assumption ensures that each platform implements the filter to improve the average trust of the players.

Assumption 4 The government ensures that any social media platform i E I that does not participate in the mechanism receives no benefits from the filters of participating social media.

D. Problem Statement

Since there is a conflict of interest between the government and the social media platforms,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.

To meet these objectives, the social planner asks each player i E J to send a message m_i from a set of feasible messages M_i. Based on the message profile m = (m_0, m_1, . . . , $m_|I|$), the social planner assigns a tax $T_i(m)$ E R for each social media platform i E I, and an investment T_0(m) E R$\geq$0 for the government. By convention, a tax Ti(m) > 0 is a payment made by player i E J , and a tax T_i(m) < 0 is a subsidy given to player i. T 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., $\sum_{iEJ} T_i(m) = 0$. The utility of each platform i E I is given by u_i(m,a_k: k EC_i) = v_i(a_k: k E C_i) - T_i(m) while the utility of the government is given by u_0(m, a_0) = v_0(a_0) - T_0(m).

Thus,the problems are :

$$max(v_0(a_0) - T_0(m) + \sum_{iEI}(v_i(a_k : kEC_i) - T_i(m))$$

subject to constraints,

$0 \leq a_i \leq 1$ for all i E J, (ensures feasibility)

a_0 - $\sum_{iEI} n_i * h_i(a_i)$ (average trust is above lower bound)

$0 \leq T_0(m) \leq b_0$ (budget is not exceeded)

$\sum_{iEJ}(T_i(m))$=0 (ensures budget balance)

By maximizing the social welfare function the utility of players is maximized and participation is maximized.Note that 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 $T_i(m)$ for each player i E J to incentivize the players to be truthful while, at the same time, maximizing the social welfare. Thus the problem statement has two parts: one concerning the maximization of the utility of the platforms while ensuring the average trust is maintained and the other one being the maximization of the utility of the Government with the investment being within a stipulated budget.

III. MECHANISM DESIGN APPROACH

The solution presents 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: Participation Step

Consider a platform i E I that chooses not to participate in the mechanism. Thus, this platform neither pays taxes nor receives any subsidies from the government, i.e., T_i(m) = 0. Furthermore, platform i is free to select the lowest value of a_i = 0 that maximizes the valuation v_i(a_k : k E C_i).Meanwhile, another competing platform l E C_i may decide to participate in the mechanism and subsequently implement a non-zero filter a_l. From Assumption 4, the government ensures that platform i receives no utility as a result of filter a_l. Thus, the utility of the non-participating platform i is given by v_i(a_k = 0 : k E C_i)

B. Step Two - The Bargaining Step

In step two, the social planner asks each player i E J to broadcast a message m_i from a set of feasible messages M_i. For each platform i E I, let D_i = C_i U 0, and D_i = D_i/(i). The message of platform i is defined as:

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

where $\tilde{h}_i$ E $R_{\geq 0}$ is the minimum average trust that platform i proposes to achieve through filtering; $\tilde{p}_i$ E $R_{\geq 0}^{D-i}$ 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 E D_-i); and $\tilde{a}_i$ = ($\tilde{a}_k^i$: k E D_i), $\tilde{a}_i$ E $R^{|D_i|}$, 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 $\tilde{p}_0$ E R$\geq$0 is the price that the government is willing to pay or receive per unit change of the average trust, and $\tilde{a}_0^0$ E R is the lower bound proposed by the government. Note that the mechanism respects the privacy of each platform i E I since she does not request either their or government's valuation function v_i or their average trust function.

C. Parameters Assigned by the Social Planner

1) The social planner allocates a filter to each platform i E 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 α_i(m) = $\sum_{kEC_i} \frac{\tilde{a_i^k}}{|C_i|}$, 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 α_0(m) = $\sum_{kEJ} \frac{\tilde{a_0^k}}{|J|}$

i.e., the average of the lower bounds proposed by all platforms and the government.

2) The social planner allocates a minimum average trust n_i(m) E [0, 1] to each platform i E I, given by

$$n_i(m) = min\left\{\frac{n_i * \tilde{h_i}}{\sum_{kEI} n_k * \tilde{h_k}} * \alpha_0(m), 1\right\}$$

Let the filter implemented by platform i be a_i.Then, platform i must ensure that $n_i \cdot h_i(a_i) \geq n_i$(m). Any violation of $n_i \cdot h_i(a_i) \geq n_i$(m) will be reported by platform l E C_i to the social planner, in order to ensure that platform i implements the largest filter a_i, and maximizes the utility $u_l(m, a_k : kEC_l)$.This prevents platforms from violating the constraint imposed by the allocated minimum average trust n_i(m).

3) The social planner allocates a price π_l^i=$\sum_{kEC_{-l}:k\neq i} \frac{\tilde{p_l^k}}{|C_l|-2}$ to each platform i E I,corresponding to the allocated filter $\alpha_l(m)$ of every other competing platform l E c_-i this price is derived as the average price proposed for the allocated filter $\alpha_l(m)$ by all competing platform in C_-l except i. Similarly social planner allocates the price π_0=$\sum_{iEI} \frac{\tilde{p_0^i}}{|I|}$ to the government.

4)The social planner allocates the following tax to each social media platform i E I,

$$T_i(m) = -\tilde{p}_0.n_i(m) - \sum_{lEC-i} \pi_i^l.\alpha_i(m) + \sum_{lEC-i} \pi_l^i.\alpha_l(m)$$

+$\sum_{lEC-iU(0)} \tilde{p}_l^i.(\tilde{a}_l^i - \tilde{a}_l^{-i})^2$,

The tax $T_i(m)$ of platform i can be interpreted as the first term in tax equation represents a subsidy given by the government to platform i for the increase in average trust among the users of platform i; the second term in tax equation is a collection of subsidies given by each competing platform lEC_-i to platform i for the increase in valuation $v_l(a_k : kEC_l)$ due to the allocated filter α_i; the third term in tax equation is a payment by platform i for the increase in valuation $v_i(a_k : kEC_i)$ due to the allocated filter α_l of each competing platform lEC_-i; and the fourth term in tax equation is a collection of penalties to platform i if either the filter proposed in message m_i for any competing platform lEC_-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 player.

Finally, the social planner allocates the following investment to the government: $T_0(m) = \pi_0.\alpha_0(m) + (\tilde{p}_0 - \pi_0)^2$

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. The strategy of platform i in the induced game is given by the message m_iEM_i , with a constraint that $\alpha_i(m)ES_i(m)$, where

$S_i(m) = a_iEA : n_i.h_i(a_i) \geq n_i(m)$

Thus, the set of feasible allocations $S_i(m)$ for iEI is a function of the messages of all social media in I and the government. A message profile m^*=(m_i^* :i E J) is the GNE of the induced game if for each i E I,

$$u_i((m_i^*, m_{-i}^*), \alpha_k(m_i^*, m_{-i}^*) : kEC_i) \geq$$

$\mathrm{u}_i((m_i^*, m_{-i}^*), \alpha_k(m_i, m_{-i}^*) : kEC_i)$

for all m_iEM_i and $\alpha_iES_i(m)$; and (ii) the message m_0^* of the government is such that

$$u_0((m_0^*, m_{-0}^*), \alpha_0(m_0^*, m_{-0}^*)) \geq u_0((m_0, m_{-0}^*), \alpha_0(m_0, m_{-0}^*))$$

, for all m_0EM_0. 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 $v_i(a_k : kEC_i)$ 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.

IV. PROPERTIES OF THE MECHANISM

The 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

- Let the message profile m^* E M be a GNE of the induced game. Then, $\tilde{p}_0^*$ = π_0^* for the government.
- Let the message profile m^* E 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 E I, for every k E D_{-i}.
- Consider any GNE m* E M of the induced game. Then, the proposed mechanism is budget balanced, i.e., $\sum_{iEJ} T_i(m^*) = 0$. Since $\sum_{iEI} n_i(m)$ =$\alpha_0(m)$,$\forall$ m E M so we have $\sum_{iEJ} T_i^*$ =

$\sum_{iEI}\left[-\sum_{lEC_{-i}}\pi_i^l * \alpha_i(m^*) + \sum_{lEC_{-i}}\pi_l^i * \alpha_l(m^*)\right]$= 0

- Every GNE message profile m^* E M leads to a filter profile $\alpha_1(m)$, . . . , $\alpha_{|I|}(m^*)$ and lower bound $\alpha_0(m^)$, which is a feasible solution. For each i E I $n_i(m) \leq n_i * h_i(\alpha_i(m)), and \sum_{iEI} n_i(m) = \alpha_0(m).Hence \sum_{iEI} h_i(\alpha_i(m)) \geq \alpha_0(m)$.
- Consider any GNE m^* E M of the induced game. Then, the allocated filter profile $\alpha_1(m^*), ..., \alpha_{|I|}(m^*)$ and the allocated lower bound $\alpha_0(m^*)$ at equilibrium is equal to the optimal solution a a^{*o} of Problem.
- Let a_0^*=($a_o^{*o}, a_1^{*o}, ...a_I^{*o}$) be the unique optimal solution of Problem. Then, there is a GNE message profile m^* E M of the induced game that guarantees that the filter profile $\alpha_1(m^*), ..., \alpha_{|I|}(m^*)$ and lower bound $\alpha_0(m^*)$ at GNE satisfy $\alpha_i(m^*) = a_i^{*o}$, for all i E J .
- The proposed mechanism is individually rational, i.e., each platform i E I prefers the outcome of every GNE of the induced game to the outcome of not participating.
- The utility $u_i(m^*)$ at any GNE m^* E M of a platform i E 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 E I is to participate in the mechanism.

V. MY LEARNINGS

While working on this research paper there were several things which I learned such as Game Theory and Mechanism Design can provide us with a solution in almost every social situation.The problem dealt in the research paper is a current real world threat to the democracies and to our lives.The social media platforms and the government both act as competing players and through mechanism design we are able to forge a solution so that both of them end up rewarding each other for the betterment of the society by removing the fake news posts from the platform.

Some further studies on the research paper can be focused on more practical 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.

Another area which could be worked upon focuses on improving the valuation and average trust functions of the social media platforms based on data.Thus, this paper does have a scope for future improvement.

VI. DISCUSSIONS AND CONCLUSIONS

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 $v_i(a_k : kEC_i)$ of any platform i E I, and the average trust function $h_i(a_i)$ of any social media platform i E I.

To meet this challenge, we present a two-step mechanism. 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 that the utility of any platform i E 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 E J to broadcast a message m_i E M_i. Based on the message profile m = (m_0, m_1, . . . , $m_|I|$), the social planner allocates a minimum average trust n_i(m), a filter α_i(m), and a tax T_i(m) to each platform i E I. Similarly, she allocates a lower bound α_0(m) and tax T_0(m) to the government. By participating in the mechanism in step one, each player i E 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 Generalised Nash Equilibrium. Then we prove the properties of the mechanism that at any GNE the game is being budget balanced, feasible and maximizes utility of players and that there exists a GNE for the game. 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.

ACKNOWLEDGMENT

I would like to record my appreciation to the CS Department of my college and especially to the course instructor Professor Shweta Jain for providing me with this opportunity to work on the topic.It greatly enhanced my knowledge in the subject and gave me new insights under the topic of "Mechanism Design".

REFERENCES

[1] Social Media and Misleading Information in a Democracy: A Mechanism Design Approach, author=Aditya Dave and Ioannis Vasileios Chremos and Andreas A. Malikopoulos,year=2021,eprint=2003.07192,