Malicious Misinformation

Jun-Yuan 'Jay' Chen

April 2024

1 Introduction

Misinformation has become an increasingly more pressing issue in society with the rapid rise of information distribution and sharing technology (like social media). The spread of misinformation has created real, tangible effects both on the economy and society. In the United States misinformation surrounding the COVID-19 vaccine slowed its admistration rate and resulted in lower economic activity [8] [11]. There are many other examples of the harms of misinformation; the insurrection on January 6th at the United States captiol is widely attributed to the misinformation surrounding the 2020 United States presidential election.

This misinformation spreads on social media sites such as Facebook and X (formerly Twitter). Peer-to-peer information exchange and individual decision making has been well-studied by the information aggregation literature both in market and non-market settings. Information aggregation traditionally has been studied with agents who are all incentivized to reveal their partial information truthfully either through an incentive compatible market structure or other mechanism. However, an extension of this method for heterogeneous agents, specifically with agents who are incentivized to act deceptively, has been far less developed. The agents acting in this specific manner in the world has become increasingly more relevant and widespread as the power of misinformation and its effects mount.

The study of malicious agents in a distributed system, especially with regards to information, is becoming more a necessity than just an area of interesting research. With the rise in popularity of the use of multiple social media platforms, the existence and regulation of the industry requires the guidance of research for effective policy design. Being able to protect social media users from misinformation is an issue that policymakers are currently facing. Experiments lend itself well as a methodology in studying misinformation since the truth is often subjective and relative. This would make it difficult to pin down and properly observe how "close" the system can get to the truth when looking at information aggregation from a general standpoint.

2 Literature Review

2.1 Malice

Malice is not precisely defined in the existing economics literature; most works refer to malicious agents as those who erode the efficiency of the system, some times defined in terms of Pareto efficiency.

The Price of Malice (PoM) is defined as the measure of to what degree the introduction of malicious agents to a distributed system will negatively impact the efficiency [15]. Using a virus inoculation game as their scenario and a peer-to-peer network, the authors established several game theoretic results namely the malicious Nash equilibrium and, the previously mentioned, PoM. Malicious Nash equilibrium (MNE) is when no non-malicious players can decrease their perceived cost by unilateral deviation; the perceived individual cost is the cost that individual players expects to incur given their knowledge of malicious players. They explore two main settings: one in which the non-malicious players are not aware of malicious players (oblivious) and one in which they are aware (non-oblivious). In the second setting, players are knowledgeable of the number of malicious nodes in the network but not their locations or strategies. The price of malice is defined as ratio between the price of anarchy in the case with zero and b malicious players in problem instance I.

$$PoM(I, b) = \frac{PoA(I, b)}{PoA(I, 0)}$$

The social costs increase monotonically with the number of malicious nodes in the network under the oblivious model, whereas in the non-oblivious model the knowledge of malicious players in the system encourages the other players to cooperate and push the price of malice below 1.

It is important to note that the previously discussed model with virus diffusion only deals with two types of agents: selfish, those who act in to maximize their own utility, and malicious, those who act to increase the overall social cost. We are interested in how information disseminates among individuals in a system, which is modeled more closely using a pairwise network and informational updating [1]. Rather than malicious agents as defined by Moscibroda, et al, this theoretic model employs "forceful" agents who do not deriving utility from increasing social cost, but simply do not update their beliefs when introduced to new information. They use a similar measure of malicious intervention by evaluating the efficiency of the informational aggregation in the system with and without these kinds of agents. The main results of their model are that the effect of misinformation can be greatly limited when the number and impact of forceful agents is small and that the spread of misinformation is sensitive to the location of the forceful agents.

Both works discussed above focus on the theoretical framework when malicious agents are involved; they focus on efficiency of the system to obtain a (social) optimum. In practice, this social optimum can be described with Pareto efficiency [3]. Now examining malice (and envy) under a more behavioral lens, malice can be defined as an individual blocking a Pareto superior allocation, if they are not the one increasing their utility and the one who would increase is lower than them at the current allocation. Beckman, et al experimentally verifies

that the presence of both malicious and envious behaviors can pull the resulting allocation from the social optimum.

We want to treat malice as defined in the theoretical bodies of work; malicious agents are those who derive utility from increasing the overall social cost and their impact evaluated by a measure of distance from the social optimum (PoM).

2.2 Information Aggregation

Information aggregation has been widely studied, particularly as it applies to asset and financial markets where the traders have differing partial information. The rational expectations (RE) model fundamentally states that the traders in a market will condition their beliefs on the market prices; this model has experimental support [13]. Additionally, the experimental support explored some robustness in market mechanisms, explicitly testing in trading Arrow-Debreu securities and double auction trading [14]. However, they showed that the market may fail under some conditions experimentally. Jointly sufficient conditions for successful convergence to the RE equilibrium are trading experience and common knowledge of dividends [10].

With information aggregation experimentally well-tested under the rational expectations model, its performance under more realistic conditions becomes the more important question. This spurred a study conducted at Hewlett-Packard (HP) where they implemented an information aggregation mechanism (IAM) to show that the IAM out-performs tradition business forecasting techniques [12]. The IAM used in their study was closely related to the existing theoretical and experimental literature stated above as to draw conclusions on the business applicability. The experiment was conducted with participants from different areas within the HP business operation to increase the likelihood of information heterogeneity among them. Their task was to buy and sell Arrow-Debreu securities based on predicted sales levels through a double auction with the payoff determined by actual realization of sales by the company. In the end, they were able to show, broadly speaking, that IAM performed better forecasting than the traditional methods of collecting information (like business meetings).

A substantial portion of the literature regarding information aggregation falls within the market paradigm. However, from the examples illustrated in the introduction, information aggregation extends beyond the scope of just markets. By tuning down the market setting and focusing on a more simplistic game, the information aggregation mechanism itself and the way that the individuals process that information can be more closely examined; the behavior the market or other IAM can be model with the micro foundations of how each individuals process and utilizes new information.

2.3 Behavioral Elements in Distributed System

Understanding the micro foundations of how each individual processes new information and signals will allow the study of information aggregation to extend past just markets. However, when changing the approach to a microlevel, the behavioral elements of each individual involved in the aggregation and other possible influx of information must be accounted for. An individual's understanding of a stochastic process (their beliefs) can

be modeled as a distribution and can be updated with Bayes rule. Risk attitudes of the individuals can affect the performance of the IAM but can be correct in the prediction when using a non-linear aggregation that takes into account their risk behavior; this requires that their risk behaviors are truthfully revealed via incentive compatible means [4]. Public knowledge can also bias the IAM predictions but, again, can be accounted by using a non-linear aggregation system that corrects the bias in a two-stage game that first reveals that bias truthfully [5].

When studying information aggregation from the bottom-up, individual issues and biases become more apparent, but also lends better towards finer changes to the system to mirror reality. A realistic change to the existing literature is the addition of heterogeneous utility for agents; the introduction malicious agents to study how they may affect the efficiency of the information aggregation system.

3 Theory

3.1 Environment

- set of players $I = \{1, ..., N\}$
- set of player types $\Theta = \{\text{Truthful}, \text{Deceitful}\}$
- set of possible states of nature Ω where $\Omega = \{A, B\}$.
- the set of actions for each player type is:
  - truthful players is $A_{\text{Truthful}} = \Omega \times \Omega$ (they choose a message and a guess).
  - deceitful players is $A_{\text{Deceitful}} = [0, 1]$ (they choose the probabilty in they send a dishonest message).
- probablity distribution over types is (p_T, p_D) corresponding to $\Theta = \{\text{Truthful}, \text{Deceitful}\}$ where $p_T + p_D = 1$.
- probability distribution over the true states $p(\omega) > 0, \forall \omega \in \Omega$ where $\sum_{\omega \in \Omega} p(\omega) = 1$.
- the payoff for each player type is:
  - truthful players is $u_i^{\text{Truthful}} : A_G \times \Theta \times \Omega \to \mathbb{R}$
  - deceitful players is $u_i^{\text{Deceitful}} : A_B \times \Theta \times \Omega \to \mathbb{R}$

Consider the following game:

1. Nature randomly determines the true state $\bar{\omega}$ according to $P(\bar{\omega} = A) = P(\bar{\omega} = B) = \frac{1}{2}$. The distribution from which the players receive their signals is dependent from the realization of the true state. The true state is not known to Truthful players, but is known to Deceitful players.
2. Nature randomly determines the type of each player according to the probabilty distribution over types.

3. Each player recieves one signal randomly drawn from the distribution F_ω denoted as s_i. We impose that the probablity of observing the the true state $\bar{\omega}$ in a signal is higher than the other state.

$$F_{\bar{\omega}}(\bar{\omega}) > F_{\bar{\omega}}(\omega), \forall \omega \neq \bar{\omega}$$

4. Each player chooses a message $m_i : \Omega \to \Omega$.

5. Every player observes the vector of all messages $M = \bigcup_{i=1}^{N} m_i$.

6. Each Truthful player chooses $x_i : \Omega \times \Omega^{n(M)-1} \to \Omega$ (they use their signals and messages to make their guess).

7. Players of type $\theta_i =$ Truthful receive payoffs according to the function:

$$u_i^{\text{Truthful}}(x) = \frac{\sum_i \mathbb{1}\{x_i = \bar{\omega} \text{ and } \theta_i = \text{Truthful}\}}{\sum_i \mathbb{1}\{\theta_i = \text{Truthful}\}}$$

   All Truthful players will recieve the same payoff; they get the fraction of Truthful players that choose the true state correctly.

8. Players of type $\theta_i =$ Deceitful receive payoffs according to the function:

$$u_i^{\text{Deceitful}}(x) = 1 - u_j^T(x) \text{ where } \theta_j = \text{Truthful}$$

   All Deceitful players will recieve the same payoff; they get the fraction of Truthful players that choose the true state incorrectly.

9. We treat the last step as "voting" where the truthful players vote on which state they believe is the true one to evaluate the performance of the system.

3.2 Two States, Two Players with Symmetric Distributions

3.2.1 Set-Up

In this setting, we set $N = 2$ and $F_T = F_A(A) = F_B(B) > F_{NT}| = F_A(B) = F_B(A)$ with $P_A = P_B = \frac{1}{2}$.

3.2.2 Bayesian Equilibrium

At equilibrium, we expect both players to maximize their payoff according to their beliefs on the true state. In the case with one signal and one messages, we can make some simplifying assumptions. First, we assume that the strategy is symmetric for all deceitful players and, second, that truthful players will tell the truth all the time (from rational expectation model). The truthful player's expect value looks like

$$\mathbb{E}[\bar{u}_i^T(x_i)|s_i, m_{-i}] = \frac{P(\bar{\omega} = x_i)F_{x_i}(s_i)[p_T F_{x_i}(m_{-i}) + p_D P(m_{-i}|\bar{\omega} = x_i, \theta_{-i} = D)]}{\sum_{\omega \in \Omega} P(\bar{\omega} = \omega)F_\omega(s_i)[p_T F_\omega(m_{-i}) + p_D P(m_{-i}|\bar{\omega} = \omega, \theta_i = D)]}$$

Since player i of type Truthful will choose x_i to maximize their expected utility, we have

$$x_i^*(s_i, m_{-i}) = \begin{cases} A & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] > \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \\ [A, B] & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] = \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \\ B & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] < \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \end{cases}$$

We set it so that if the expect value of each choice is equal, they will mix over the states with any probability distribution. We can evaluate the inequality explicitly.

$$\begin{aligned} &F_A(s_i)[p_T F_A(m_{-i}) + p_D p_A(m_{-i})] \square F_B(s_i)[p_T F_B(m_{-i}) + p_D p_B(m_{-i})] \\ &F_A(s_i)[p_T F_A(m_{-i}) + p_D p_A(m_{-i})] \square F_B(s_i)[p_T F_B(m_{-i}) + p_D(1 - p_A(m_{-i}))] \end{aligned}$$

where $p_{x_i}(m_i) = P(m_{-i}|\bar{\omega} = x_i, \theta_{-i} = D)$. Then, for the deceitful player to maximize their utility, they will minimize the probability that the truthful player chooses the correct state.

$$\max_{P(m_i=B)} P(x_{-i} = B) = \max_{P(m_i=B)} P\left(\mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] \leq \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}]\right)$$

3.2.3 Truthful Player Best Response

Since the utility of the truthful players depend on other truthful players correctly choosing the true state, we assume that they will always reveal their information truthfully, $m_i^T(s_i) = s_i$ and $P(m_i^T(s_i)) = F_{\bar{\omega}}(s_i)$. The utility of truthful players is separable in x_i, then since we hold all other $x_{i'}, i' \neq i$ fixed, we can just evaluate the term in the u_i^T that varies with x_i since constant shifts are irrelevant in the comparison between expect utility between choices of x_i. Thus, we will evaluate:

$$\bar{u}_i^T(x_i) = \begin{cases} 1 & \text{if } x_i = \bar{\omega} \\ 0 & \text{if } x_i \neq \bar{\omega} \end{cases}$$

The best response maps to $\{A, B\}$ if the expected value of the two choices are equivalent since the player will be indifferent to each choice and any mixture over the two options will be the best response. Otherwise, the player will always have a pure strategy.

3.2.4 Deceitful Player Best Response

Deceitful players seeks to minimize the utility of the truthful players in order to maximize their own; they want to send messages that would increase the chances that the truthful players choose the true state incorrectly. However, the deceitful players can only do this through their message which they send before observing all other messages. They will have to expect the expected utility of truthful players over all possible signals and message sets.

When $N = 2$, the deceitful player does not know if there is another deceitful player but we have assumed that they will choose the same probability of lying regardless. Their objective function is to maximize the probability that the expected utility of the incorrect state is higher than the expected utility of the true state for the truthful players. Since again

we have assumed homogeneity in the truthful players, maximizing the probability of one of them will be the same for all of them.

$$\max_{P(m_i=\omega),\forall\omega\in\Omega} P(x_{-i} \neq \bar{\omega})$$

Since in our setting, we only have two possible states of nature $\{A, B\}$, the objective function can be written more explicitly. We let $\bar{\omega} = A$ (without loss of generality) and with only two possible states, let $p = P(m_i = B)$ which is the probability of lying. Furthermore, we simplify our setting more by setting $F_A(A) = F_B(B) = F_T$ and $F_A(B) = F_B(A) = F_{NT}$; that is, the probability distribution is symmetric between for either state being true. The deceitful player objective function is

$$\max_{P(m_i=B)} P(x_{-i} = B) = \max_{P(m_i=B)} P\left(\mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] \leq \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}]\right)$$

Claim 3.1. *The objective function of the deceitful player can be written as*

$$Q_D(p) = \max_p (1-p)F_A(A)E_{AA}(p) + pF_A(A)E_{AB}(p) + (1-p)F_A(B)E_{BA}(p) + pF_A(B)E_{BB}(p)$$

where $p = P(m_i = B|\theta_i = D)$.

Proof. We have that the deceitful player objective function is

$$Q_D(p) = \max_p P\left(\mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] \leq \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}]\right)$$

We can see that there are 4 possible realizations of the $\mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] \leq \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}]$, which will correspond to the realizations of $(s_i, m_i) \in \Omega \times \Omega$. The 4 cases have the following probabilities

$$\begin{aligned}
P(s_i = A, m_{-i} = A) &= F_A(A)P(m_{-i} = A|\theta_{-i} = D) = F_A(A)(1-p) \\
P(s_i = A, m_{-i} = B) &= F_A(A)P(m_{-i} = B|\theta_{-i} = D) = F_A(A)p \\
P(s_i = B, m_{-i} = A) &= F_A(B)P(m_{-i} = A|\theta_{-i} = D) = F_A(B)(1-p) \\
P(s_i = B, m_{-i} = B) &= F_A(B)P(m_{-i} = B|\theta_{-i} = D) = F_A(B)p
\end{aligned}$$

and corresponding to the events:

$$E_{AA}(p) = \begin{cases} 1 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = A, m_{-i} = A] < \mathbb{E}[u_i^T(x_i = B)|s_i = A, m_{-i} = A] \\ 1/2 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = A, m_{-i} = A] = \mathbb{E}[u_i^T(x_i = B)|s_i = A, m_{-i} = A] \\ 0 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = A, m_{-i} = A] > \mathbb{E}[u_i^T(x_i = B)|s_i = A, m_{-i} = A] \end{cases}$$

$$E_{AB}(p) = \begin{cases} 1 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = A, m_{-i} = B] < \mathbb{E}[u_i^T(x_i = B)|s_i = A, m_{-i} = B] \\ 1/2 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = A, m_{-i} = B] = \mathbb{E}[u_i^T(x_i = B)|s_i = A, m_{-i} = B] \\ 0 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = A, m_{-i} = B] > \mathbb{E}[u_i^T(x_i = B)|s_i = A, m_{-i} = B] \end{cases}$$

$$E_{BA}(p) = \begin{cases} 1 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = B, m_{-i} = A] < \mathbb{E}[u_i^T(x_i = B)|s_i = B, m_{-i} = A] \\ 1/2 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = B, m_{-i} = A] = \mathbb{E}[u_i^T(x_i = B)|s_i = B, m_{-i} = A] \\ 0 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = B, m_{-i} = A] > \mathbb{E}[u_i^T(x_i = B)|s_i = B, m_{-i} = A] \end{cases}$$

$$E_{BB}(p) = \begin{cases} 1 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = B, m_{-i} = B] < \mathbb{E}[u_i^T(x_i = B)|s_i = B, m_{-i} = B] \\ 1/2 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = B, m_{-i} = B] = \mathbb{E}[u_i^T(x_i = B)|s_i = B, m_{-i} = B] \\ 0 & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i = B, m_{-i} = B] > \mathbb{E}[u_i^T(x_i = B)|s_i = B, m_{-i} = B] \end{cases}$$

Then, we can write

$$Q_D(p) = \max_p (1-p)F_A(A)E_{AA}(p) + pF_A(A)E_{AB}(p) + (1-p)F_A(B)E_{BA}(p) + pF_A(B)E_{BB}(p)$$

□

Each of these events is dependent on p, so the objective of the deceitful player is to choose p such that the summation of satisfied events' probabilities are the highest. To simplify the notation, we set the bounds of p for each event which can be simplified using $F_A(A) = F_B(B) = F_T$ and $F_A(B) = F_B(A) = F_{NT}$:

Claim 3.2. *The solution to the optimization problem the deceitful player is indifferent among the set of* $p^* \in [F_{NT}, F_T + \frac{p_T}{p_D}(2F_T - 1)$ *with an objective function value of* F_{NT}.

Proof. With some algebra, we find that the p which satisfies the events are:

$$E_{AA}(p) = \begin{cases} 1 & \text{if } F_T + \frac{p_T}{p_D}(2F_T - 1) < p \leq 1 \text{ and } \frac{F_T - F_{NT}}{F_T} < p_D \leq 1 \\ 1/2 & \text{if } p = F_T + \frac{p_T}{p_D}(2F_T - 1) \text{ and } \begin{cases} \frac{F_T - F_{NT}}{F_T} \leq p_D \leq 1 \\ p_T = 0 \end{cases} \\ 0 & \text{otherwise (including } p_T = 1) \end{cases}$$

$$E_{AB}(p) = \begin{cases} 1 & \text{if } 0 \leq p < F_{NT} \\ 1/2 & \text{if } \begin{cases} 0 \leq p \leq 1 \text{ and } p_D = 0 \\ p = F_{NT} \text{ and } 0 \leq p_T < 1 \end{cases} \\ 0 & \text{otherwise} \end{cases}$$

$$E_{BA}(p) = \begin{cases} 1 & \text{if } F_{NT} < p \leq 1 \\ 1/2 & \text{if } \begin{cases} 0 \leq p \leq 1 \text{ and } p_D = 0 \\ p = F_{NT} \text{ and } 0 \leq p_T < 1 \end{cases} \\ 0 & \text{otherwise} \end{cases}$$

$$E_{BB}(p) = \begin{cases} 1 & \text{if } \begin{cases} 0 \leq p \leq 1 \text{ and } 0 \leq p_D \leq \frac{F_T - F_{NT}}{F_T} \\ 0 \leq p < F_T + \frac{p_T}{p_D}(2F_T - 1) \text{ and } \frac{F_T - F_{NT}}{F_T} \leq p_D \leq 1 \end{cases} \\ 1/2 & \text{if } p = F_T + \frac{p_T}{p_D}(2F_T - 1) \text{ and } \begin{cases} p_T = 0 \\ 0 < p_T < 1 \text{ and } F_T \leq \frac{1}{1+p_T} \end{cases} \\ 0 & \text{otherwise} \end{cases}$$

Now, the optimal p can be solved for with this explicit optimization. Each event has some bounds for p in which they are > 0. We will call these bounds

$$C_1 = F_T + \frac{p_T}{p_D}(2F_T - 1) \text{ and } C_2 = F_{NT}$$

and we have that $C_1 > C_2$. Thus, we can evaluate the Deceitful player's objective function based on the region of p.

- $p_1^* = \arg \max_{C_1 < p \leq 1} (1-p)F_T + (1-p)F_{NT} = C_1 + \varepsilon$
- $p_2^* = \arg \max_{p=C1} (1-p)F_T/2 + (1-p)F_{NT} + pF_{NT}/2 = C_1$
- $p_3^* = \arg \max_{C_2 < p < C_1} (1-p)F_{NT} + pF_{NT} = (C_2, C_1)$
- $p_4^* = \arg \max_{p=C2} pF_T/2 + (1-p)F_{NT}/2 + pF_{NT} = C_2$
- $p_5^* = \arg \max_{0 \leq p < C_2} pF_T + pF_{NT} = C_2 - \varepsilon$

where $\varepsilon > 0$. We have the value of the objective function at these points to be

$$\begin{aligned}
Q_D(p_1^*) &= 1 - (C_1 + \varepsilon) \\
&= 1 - F_T - \frac{p_T}{p_D}(2F_T - 1) - \varepsilon \\
Q_D(p_2^*) &= (1 - C_1)F_T/2 + (1 - C_1)F_{NT} + C_1F_{NT}/2 \\
&= \frac{1}{2}(1 - C_1) \\
Q_D(p_3^*) &= C_2 = F_{NT} \\
Q_D(p_4^*) &= C_2F_T/2 + (1 - C_2)F_{NT}/2 + C_2F_{NT} \\
&= C_2 \\
Q_D(p_5^*) &= C_2 = F_{NT} - \varepsilon
\end{aligned}$$

We find that $Q_D(p_3^*) = Q_D(p_4^*) > Q_D(p_1^*) > Q_D(p_2^*) > Q_D(p_5^*)$ so the solution is $p^* \in [C_2, C_1)$. □

We see that the expected value of the $p^* \in [C_2, C_1)$ is F_{NT}, the probability of the false state. Meaning that even with $C_1 > p > F_{NT}$ (slightly lying), the deceitful player still cannot do better than not lying at all. Since $Q(p^*) = F_{NT}$, the best response of the deceitful player is not increase their expected value past just the chance of the truthful player listening to their signal.

Let us consider two different tie breaker scenarios: one where the truthful player will simply choose A and one where they will simply choose B. First we examine if they choose A in a tie, the truthful players will choose with

$$x_i^{A*}(s_i, m_{-i}) = \begin{cases} A & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] \geq \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \\ B & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] < \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \end{cases}$$

The deceitful player best response analysis will be the same until we evaluate the regions of the objective function. Now, all the event functions we had will be zero at indifference rather than 1/2. Then, the solution will be (C_2, C_1) losing inclusion of the lower bound in the solution. Conversely, if the truthful player will choose B in a tie,

$$x_i^{B*}(s_i, m_{-i}) = \begin{cases} A & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] > \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \\ B & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] \leq \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \end{cases}$$

Again, the deceitful player best response analysis is very similar, but now yielding $p^* = C_2$.

In summary, in the $N = 2$ case the deceitful player is unable to sway the truthful player at all; therefore their best strategy is just trying to make sure they do not give away too much information in their signal accidentally. This theoretical result would align with intuition since they should trust their own signal more than a single message which could be a lie. However, this leads into whether or not multiple messages (which have some probability of being lies) may override a truthful player's own signal.

3.3 General N and Two States

Now, extending our $N = 2$ model to a general number of N is fairly straightforward; there are three major changes when generalizing the number of players. The first is that the truthful player expected value must include updating from all the additional players all of which also have the possibility of being deceitful.

$$\mathbb{E}[\bar{u}_i^T(x_i)|s_i, m_{-i}] = \frac{P(\bar{\omega} = x_i)F_{x_i}(s_i)\prod_{m\in m_{-i}}[p_T F_{x_i}(m) + p_D P(m|\bar{\omega} = x_i, \theta_{-i} = D)]}{\sum_{\omega\in\Omega} P(\bar{\omega} = \omega)F_\omega(s_i)\prod_{m\in m_{-i}}[p_T F_\omega(m_{-i}) + p_D P(m_{-i}|\bar{\omega} = \omega, \theta_i = D)]}$$

It can be observed that the probability of each message is simply raised to the $N - 1$.

Claim 3.3. *Every truthful player will have $s_i = m_i$.*

Proof. Each truthful player's utility is a function of all other truthful player's choice of x_i; specifically, they receive a higher payoff for each truthful player that guesses the true state correctly. Since, all truthful players are homogeneous, we can take a pair of two truthful agents that would be representative of an interaction between any pair of them (WLOG we will denote them as $i = 1$, $i = 2$).

They each will have a signal before they send their messages, s_1 and s_2. When player 1 receives their signal, by Bayes rule, we will have $\mathbb{E}[\bar{u}_1^T(x_1 = s_1)|s_1] > \mathbb{E}[\bar{u}_1^T(x_1 \neq s_1)|s_1]$ since the prior is uniform; player 1 believes that s_1 is the most likely true state. Player 1 will receive a higher payoff if player 2 chooses the correct state and their message m_1 will influence that choice. So, player 1 will choose a message that increases the probability that player 2 will choose s_1 since at this stage player 1 believes s_1 is the most likely true state.

After receiving signal s_1, let player 1 choose their message $m_1 = s_1$ with probability q. When $q = 1$ player 1 send s_1 for sure and when $q = 0$ they will not send s_1 at all. We just add one more possibility to truthful player signal branch and we get that player 2's expect value of choose $x_2 = s_1$ with q from player 1's perspective is

$$\frac{P(\bar{\omega} = s_1)F_{s_1}(s_2)\prod_{m\in m_{-i}/m_1}[p_T F_{s_1}(m) + p_D p(m)] \times [p_T(F_{s_1}(s_1)q + (1 - F_{s_1}(s_1))(1 - q)) + p_D p(m_1)]}{\sum_{\omega\in\Omega} P(\bar{\omega} = \omega)F_\omega(s_2)\prod_{m\in m_{-i}/m_1}[p_T F_\omega(m) + p_D p(m)] \times [p_T(F_\omega(s_1)q + (1 - F_\omega(s_1))(1 - q)) + p_D p(m_1)]}$$

$$\frac{P(\bar{\omega} = s_1)F_{s_1}(s_2)\prod_{m\in m_{-i}/m_1}[p_T F_{s_1}(m) + p_D p(m)] \times [p_T((2F_{s_1}(s_1) - 1)q + 1 - F_{s_1}(s_1)) + p_D p(m_1)]}{\sum_{\omega\in\Omega} P(\bar{\omega} = \omega)F_\omega(s_2)\prod_{m\in m_{-i}/m_1}[p_T F_\omega(m) + p_D p(m)] \times [p_T((2F_\omega(s_1) - 1)q + 1 - F_\omega(s_1)) + p_D p(m_1)]}$$

This expression is very complicated but we are only interested in the behavior with respect to q, so we can rewrite this expectation as

$$\frac{\varphi_{s_1} q + A_{s_1}}{\sum_{\omega\in\Omega}(\varphi_\omega q + A_\omega)}$$

where

$$\varphi_\omega = P(\bar{\omega} = \omega)F_\omega(s_2) \prod_{m \in m_{-i}/m_1} [p_T F_\omega(m) + p_D p(m)] \times p_T((2F_\omega(s_1) - 1)$$

$$A_\omega = P(\bar{\omega} = \omega)F_\omega(s_2) \prod_{m \in m_{-i}/m_1} [p_T F_\omega(m) + p_D p(m)] \times [p_T(1 - F_\omega(s_1)) + p_D p(m_1)]$$

and has first order condition

$$\begin{aligned}\frac{(\sum_{\omega\in\Omega} \varphi_\omega q + A_\omega)\varphi_{s_1} - (\varphi_{s_1} q + A_{s_1})(\sum_{\omega\in\Omega} \varphi_\omega)}{(\sum_{\omega\in\Omega} \varphi_\omega q + A_\omega)^2} &= \frac{(\sum_{\omega\in\Omega} A_\omega)\varphi_{s_1} - A_{s_1}(\sum_{\omega\in\Omega} \varphi_\omega)}{(\sum_{\omega\in\Omega} \varphi_\omega q + A_\omega)^2} \\ &= \frac{(\sum_{\omega\in\Omega} A_\omega)\varphi_{s_1} - A_{s_1}(\sum_{\omega\in\Omega} \varphi_\omega)}{(\sum_{\omega\in\Omega} \varphi_\omega q + A_\omega)^2} > 0\end{aligned}$$

since $\sum_{\omega\in\Omega}(1 - F_\omega(s_1)) = 1$ and $\sum_{\omega\in\Omega}(2F_\omega(s_1) - 1) = 0$. Then, the first order condition is always greater than 0, so the corner solution $q = 1$ maximizes the utility under the s_1 signal and truth telling is the equilibrium action. □

Using these beliefs, the optimal x^* choice remains the same. Since player i of type Truthful will choose x_i to maximize their expected utility, we have

$$x_i^*(s_i, m_{-i}) = \begin{cases} A & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] > \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \\ [A, B] & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] = \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \\ B & \text{if } \mathbb{E}[u_i^T(x_i = A)|s_i, m_{-i}] < \mathbb{E}[u_i^T(x_i = B)|s_i, m_{-i}] \end{cases}$$

Second, the deceitful player must still expect over all possible message and signal states but must also additionally expect over the possible player types; in the case when $N = 3$, they do not know if the third additional player is truthful or deceitful. Furthermore, since the deceitful players know for sure their own type and the case in which all players are deceitful is irrelevant, we shrink the possible player type combinations only to those consisting of when the deceitful player is deceitful and at least one other player is truthful.

$$Q_D(p) = \sum_{\theta\in\Theta^{N-2}} P(\theta \in \Theta^{N-2}) \left(\sum_{\omega\in\Omega} P(\omega) \left(\prod_{m\in\Omega^{N-1}} P(m_i|\theta_i) \times E_{\omega m} \right) \right)$$

where

$$E_{\omega m} = \begin{cases} 1 & \text{if } \mathbb{E}[u_i^T(x_i = \bar{\omega})|s_i = \omega, m_{-i} = m] < \mathbb{E}[u_i^T(x_i = \omega_n)|s_i = \omega, m_{-i} = m] \\ 1/2 & \text{if } \mathbb{E}[u_i^T(x_i = \bar{\omega})|s_i = \omega, m_{-i} = m] = \mathbb{E}[u_i^T(x_i = \omega_n)|s_i = \omega, m_{-i} = m] \\ 0 & \text{if } \mathbb{E}[u_i^T(x_i = \bar{\omega})|s_i = \omega, m_{-i} = m] > \mathbb{E}[u_i^T(x_i = \omega_n)|s_i = \omega, m_{-i} = m] \end{cases}$$

with ω_n used to denote the non-true state. It is important to bear in mind $P(\omega)$ is known to the deceitful player since they are aware of the true state. Third, our formulation implies that we are only solving for the symmetric equilibrium, which must exist given this is a finite symmetric game. The game becomes more and more difficult to solve with increasing N since the truthful player's expected value function will be of order $N - 1$ with respect to p, the probability of the deceitful player lying. So, we turn to a numerical solution.

4 Numerical Results

4.1 Numerical Equilibrium

The parameters are set to $p_D = 0.3$ and $F_T = 0.60$. First, we can look at an example. We can plot the expected values of the deceitful player with respect to their action choice p in Figure 1. From Figure 1, we can see that this expect utility function is discontinuous

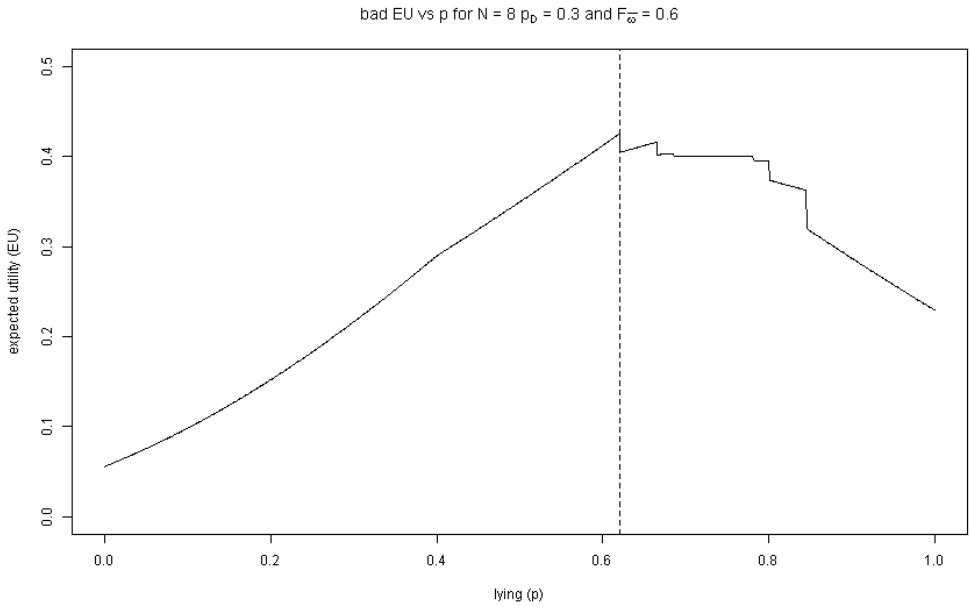

Figure 1: A plot of the deceitful player(s) expected utility when $N = 8$.

with respect to p. The discontinuities occur when p reaches a sufficiently high level where the representative truthful player will no longer choose the incorrect state in a particular signal-message realization. For example, if a truthfuly player sees 5 messages out of 7 going against their own signal. Without deception in the messages, they would choose to go with the messages. However, if deception is sufficiently high they would instead choose to disregard the messages and go with their own signal. As p increases, more of the choices in these signal-message realization flip. In our example illustrated in Figure 1, we have that in $N = 8$ the best strategy is to choose $p^* = 0.6208652$. Lying above $1 - F_{\bar{\omega}}(\bar{\omega})$ means that the deceptive players are indeed introducing more noise into the messages. In this discrete message distribution, the deceptive player will always shade a discontinuity be $\epsilon > 0$.

Figure 2 shows how the information aggregation (IA) efficiency degrades as the probability of a player is deceitful increases. This efficiency is calculated ex-ante to the signals and messages (similarly to how the deceptive players take the expectation of all possible signals and messages) and measures the probability of any truthful player deducing the correct state when the deceptive players are lying optimally.

$$\text{efficiency} = \sum_{\theta \in \Theta^{N-1}} P(\theta \in \Theta^{N-1}) \left(\sum_{\omega \in \Omega} F_{\bar{\omega}}(\omega) \left(\prod_{m \in \Omega^{N-1}} P(m_i|\theta_i) \times E(\omega, m) \right) \right)$$

where

$$E(s,m) = \mathbb{1}\left\{\mathbb{E}[\bar{u}_i^T(\bar{\omega})|s_i = \omega, m_{-i} = m] \geq \mathbb{E}[\bar{u}_i^T(\Omega/\bar{\omega})|s_i = \omega, m_{-i} = m]\right\}$$

Essentially, the efficiency tells us, ex-ante to any information exchange while being lied to optimally, how often a representative truthful player will be able to guess the true state. In the case worst case scenerio, the representative truthful player will only be able to rely on their own signal; thus, the efficiency will be bounded below by $F_{\bar{\omega}}(\bar{\omega})$. Furthermore, IA efficiency tends to increase when $p_D = 0$ and N increases since there will be more information in the system. As p_D increases, the IA efficiency will decreases towards its lower bound.

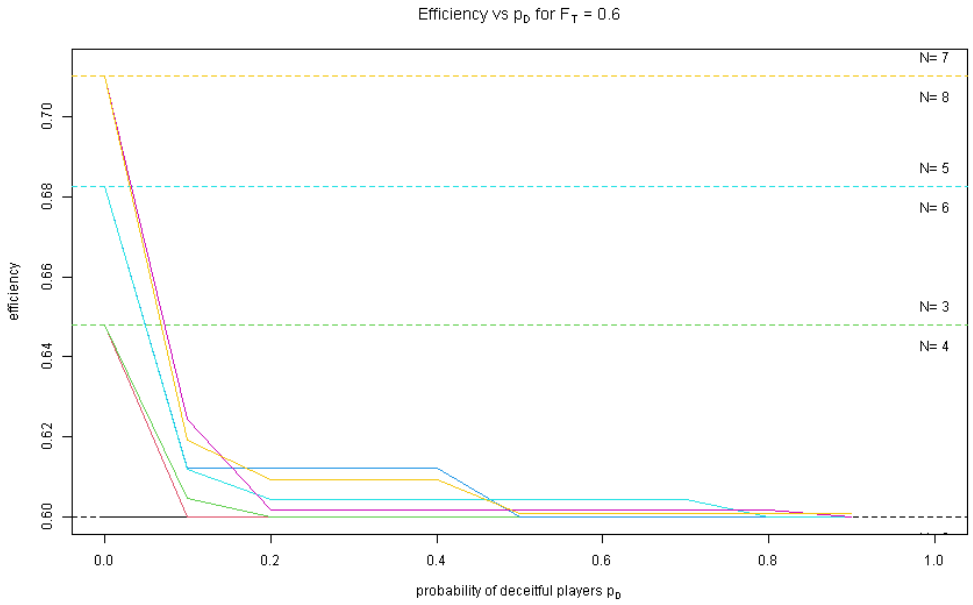

Figure 2: The information aggregation efficiency decreases as the presence of deceptive players increases.

While these expected utiltiy functions accurately reflect our setting, the discontinuities make evaluation, calculation, and interpretation more difficult. So, what if we wanted something that was smoother and more continuous.

4.2 Continuous Message Distribution

Let us characterize the messages recieved by a truthful player as how many messages A they will get: m_A. Currently, m_A follows a binomial distribution (with A as 1 and B as 0). This distribution only has support over $\{0, ...M\}$ which makes sense since you cannot have fractional messages. But this results in a discontinous expected utility function for deceitful players. So, we can replace the that message m distribution with one that is support over $[0, M]$ and we can mimic a continuous version of a binomial distribtution with a trucated normal approximation. That is a normal approximation only supported over $[0, M]$ with $\mu = Np$ and $\sigma = \sqrt{Npq}$. This approximation ends up being fairly good even at low N and while this is an approximation right now, this analysis can be generalized to any continuous distribution.

Now that fractional messages are allowed and m_A is continuous between 0 and M, the truthful player expected utility is smooth and continous with respect to m_A. In Figure 3, we can see the expected utilities of choosing $x = A$ and $x = B$ under the signal $s = A$. From the truthful player's perspective, they will always choose $x = B$ so long as $EU_T(x = B) > EU_T(x = A)$. In the discrete version, we can see from Figure 3, that they will choose $x = B$, until there are at least 4 messages are A (in this $N = 10$, $M = 9$ example). In the continuous version, we get a little more granularity. Let us define the value of m_A such that $EU_T(x = B) = EU_T(x = A)$ as c. Thus, every m_A to the left of c, the truthful player chooses $x = B$ and every m_A to the right of c, they choose $x = A$.

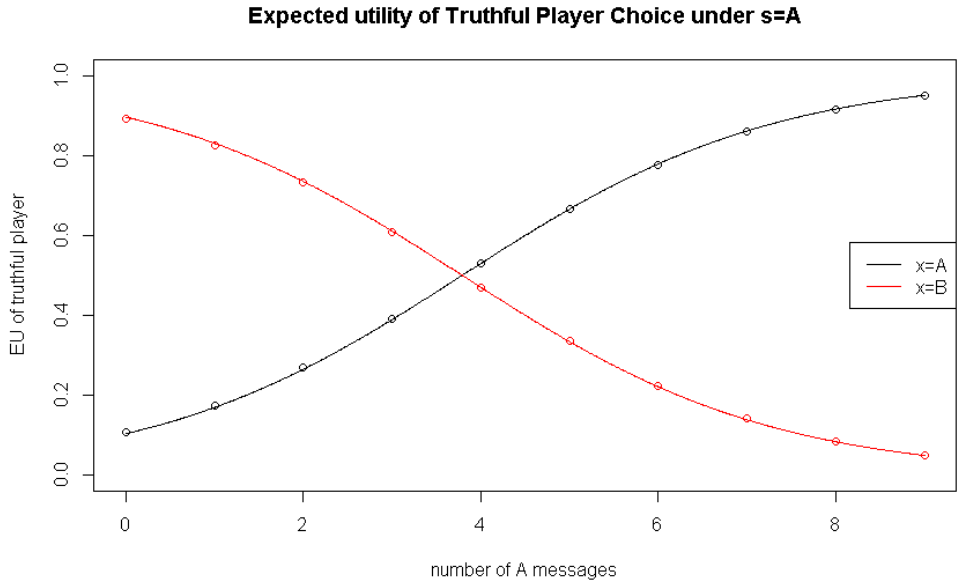

Figure 3: The plot of the expected value of a truthful player choosing $x = A$ (in black) and $x = B$ (in red) after already recieving a signal $s = A$ with repsect to increasing the number of A messages received. The black and red points are the expected utilties under the binomial distribution for m_A for $x = A$ and $x = B$, respectively.

Once, we have these "cutoff" points c where the truthful player will switch their choice,

we can use them to assemble the expected utility of the deceitful players. If you recall, the deceitful players maximize the probability that the truthful players choose $x = B$. If we can define the probability distribution of the A messages from the perspective of the deceitful player, then we can use the cutoff c to add up all the probabilities that will result in a choice $x = B$. A numerical calcuation of this continuous distribution compared to the discrete case can be seen in Figure 4.

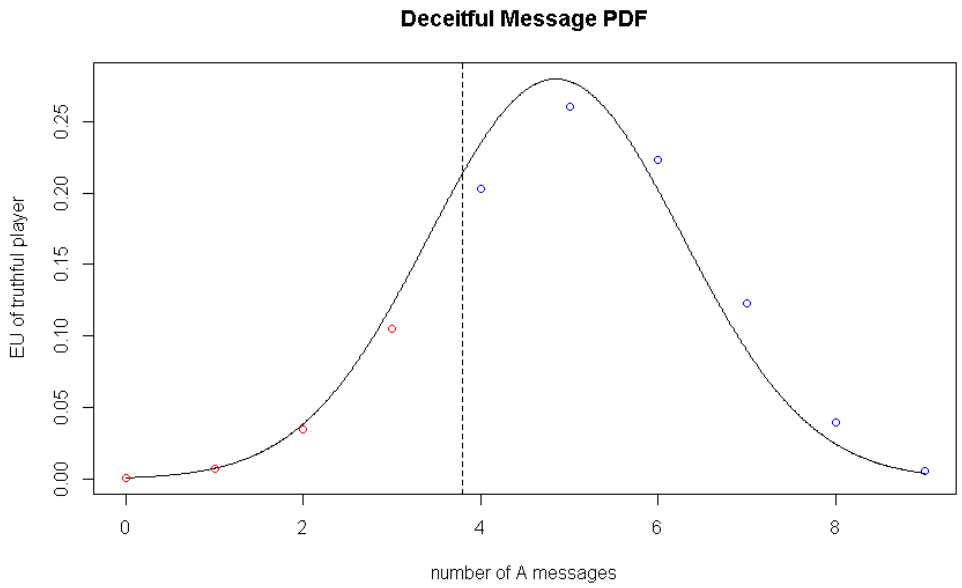

Figure 4: The plot of the probability density distribution of the messages seen by the truthful player from the deceitful player's persective. The points on the plot are the probabilities from the discrete case where the red and blue points are when the truthful player will choose B and A, respectively. The vertical dashed line is the cutoff c where $EU(A) = EU(B)$ in the continuous case.

In Figure 4, the red points and the blue points represent choosing B and A, respectively. The vertical dashed line is the cutoff c that shows where the truthful player will switch from $x = B$ to $x = A$. If this is the case, this vertical dashed line should always divide the red points from the blue points, which it does. Just as in the original model, we would sum over all the red points (all the points where the truthful player would choose $x = B$). We can integrate the PDF up c to capture the probabilties of all the m_A in which the truthful player chooses $x = B$. By assembling the expected utility of the deceitful player in this fashion, we can plot it against the discrete version in Figure 5.

Figure 5 shows us a very nicely smoothed version of the deceitful player's expected utilities without the discontinuities. However, the true benefits are when we observed the expected utility plot for several values of N which show that the reason that the optimal lying value jumps around for N is that it finds the closest discontinuity to the optimal lying value in the smoothed version (these plots are included at the end).

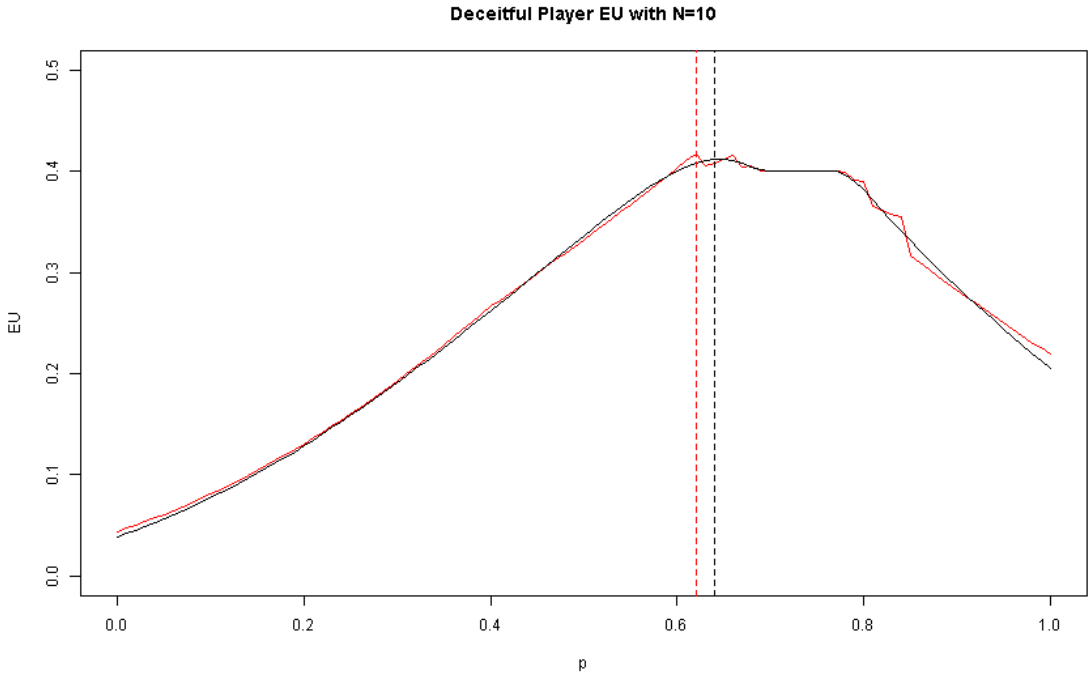

Figure 5: The plot of expected utility of the deceiful player over different levels of lying (p). The red and black lines describe the discrete and continuous versions, respectively, while the dashed lines are their optimal values.

And now, Figure 6, shows the relationship between optimal lying and N is now strictly monotonic non-decreasing in N for both the discrete and continuous cases. The optimal lying seems to jump around with a lot of noise but roughly following the very apparent increasing trend in optimal lying of the continuous model. This seemingly noisy pattern stems of the discontinuities in the expected utility induced by the discrete nature of the messages. The continuous case does not suffer from this issue.

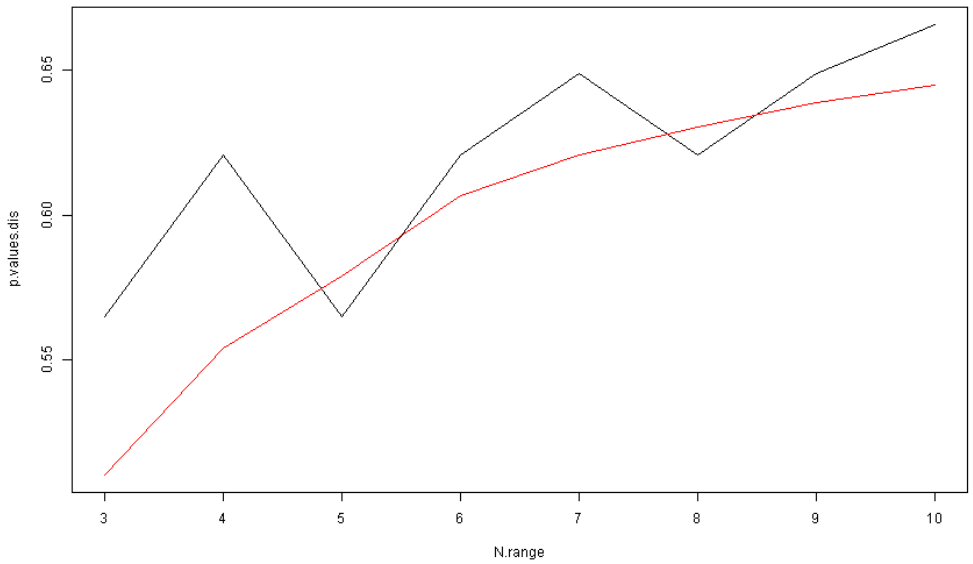

Figure 6: The plot of optimal lying (p) on the y-axis against the number of group members (N) on the x-axis with the red line as the model with continuous messages and the black line as the model with discrete messages.

4.3 Hypotheses

We form the following hypotheses based on our theoretical model:

1. Information aggregation efficiency decreases when deceptive agents are introduced (when $p_D = 0$ to $p_D = 0.3$), the percentage of truthful player who correctly choose the true state will decrease).

2. Information aggregation efficiency stays about the same when $p_D = 0.3$ but N increases from 4 to 8.

3. As the group size (N) increases, the optimal level lying for deceptive players increases (under the smoothed message probability).

4. As the group size increases from $N = 4$ to $N = 8$, the optimal level lying for deceptive players will stay the same (under discrete message probabilties).

The first two hypotheses are directly testing features of our model while the last two test differences between our continuous and discrete messaging models.

5 Experimental Investigation

5.1 Method

The experiment will be run at the Interdisciplinary Experimental Laboratory at Indiana University Bloomington (IU). Subects will be recruit via Online Recruitment System for

Economic Experiments (ORSEE) from a pool of opt-in college students enrolled at IU. Subjects will be taken through the instructions and take a comprehension quiz before being allowed to proceed to the experiment.

5.2 Design

The experiment consistents of several rounds with rematched groups between every round. Subjects will be assigned a player type (either truthful or deceitful) at the beginning of the experiment according the parameter p_D. Their player types do not change. The subjects participated in the following experiment:

In each decision round, the computer will randomly choose one of two bags with equal chance, an 'up' bag and a 'down' bag. Each bag contains 100 balls that are labelled either 'up' or 'down'. The 'up' has $F_T \times 100$ balls labelled 'up' and $(1 - F_T) \times 100$ balls labelled 'down'. The 'down' bag follows a similar pattern such that it is more likely to draw a ball matching the label of the bag.

After a bag has been selected, the deceitful players will be told which bag has been selected with the truthful players will only get to observe a single ball randomly drawn (with replacement) out of the bag. Then, each player will send one message (either 'up' or 'down') to all other players.

Truthful players will then observe all messages sent by their fellow players and choose which bag they believe to have been selected by the computer. Truthful players will receive a higher payoff the more truthful players that guess correctly (including themselves while deceitful players will recieve a higher payoff the less truthful players that guess correctly.

5.3 Treatments

To test our hypotheses, we have the following treatments holding $F_T = 0.60$ constant:

	$N = 4$	$N = 8$
$p_D = 0$	$N = 4, p_D = 0$	$N = 8, p_D = 0$
$p_D = 0.3$	$N = 4, p_D = 0.3$	$N = 8, p_D = 0.3$

6 Behavioral Model

6.1 Re-weighted Updating Process

As previous derived, the updating process for a Bayesian agent will have expected value as

$$\mathbb{E}[\bar{u}_i^T(x_i)|s_i, m_{-i}] = \frac{P(\bar{\omega} = x_i)F_{x_i}(s_i) \prod_{m' \in m_{-i}} [p_T F_{x_i}(m) + p_D p_{x_i}(m')]}{\sum_{\omega \in \Omega} P(\bar{\omega} = \omega)F_{\omega}(s_i) \prod_{m' \in m_{-i}} [p_T F_{\omega}(m_{-i}) + p_D p_{\omega}(m')]}$$

which can be augmented to

$$\mathbb{E}[\bar{u}_i^T(x_i)|s_i, m_{-i}] = \frac{P(\bar{\omega} = x_i)F_{x_i}(s_i)^{1+\alpha} \prod_{m' \in m_{-i}} [p_T F_{x_i}(m) + p_D p_{x_i}(m')]^{1+\beta}}{\sum_{\omega \in \Omega} P(\bar{\omega} = \omega)F_{\omega}(s_i)^{1+\alpha} \prod_{m' \in m_{-i}} [p_T F_{\omega}(m_{-i}) + p_D p_{\omega}(m')]^{1+\beta}}$$

By construction, α augments the weight of the signals while β augments the weight of the messages. The values of β have the following interpretations:

- $\alpha, \beta > 0$: information is weighted more than Bayesian benchmark
- $\alpha, \beta = 0$: equivalent to Bayesian benchmark
- $\alpha, \beta < 0$: information is weighted less than Bayesian benchmark

So, when β is positive, the subject would be over-weighting the messages and believing them more than they should. However, if β is negative they discount the messages too much and disregard the information that the messages could provide. α has a similar interpretation for the signal. We put a separate distortion parameters on the signal (private information) and the messages (public information) since there may be reaction to second-hand information. The identification of the two parameter comes when in the data the signals and messages go in opposing directions and how that ultimately influnces the choice made.

7 Conclusion

Still working on it.

References

[1] D. Acemoglu, A. Ozdaglar, and A. ParandehGheibi. Spread of (mis) information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010.

[2] C. Ba, J. A. Bohren, and A. Imas. Over-and underreaction to information. *Available at SSRN 4274617*, 2022.

[3] S. R. Beckman, J. P. Formby, W. J. Smith, and B. Zheng. Envy, malice and pareto efficiency: An experimental examination. *Social Choice and Welfare*, 19:349–367, 2002.

[4] K.-Y. Chen, L. R. Fine, and B. A. Huberman. Forecasting uncertain events with small groups. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, pages 58–64, 2001.

[5] K.-Y. Chen, L. R. Fine, and B. A. Huberman. Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science*, 50(7):983–994, 2004.

[6] C. Cornand and F. Heinemann. Measuring agents' reaction to private and public information in games with strategic complementarities. *Experimental Economics*, 17:61–77, 2014.

[7] V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451, 1982.

[8] P. Deb, D. Furceri, D. Jimenez, S. Kothari, J. D. Ostry, and N. Tawk. The effects of covid-19 vaccines on economic activity. *Swiss Journal of Economics and Statistics*, 158 (1):3, 2022.

[9] S. Fehrler and N. Hughes. How transparency kills information aggregation: theory and experiment. *American Economic Journal: Microeconomics*, 10(1):181–209, 2018.

[10] R. Forsythe and R. Lundholm. Information aggregation in an experimental market. *Econometrica: Journal of the Econometric Society*, pages 309–347, 1990.

[11] K. Kricorian, R. Civen, and O. Equils. Covid-19 vaccine hesitancy: misinformation and perceptions of vaccine safety. *Human vaccines & immunotherapeutics*, 18(1):1950504, 2022.

[12] C. R. Plott and K.-Y. Chen. Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem. 2002.

[13] C. R. Plott and S. Sunder. Efficiency of experimental security markets with insider information: An application of rational-expectations models. *Journal of political economy*, 90(4):663–698, 1982.

[14] C. R. Plott and S. Sunder. Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica: Journal of the Econometric Society*, pages 1085–1118, 1988.

[15] R. Wattenhofer, T. Moscibroda, and S. Schmid. The price of malice: A game-theoretic framework for malicious behavior in distributed systems. *Internet Mathematics*, 6(2), 2009.