CS136 Project: Economic Mechanism to Prevent Misinformation

Ostap Stefak, Tejovan Parker, Jacob Cremers

December 7, 2023

Contents

1 Motivation

1.1 The state of the world

The spread of falsehoods is becoming an increasing problem on social media.

According to Statistica, 67% of Americans have encountered fake news on social media, and 10% of U.S. adults have knowingly shared fake news[1]. According to a study done at MIT in 2018, false news on Twitter (now called X) is 70% more likely to be retweeted, and retweet cascades (where a tweet is retweeted multiple times in a row) grow 10 to 20 times faster than truthful tweets[2].

To address this issue of increasingly prevalent and viral falsehoods, Prof. Marshall Van Alstyne and his team at BU have been conducting research on how to use mechanism design to place friction on false and harmful ideas on social media platforms[3]. A mechanism that can successfully produce these desired effects would increase truthfulness and informedness across society.

In this paper we state simplified versions of the mechanism proposed by Van Alstyne in formal game-theoretic terms and investigate their properties through some theoretic proofs and computational simulation.

1.2 Mechanism Goals

Our goal is to make a system in which it is more beneficial for agents to tell the truth than to tell lies, and where truthtellers are more confident in stating the truth than liars.

Formally, we would like:

1. Speaker's utility is negative in expectation (opportunity gain compared to simple paid advertising or not speaking) for warranting something they believe is false.
2. Speaker's utility (opportunity gain compared to simple paid advertising or not speaking) for warranting something they believe is true is positive, with very-low/bounded risk.
3. A post being warranted is a strong diagnostic, a.k.a. a credible signal, of a claim being true.
4. The mechanism would be scalable. Especially considering the total cost of accurate arbitrations remaining low over all claims.

To do this, we use money. One puts their money where their mouth is, as the saying goes.

In our mechanism, the unit of analysis is a claim. This is a statement/post that somebody writes on the social network. In our system, those who are more confident that their claims is true will stake more on the truthfulness of their claim, as a signal of confidence that their claims is true. We also, however, want such high staking to harm liars when it is determined that they are in fact lying.

[1] https://news.temple.edu/news/2021-11-09/study-shows-verified-users-are-among-biggest-culprits-when-it-comes-sharing-fake

[2] https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[3] See section "VIII. Solutions Derived from Coase" from [1] "Free Speech, Platforms & The Fake News Problem", Marshall Van Alstyne, SSRN, `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3997980`

1.3 3 Components

From a birds-eye view, we envisage three components of a mechanism to counter misinformation:

1. Filtering: determining which claims get to arbitration. Arbitration is costly. Therefore, it is not feasible to have every statement posted to a platform arbitrated. Filtering is the process by which certain statements are selected to be arbitrated. The filter could be random, or it could be done through some voting process, or some anti-staking or betting process. In the model proposed in Section 2, the filter is based on the ground truth of a claim. The assumption here is that individuals are more likely to challenge claims that they believe are false, because they receive a proportion of the warranted amount if they do so. However, a detailed model of this process is omitted.

2. Arbitration (for some claims): If a claim is selected for arbitration, this step is executed on that claim. By some process, a statement is judged as either True or False. The arbitration has some level of accuracy but can be assumed to not be completely accurate. The arbitrations process could be designed in many different ways. To avoid designing a highly complex entire peer prediction system, in this paper we mainly abstract away the arbitration step, simplifying it by only considering the accuracy, and the final True/False output.

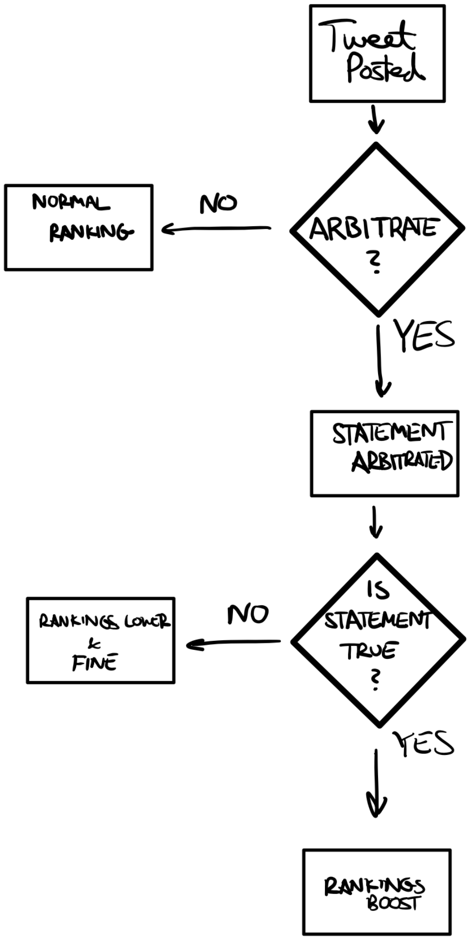
3. Ranking and amplification of statements: This mechanism component determines how many views the statement receives. A statement that is arbitrated True should get a boost. A statement arbitrated False should be docked severely. A statement that does not reach arbitration should be scaled by some intermediate amount. The warranted amount should also play a role here. Claims that have more money placed on them should be ranked higher, to indicate their credibility. In the model proposed in Section 2, the warranted amount is the same for all claims, so only the arbitration decision plays a role in the number of views a claim gets.

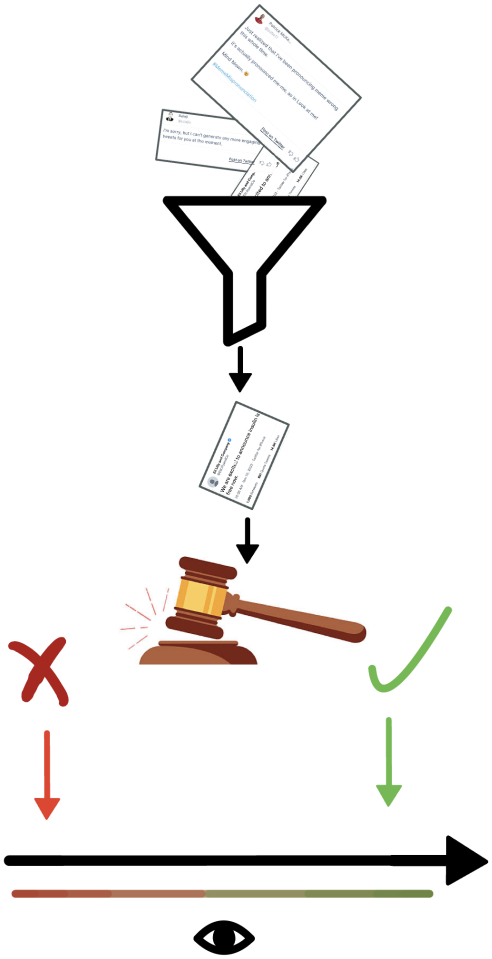
1.4 Paper outline

In Section 2 we set up a bare-bones mechanism that nonetheless has three key components that we envisage in a system of this kind (filtering, arbitration, ranking) and investigate how a mechanism designer would need to set the choice variables to guarantee that False claims are punished, while True claims are promoted. In Section 4 we make this mechanism slightly more realistic, and simulate its outcome in a favorable and unfavorable setting. We also compare it to a simulated social media landscape without any mechanism to remove misinformation.

In Section 3 we propose an alternative mechanism for the ranking step, which uses the Ad Auction we discussed in class as a starting point. This is a richer model of the interaction between claims which the model in Section 2 does not account for.

1.5 Diagram

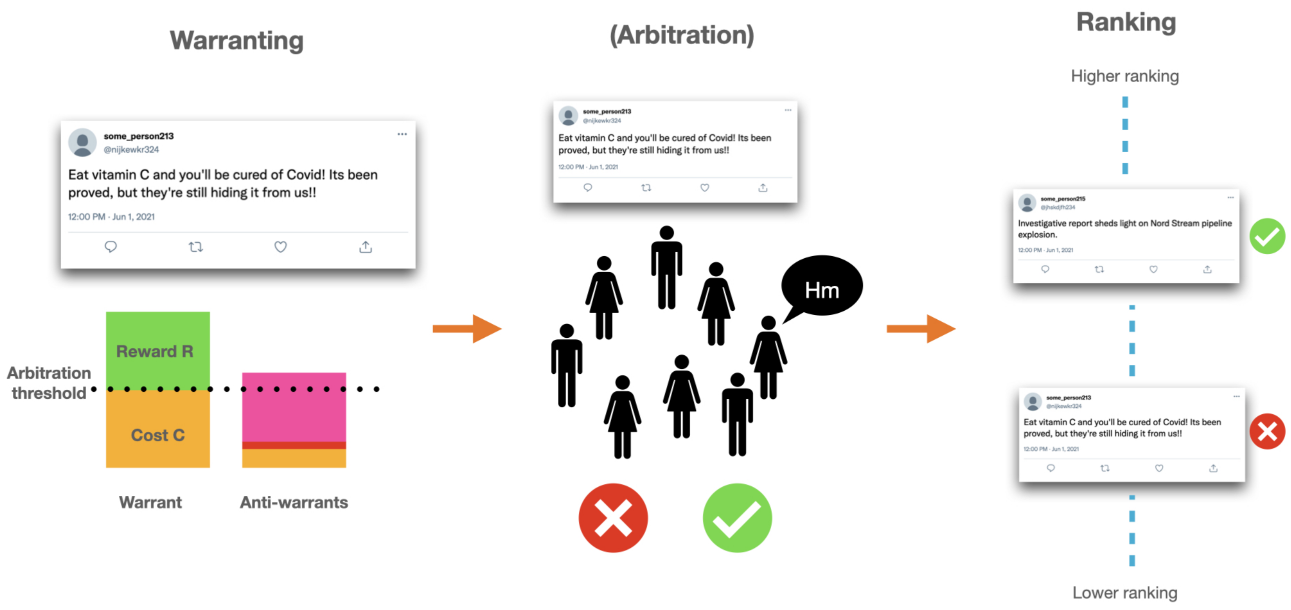




The chart on the left is a flow chart of the arbitration process. Tweets that aren't selected to be arbitrated are given a normal ranking. Those that are arbitrated are given higher rankings if they are arbitrated to be true, and lower rankings along with a fine if they are arbitrated to be false.

The chart on the right visualizes the simulation, showing how only certain statements are selected for arbitration, and showing how those that are arbitrated to be true get more views than those arbitrated to be false.

A third visual of the mechanism is the Figure below. It highlights that the arbitration step is a peer prediction problem, and gives some indication of what the Filtering / Warranting step might look like, with the warranted amount set to cover the cost of arbitration and some potential reward R for individuals who successfully challenge it.



some_person213
@nijkewkr324
Eat vitamin C and you'll be cured of Covid! Its been proved, but they're still hiding it from us!!
12:00 PM · Jun 1, 2021

some_person213
@nijkewkr324
Eat vitamin C and you'll be cured of Covid! Its been proved, but they're still hiding it from us!!
12:00 PM · Jun 1, 2021

some_person215
@jhskdjfh234
Investigative report sheds light on Nord Stream pipeline explosion.
12:00 PM · Jun 1, 2021

some_person213
@nijkewkr324
Eat vitamin C and you'll be cured of Covid! Its been proved, but they're still hiding it from us!!
12:00 PM · Jun 1, 2021

Figure 1: Mechanism Schematic

2 Theory I: Simple, binary model

2.1 Variables

Exogenous variables

- a_T: Probability that True claim reaches arbitration. Note: the probability a claim goes to arbitration will be determined by willingness of individuals on the social network to stake money against the claim to challenge it. This is out of the control of the mechanism designer.
- a_F: Probability that False claim reaches arbitration.
- s_T: The probability that a True claim is arbitrated True. Note: accuracy of arbitration step is assumed to be out of control of the mechanism designer.
- s_F: The probability that a False claim is arbitrated False.

Endogenous variables

- $G \in \{0,1\}$: Claim ground truth. 1 if True, 0 if False. Sampled from a distribution.
- $A \in \{0,1\}$: Filtering outcome. Variable indicating whether a claim reaches arbitration or not. 1 if claim reaches arbitration, 0 otherwise. Outcome of filtering process.
- $S \in \{0,1\}$: Arbitration outcome. 1 if claim is judged to be True, 0 if judged False.
- $R \in \{r_u, r_t, 0\}$: Ranking outcome.
- U: Final claiming utility.

Choice variables

- r_u: Number of views given to claims that do not reach arbitration (unverified claims).
- r_t: Number of views given to claims that reach arbitration, and are judged to be True.
- f: Dollar amount that needs to be warranted for a claim to be shown.

2.2 Assumptions

For theoretical analysis, we make several simplifying assumptions:

1. f is fixed, and the same for all claims. The warranted amount is set by the mechanism designer.

2. All posts are warranted posts.

3. The only difference between claims is their truthiness G. In all other respects (virality, utility per impression etc) they are the same.

In the theoretical analysis, we want to show that the mechanism designer can guarantee that making a False claim has negative utility, and making a True claim has positive utility.

2.3 Model components

2.3.1 Arbitration Occurrence (Filtering) A

We model that True and False claims have different probability of reaching the arbitration stage. In particular, the assumption is that $a_F \geq a_T \geq 0$. This is because individuals are motivated to receive the reward of the warranted amount f. They only receive this if the claim is judged False by the arbitration. So they have an incentive to preferentially challenge claims they believe are False, and which will likely ultimately be judged False.

	$A = 0$	$A = 1$
$G = 1$	$1 - a_T$	a_T
$G = 0$	$1 - a_F$	a_F

Table 1: Probability of reaching arbitration for True and False claims

$$P(A(G) = 1) = \begin{cases} a_T & \text{if } G = 1 \\ a_F & \text{if } G = 0 \end{cases}$$

$$P(A(G) = 0) = \begin{cases} 1 - a_T & \text{if } G = 1 \\ 1 - a_F & \text{if } G = 0 \end{cases}$$

2.3.2 Arbitration S

In the general form of the model, arbitration is assumed to be imperfect. If arbitration is perfect (i.e. it recovers ground truth of G) then we would have $s_T = s_F = 1$.

	$S = 0$	$S = 1$
$G = 1$	$1 - s_T$	s_T
$G = 0$	s_F	$1 - s_F$

Table 2: Probability of arbitration outcomes S for True and False claims

$$P(S(G) = 1) = \begin{cases} s_T & \text{if } G = 1 \\ 1 - s_F & \text{if } G = 0 \end{cases}$$

$$P(S(G) = 0) = \begin{cases} 1 - s_T & \text{if } G = 1 \\ s_F & \text{if } G = 0 \end{cases}$$

2.3.3 Ranking R

Note that the amplification system does not have access to the ground truth of whether a claim is True or False. All that it has access to is information about whether a claim reached arbitration A, and, if so, the arbitration outcome S.

The table indicates that there are three cases:

	$S = 0$	$S = 1$
$A = 0$	r_u	r_u
$A = 1$	$r_F = 0$	r_T

Table 3: Number of views given to claim by ranking system

1. **Claim does not reach arbitration**. In this case we make an assumption that it is likely true, or at least not blatant misinformation. So it is receives some positive amount of views r_u, where u stands for unverified claim.
2. **Claim reaches arbitration, and is judged False**. Such a claim receives no amplification in the social network. This means it is not recommended to anyone, and reposts are not allowed. The poster gets 0 utility from this outcome.
3. **Claim reaches arbitration, and is judged True**. Such a claim receives number of views r_T. We will set $r_T > r_u > 0$ to reward claims that pass arbitration and receive a True judgment.

2.4 Perfect arbitration

In perfect arbitration, we have $s_T = s_F = 1$. This means that the arbitration is perfect, with 0 false positive and false negative rate. This means that we learn the ground truth of any claim that goes to arbitration. In this setup:

- Utility of making a false claim:

$$E[U_F] = E[U_F|A = 1]P(A = 1) + E[U_F|A = 0]P(A = 0) \tag{1}$$
$$= (-f)a_F + r_u(1 - a_F) \tag{2}$$

- Utility of making a true claim:

$$E[U_T] = E[U_T|A = 1]P(A = 1) + E[U_T|A = 0]P(A = 0) \tag{3}$$
$$= (r_T)a_T + r_u(1 - a_T) \tag{4}$$

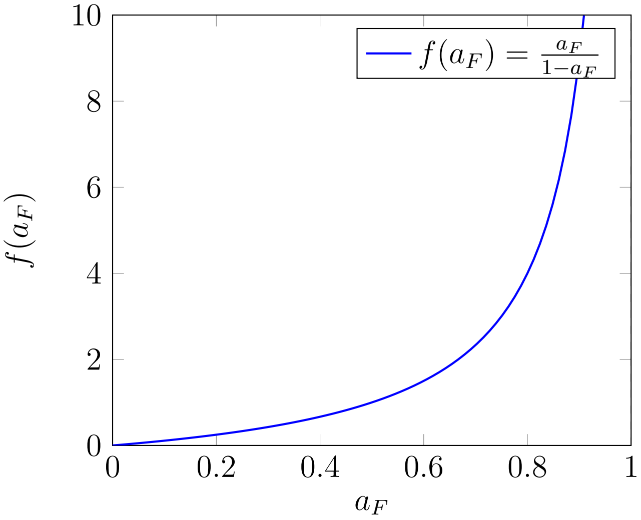
In the case of perfect arbitration, there is no risk of the individual making a true claim losing money, since they can only lose money if the claim is arbitrated false and this will never happen. So we automatically satisfy the requirement that $E[U_T] > 0$. Then, we just need to satisfy $E[U_F] < 0$. There are two choice variables that we can modify in the expression for $E[U_F]$, so there is no unique critical value that we can solve for. However, if we take f to be fixed, then we can solve for the maximum r_u that we can have for the expected utility of a False claim to still be negative:

$$(-f)a_F + r_u(1 - a_F) \leq 0$$
$$r_u \leq \frac{fa_F}{(1 - a_F)}$$

This expression shows us the tradeoffs at play in this simplest case setup:

1. The higher we set f, the higher we can set r_u. That is, the higher warrant amount we require, the more we can boost unverified claims without giving False claims a positive expected utility. This makes intuitive sense, since higher f means False claims have more to lose.

2. Notice the shape of the $f(a_F) = \frac{a_F}{1-a_F}$ graph in the range $\{0, 1\}$:



   It has an asymptote at $a_F = 1$. So, as the False claim approaches probability 1 of getting to arbitration, the restriction on r_u becomes irrelevant. We can set r_u to be as high as we like, because any False claims will be caught by the arbitration process anyway and receive 0 views, so only True claims will receive r_u.

2.5 Imperfect arbitration

2.5.1 Utilities

Imperfect arbitration means that arbitration does not return the ground truth G. Instead it has some error. It judges a True claim True with probability s_T, and a False claim False with probability s_F, where $s_T \neq 0$ and $s_F \neq 0$.

This complicates the theoretical analysis because it means that there is a probability that both the True or False claim pay the fee $-f$, if they are (correctly or incorrectly) determined to be False by the arbitration process.

So we have:

- Utility of making a false claim:

$$\begin{aligned}E[U_F] = {} & E[U_F|A=1, S=1]P(A=1, S=1) + E[U_F|A=1, S=0]P(A=1, S=0) \\ & + E[U_F|A=0, S=0]P(A=0, S=0) + E[U_F|A=0, S=1]P(A=0, S=1) \\ & = -f \cdot (a_F) \cdot (s_F) + r_T \cdot (a_F) \cdot (1 - s_F) + r_u \cdot (1 - a_F)\end{aligned}$$

- Utility of making a true claim:

$$\begin{aligned}E[U_T] = & E[U_T|A=1,S=1]P(A=1,S=1) + E[U_T|A=1,S=0]P(A=1,S=0) \\ & + E[U_T|A=0,S=0]P(A=0,S=0) + E[U_T|A=0,S=1]P(A=0,S=1) \\ & = r_T \cdot (a_T) \cdot (s_T) - f \cdot (a_T) \cdot (1-s_T) + r_u \cdot (1-a_T)\end{aligned}$$

Now, we see that it is not certain that making a True claim has a positive utility. It depends on the relative sizes of f, a_T, s_T, r_u and r_T.

2.5.2 Conditions on f and r_T for $E[U_F] < 0$

To simplify slightly, we can let $r_u = 1$, since by setting r_T we can control the ratio r_T/r_u, which is the important quantity in this setup.

We can also consider initially of f as being a fixed quantity (although in principle the mechanism designer can change it to anything they like). This leaves r_T as the only choice variable left to solve for. An important question is whether setting r_T to the maximum quantity for $E[U_F]$ to still have a negative utility, leaves $E[U_T] > 0$ or not.

Solving for r_T in this way we get:

$$\begin{aligned}-f \cdot (a_F) \cdot (s_F) + r_T \cdot (a_F) \cdot (1-s_F) + r_u \cdot (1-a_F) &< 0 \\ -f \cdot (a_F) \cdot (s_F) + r_T \cdot (a_F) \cdot (1-s_F) + (1-a_F) &< 0 \\ r_T \cdot (a_F) \cdot (1-s_F) &< f \cdot (a_F) \cdot (s_F) - (1-a_F) \\ r_T &< \frac{f \cdot (a_F) \cdot (s_F) - (1-a_F)}{(a_F) \cdot (1-s_F)} \\ r_T &< \frac{f \cdot (s_F) - (1-a_F)}{(1-s_F)}\end{aligned}$$

So in the final calculation

$$r_T < \frac{f \cdot (s_F) - (1-a_F)}{(1-s_F)}$$

is the condition for r_T. Given that we are assuming a large value of f (on the order of 100) so that it can cover the cost of arbitration, and a_F, s_F between 0 and 1, this is not a particularly restrictive condition.

Another important fact is that we require $r_T > 1$, given that we have set $r_u = 1$. So we also have the requirement on f that

$$1 < \frac{f \cdot (s_F) - (1-a_F)}{(1-s_F)}$$

Solving for f, this gives that

$$f > \frac{(1-s_F)+(1-a_F)}{s_F}$$

So there is also this lower bound on f. This is also not a very restrictive condition, given reasonable values of a_F and s_F. For example, given $s_F = 0.9, a_F = 0.9$ (arbitration system working well), we get that f must exceed 2/9. In a less favorable case of $s_F = 0.6, a_F = 0.6$, we get that f must exceed 4/3.

2.5.3 Maximize $E[U_T]$ subject to $E[U_F] < 0$

Another relevant goal may be to maximize $E[U_T]$ subject to the constraint that $E[U_F] < 0$. We can do this using linear programming techniques. Our linear program is:

$$\begin{aligned} \text{Maximize } & r_T \cdot a_T \cdot s_T - f \cdot a_T \cdot (1-s_T) + (1-a_T) \\ \text{S.T } & -f \cdot a_F \cdot s_F + r_T \cdot a_F \cdot (1-s_F) + (1-a_F) \leq 0 \\ & r_T \geq 1 \end{aligned}$$

We are interested in how $\max E[U_T]$ changes as the exogenous variables a_T, a_F, s_T, s_F change, subject to $f = 100$ and the optimal value of r_T being chosen. Below is a plot of the output of the linear program. In this linear program, 'Arbitration Accuracy' refers to s_T, s_F and 'Probability of Arbitration' refers to a_T, a_F.

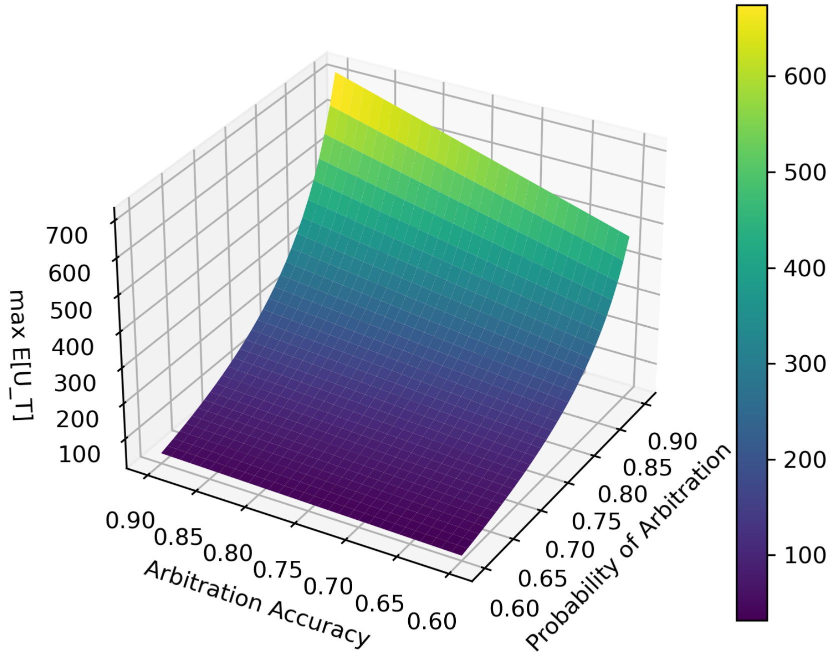


Figure 2: Linear Programming Formulation

In this plot, we see the expected relationship between probability of arbitration and arbitration accuracy, with the highest expected utility of a True claim being reached when both are high. It is also encouraging that even with fairly low arbitration accuracy and arbitration probability it is possible to choose r_T such that $E[U_F] \leq 0$ while $E[U_T]$ is significantly larger than zero.

3 Theory II: Views via Simplified Ad Auction

In this section we propose a more complex model for the ranking step R which considers interaction between claims, and analyse its theoretical properties.

3.1 Assumptions & Setup

Initially, we consider a two-bidder, two-slot, ad auction setting with identical qualities, $Q_i = 1\ \forall i$. At the end of the analysis we also simplify to a single-slot, making it effectively a single-item auction. We use a slightly modified VCG mechanism, where the payment for an allocated advertiser also includes the cost of any warrant they lose. We assume arbitration of the claim always happens, and that the arbitration a claim is always accurate. With these assumptions, the cost of thew warrant becomes a fee for false advertisements which always occurs for false ads.

Additional Variables

- Two slots, $Pos_a,\ Pos_b$
- $Q_i = 1:$ all quality is the same
- Two agents: a_1, a_2 , with value per view w_i, and reported/bidded value per view b_i.

The naive-modified VCG for warrants, stated generally enough for multi-slot ad auctions.

$$\begin{aligned} x_i &= \arg\max_{a\in A} \sum_{i\in N} \cancelto{1}{Q_i} b_i = \begin{cases} 1, & \text{if } b_i > b_{-i} \\ 0, & \text{otherwise} \end{cases} \\ t_i &= (Pos_{[i]} - Pos_{[i]-1})b_{-i} + Pos_{[i]}(1 - G_i)f \\ u_i &= Pos_{[i]}w_i - (Pos_{[i]} - Pos_{[i]-1})b_{-i} - Pos_{[i]}(1 - G_i)f \end{aligned}$$

3.2 Analysis

3.2.1 Warrant/fee condition to disincentivize false warrants

- IF $G_i = 1$, $u_i > 0$:

If the claim is true, we want the utility to be positive. Since the payment reduces to the same as regular VCG, which is IR, we get positive utility.

- IF $G_i = 0$, $u_i < 0$:

If the claim is false, then we want the claimant's utility to be negative. The warrant/fee must be large enough for this to hold.

$$\begin{aligned} 0 &> Pos_{[i]}w_i - (Pos_{[i]} - Pos_{[i]-1})b_{-i} - Pos_{[i]}(1 - \cancelto{0}{G_i})f \\ Pos_{[i]}f &> Pos_{[i]}w_i - (Pos_{[i]} - Pos_{[i]-1})b_{-i} \\ f &> w_i - \frac{Pos_{[i]} - Pos_{[i]-1}}{Pos_{[i]}} b_{-i} \end{aligned} \tag{5}$$

The above condition, (5), must be implemented by the mechanism designer. We see a few ways this could be done.

1. Use self-bid b_i to estimate w_i, and set $f = b_i - \frac{Pos_{[i]} - Pos_{[i]-1}}{Pos_{[i]}} b_{-i}$
2. Assume knowledge of the distribution of w_i's, such as the maximum $\bar{w}$. Set $f = \bar{w} - \frac{Pos_{[i]} - Pos_{[i]-1}}{Pos_{[i]}} b_{-i}$
3. Learn bidder's values for claims over multiple iterations of the ad auction. Increasing f each iteration until they stop bidding.

The below analysis shows non-strategyproofness of Option 1 in the single-slot (single item auction) setting, with $Pos_b = 0$.

3.2.2 Option 1 is not strategy proof

Assuming other bidders are truthful and have truthful claims, $G_{-i} = 1$

- $G_i = 1$ (i's claim is true):

Regular VCG setting which is strategyproof. If the other bidders were warranting false claims, then strategyproofness may or may no still hold.

- $G_i = 0$ (i's claim is false):

For simplicity assume $Pos_a = 1$, or think about analysing the per view utility rather than total utility.

$$u_i = Pos_{[i]} w_i - (Pos_{[i]} - \cancelto{0}{Pos_{[i]-1}}) b_{-i} - Pos_{[i]}(1 - \cancelto{0}{G_i})(w_i - \frac{Pos_{[i]} - \cancelto{0}{Pos_{[i]-1}}}{Pos_{[i]}} b_{-i})$$

$$u_i = Pos_{[i]}\left(w_i - b_{-i} - (b_i - b_{-i})\right) = \cancelto{1}{Pos_{[i]}}(w_i - b_i))$$

For simplicity assume $Pos_a = 1$, or think about analysing the per view utility rather than total utility.

- IF $w_i < w_{-i}$:
  * $b_i = w_i \implies u_i = 0$
  * $b_{-i} > b_i = w_i \implies u_i = 0$
  * $b_i > b_{-i} > w_i \implies u_i = w_i - b_i < 0$
- IF $w_i > w_{-i}$:
  * $b_i > w_i \implies u_i = w_i - b_i < 0$
  * $b_i = w_i \implies u_i = w_i - w_i = 0$
  * $w_i > b_i > b_{-i} \implies \boxed{u_i = w_i - b_i > 0}$
  * $b_{-i} > b_i = w_i \implies u_i = 0$

The boxed case shows that in the naive-warranted VCG mechanism, using Option 1 to estimate the necessary warrant/fee, f, when the bidder with the higher value wants to make a false claim, it is worthwhile for them to bid below their value (but above the next person's value) in order to reduce the warrant/fee charged to them s that they get an overall positive utility. Thus, it is neither strategyproof, nor negative utility for false claims.

This could potentially be fixed using 'more competition'. If the top bidder does not know how far below their bid the next bidder's value is, then they cannot reduce their bid without risk of losing the top slot. If there is a lot of competition, then this gap between the top valuation and the second valuation is likely small. Thus, the top bidder has a very small and risky window within which to strategically manipulate their bid. Under these conditions, the mechanism could become approximately strategyproof and approximately 0 utility for false claims.

The payment rule of the naive-warranted VCG mechanism (with guaranteed arbitration and perfectly accurate arbitration) looks nearly like the 'affine' or 'weighted' VCG mechanism. If the choice rule were also modified to the affine VCG, then we could entirely guarantee strategyproofness. However, this would require the affine shift (in our case f), to not depend in the bids. However, f could depend on which allocation is chosen. Thus, we could use Option 2 or 3 for setting f, and thus get non-positive utility for false claims. However, option 1 would once again likely break strategyproofness.

4 Simulation

The simulation is based on the model described in Theory I. Simulation allows us to make this model slightly more complex, and relax some strict, unrealistic assumptions we made in that section. Note: we retain the strong assumption that claims do not really interact with one another; there is no competition for viewership 'slots', for instance (as described in Theory II).

4.1 Moderate realism

Here, we still have a fixed warranting amount at $f = 100$.

We add two complexities to the Theory I model:

Additions to Theory I model

1. Utility per view (θ). This follows a random normal distribution independent of the truth of the claim. $\theta \sim N(0.5, 0.05)$. In Theory I, it is as if we had $\theta = 1$.
2. Virality (z): False claims are 1.5x more viral than True claims. For True claims, $Z \sim N(1, 0.1^2)$ and for False claims, $Z \sim N(1.5, 0.1^2)$.

Mathematically, this means that we replace the R ranking by

$$R' = R \cdot \theta \cdot z$$

So the original number of views R is scaled by virality and utility per view, and R' is used instead of R to calculate agents' utility.

Using this model, we consider three setups:

1. No arbitration, no warranting. This represents the current state of social media.
2. Arbitration, favorable exogenous variables.
3. Arbitration, unfavorable exogenous variables.

In all three cases, we simulation with numebr of claims $n = 10^4$, generated independently.

4.1.1 No arbitration

With no arbitration, we see the expected situation of False claims generally having higher utility than True claims. This is because they tend to be more viral.

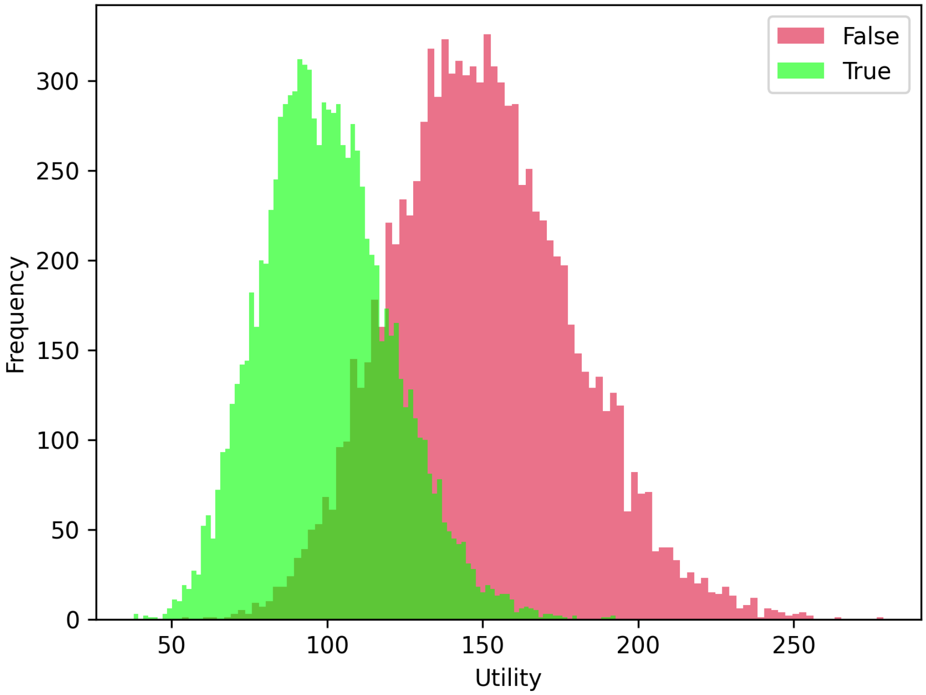


Figure 3: Utilities (no Arbitration)

4.1.2 Arbitration, favorable conditions

Favorable conditions

1. False claims have a higher probability of getting to arbitration than True claims (90% vs 40%).
2. Arbitration is imperfect, but has low false positive/false negative rates. $S_T = S_F = 0.9$.
3. Arbitrated True claims receive a 5x boost in views compared to unverified claims. $r_u = 100$ while $r_T = 500$.

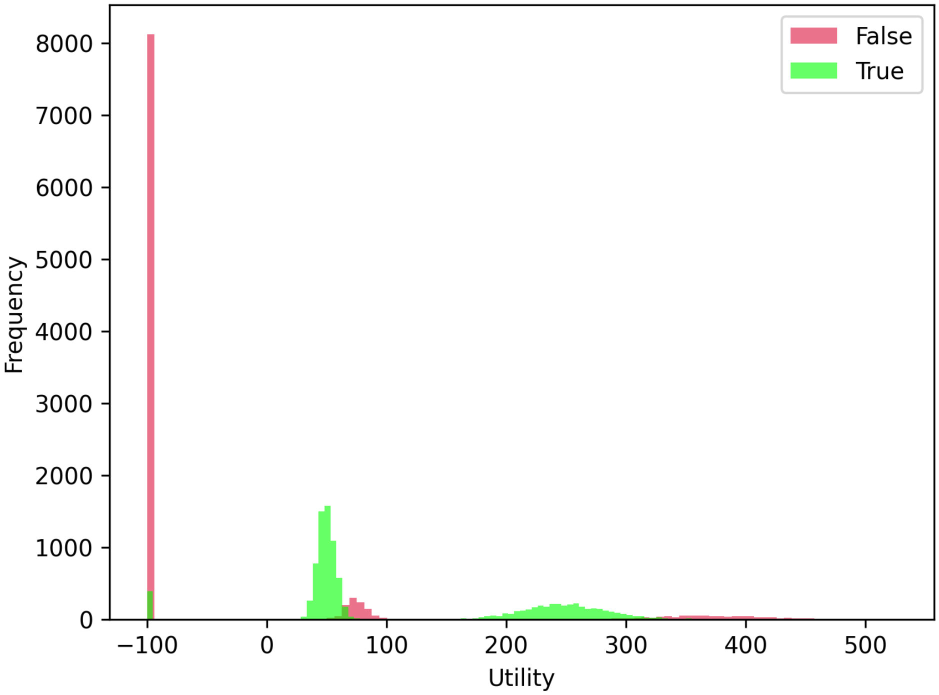


Figure 4: Utilities (Arbitration, Favorable)

Here, the distribution of utilities is very interesting. In particular, we see three clusters at which there are peaks.

- Cluster 1: The peak at -100 represents claims that were arbitrated, judged False, and had to pay fine $-f$. There are some True claims in this stack, but the vast majority are False claims since the arbitration is working well.
- Cluster 2: Then there are two peaks with utilities between 25 and 100 or so. These are claims which did not reach arbitration. They get r_u views, scaled by their virality and utility per view. Since False claims are more viral, their utility distribution is shifted to the right compared to True claims. But since most of them get to arbitration and are determined to be False, there are far fewer of them here compared to True claims.
- Cluster 3: There are two peaks on the right. These are claims that were arbitrated, judged True, and received r_T base views scaled by virality and utility per view. There are a few False claims that slip through this process, but the vast majority of claims that end up having this high utility are True claims, as we would like. This is a big difference to the setup without warranting and arbitration, and provides a hope for how this mechanism would work in an ideal case, punishing falsehoods and promoting truth.

In this case we also have $\hat{U_F} = -41.62$ and $\hat{U_T} = 115.62$. So we satisfy the requirement that there is a negative utility to posting False claims, and positive utility to posting True claims.

4.1.3 Arbitration, unfavorable conditions

Unfavorable conditions

1. False claims have the same probability of getting to arbitration as True claims (50% each)
2. Arbitration is only slightly better than chance: $s_T = s_F = 0.6$.
3. Arbitrated True claims receive a 5x boost in views compared to unverified claims. $r_u = 100$ while $r_T = 500$.

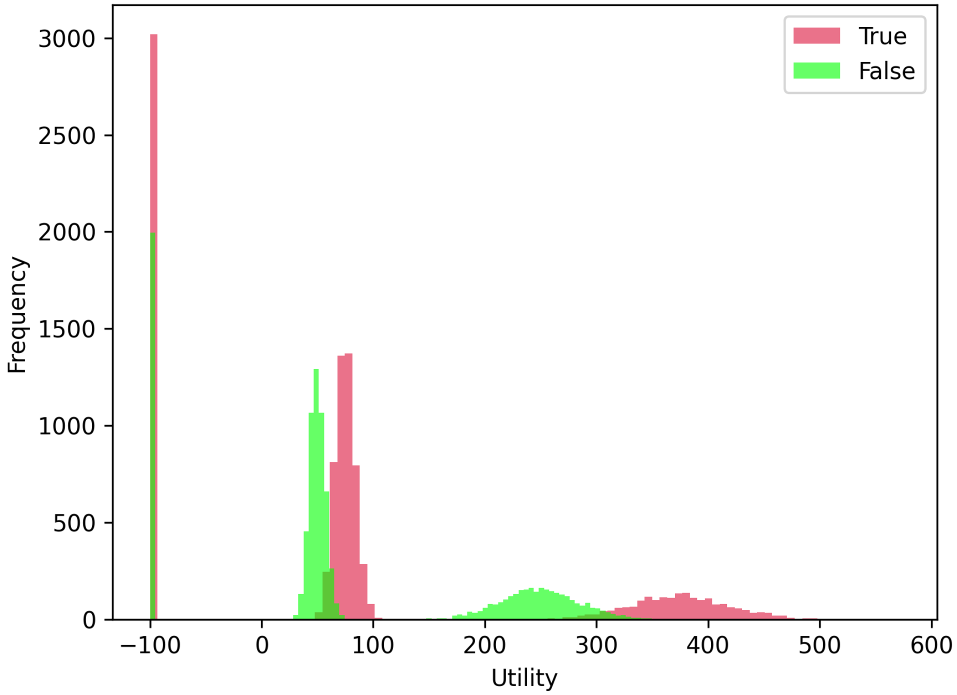


Figure 5: Utilities (Arbitration, Unfavorable)

Here, we see a similar situation in terms of the placement of peaks as we had with favorable exogenous variables, but now many False claims are receiving high utilities, since they are not stopped by accurate arbitration. Many true claims are also unfairly punished and pay the fine f. In this case $\hat{U_F} = 81.79$ and $\hat{U_T} = 78.75$. So this represents how the mechanism might fail.

4.2 Budget-constrained truth-tellers, wealthy liars

Now, we consider the situation where truth-tellers are budget constrained, but those spreading falsehoods have more money to spend on warranting their claims to give them the appearance of credibility and boost their rankings. This is a concerning situation and we are interesting in how our system would handle it.

Formally, we make the following modifications to the model:

1. Truth-tellers warrant their claim for $f = 100$. False-tellers warrant their claims for $f = 300$.
2. f plays a role in the ranking step. Numbers of views are scaled proportionally to f. So warranting more gives you more views.
3. f helps to determine whether a claim is arbitrated or not. Claims with higher f have a higher probability of getting to arbitration step. We now have

   $$a'_F = a_F + (1 - a_F) \times \left(\frac{1}{1 + e^{-0.05 \cdot (f - 200)}} \right)$$

   This uses a recentered and rescaled sigmoid function to ensure that false claims that have a high amount warranted on them will go to arbitration with a higher probability.

Surprisingly, in this setup the model becomes even more effective than before.

4.2.1 Arbitration, favorable conditions

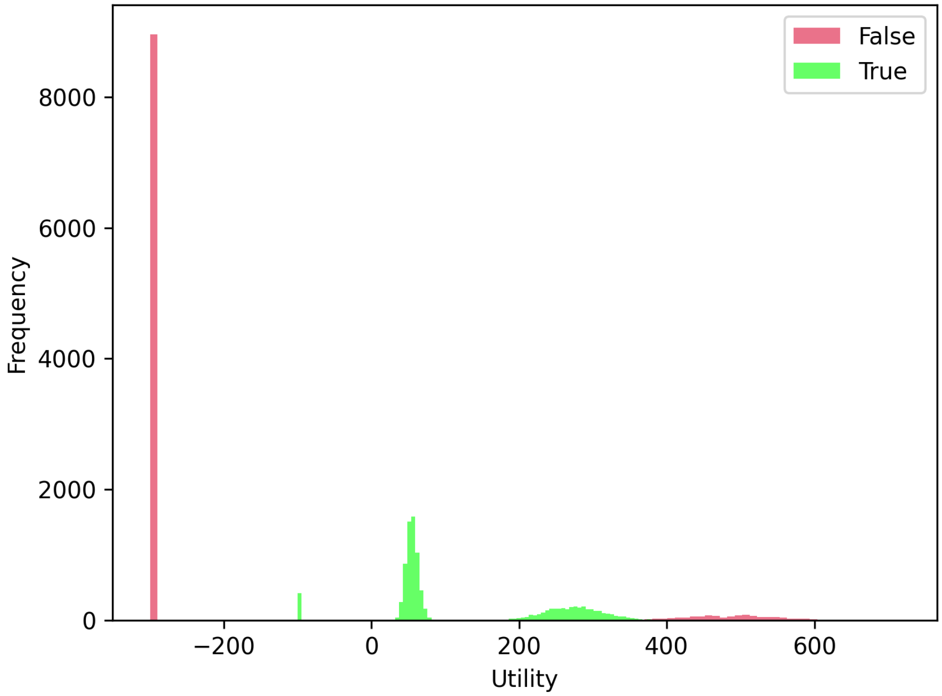


Figure 6: Utilities (Arbitration, Favorable, Budget-Constrained Truth Tellers)

In this case $\hat{U_T} = 128.72$ and $\hat{U_F} = -218.44$. This is an even bigger difference than in the favorable case we had previously, where both truth-tellers and liars were warranting the same amount.

It must be noted that this happens because of how aggressively False and high-warranted claims are arbitrated. However, the intuition that the mechanism can avoid favoring wealthy spreaders of misinformation if higher rankings are counterbalanced by greater probability of arbitration is valuable. It suggests that the design of the filtering system is important, so that it indeed has this property and incentivizes social network participants to preferentially challenge claims with large warrants.

4.2.2 Arbitration, unfavorable conditions

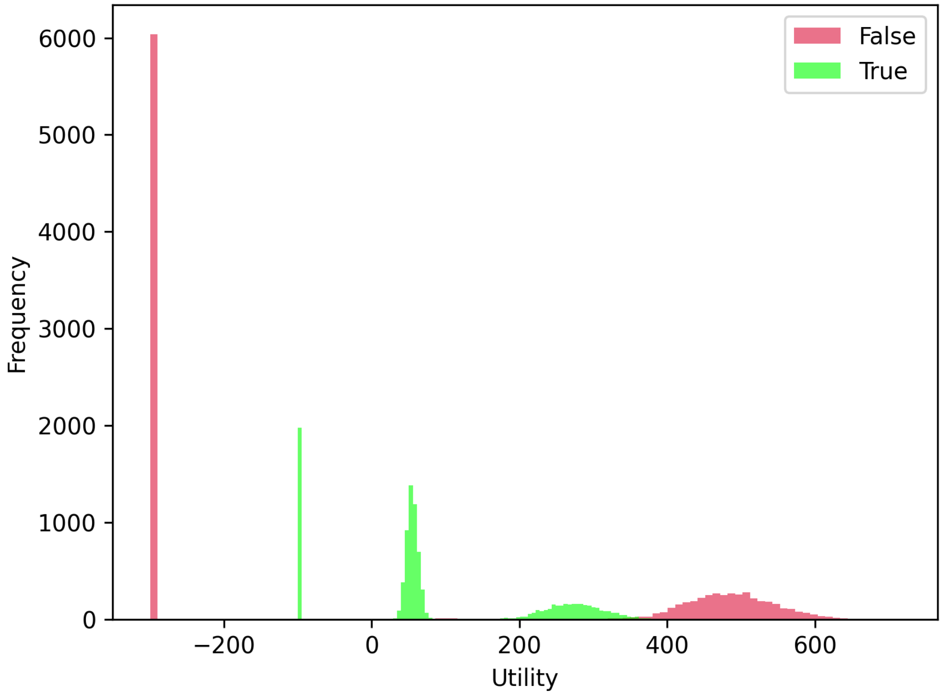


Figure 7: Utilities (Arbitration, Unfavorable, Budget-Constrained Truth Tellers)

In this case $\hat{U_T} = 89.53$ and $\hat{U_F} = 10.14$. Although there is still positive utility to making false claims, it is much smaller than before, because the benefit of the claim having more views is outweighed by its higher probability of going to arbitration and potentially losing the entire warranted amount f.

5 Next Steps & Ideas

In this project, we considered simplified versions of a truth warranting system. To continue this research we would look at more complicated aspects.

5.1 Arbitration Mechanism

While in this project we decided to assume the existence of an arbitration mechanism with a certain rate of accuracy, the real world would be more complicated. Things we may go on to consider would be the cost of the arbitration process, how to pay for it, how to design it to account for different levels of warranting.

5.2 Anti-Warranting & Filtering

Right now we simplify which statements are filtered based on their ground truth signal, and we assume that false statements are more likely to be arbitrated. In future we may want to set up a system of anti-warranting, which filters statements on the number of agents who stake money on the other side of the arbitration. A theoretical investigation of the properties of such a system would be of interest.

5.3 Simulation

Currently, actors in the simulation do not interact with each other and only make one statement. We could develop the simulation that allows multiple statements and thereby allows strategic competition and collaboration between players.

5.4 Price of misinformation

The harm of a false claim is not always the same. Perhaps we could explore pricing of false claims using the VCG mechanism, using some elicitation by members of the social network?

5.5 Reputation Systems

If some actors are continuously making false claims (even if they are paying the cost for them), they should probably be punished at the level of the actor also, rather than at the level of a claim. A reputation system of sorts could be introduced to give longer-term viewership penalties to those who consistently spread lies.

References

[1] Marshall Van Alstyne. Free speech, platforms & the fake news problem. *SSRN*, 2022. Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3997980`.