DEPARTMENT OF COMPUTER SCIENCE

Counteracting against Bots within Social Media

Can Misinformation Intervention Create a Wise Society

Pai Chen

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Engineering in the Faculty of Engineering.

Tuesday 5th September, 2023

Abstract

Misinformation and polarization are two persistent issues that endanger social welfare by turning society unwise, we investigate how effective are two intervention strategies, social inoculation and account banning, in tackling the harmful impact caused by bots, which act as spreaders of fake information. We also evaluate how resistant the interventions are under different societal environments. We have found that (i) an optimal implementation of account banning (applied regularly and early) outperforms social inoculation. (ii) Social inoculation achieves a moderate degree of success in countering misinformation but does poorly against polarization, whereas account banning performs strongly for both metrics. (iii) Social inoculation performs better in a more misinformed society.

Dedication and Acknowledgements

I would like to thank my supervisor Dr Nirav Ajmeri for offering me guidance and help throughout the project. I would also like to thank Professor Seth Bullock for giving me inspiration for the topic of this research. Many thanks to Professor Marcos Ross Fernandes, for answering the questions I have regarding his paper on fake news and social media. And finally, I would like to thank my friends and family for all the support given to me to pull through this difficult time.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

Pai Chen, Tuesday 5th September, 2023

Contents

List of Figures

List of Tables

Ethics Statement

This project did not require ethical review, as determined by my supervisor, Nirav Ajmeri.

Chapter 1

Introduction

The term web 2.0 was first coined in 1999, and with it, internet users have promoted their roles from passive content consumers to active content producers [21]. However, with the minimal effort of creating an account on any social media, users can spread their views to a wider audience without any responsibilities attached. Within a surveyed population of 2210 U.S. adults, 37% of whom update their understanding of the world through social media [12]. As more and more people turn to social media these days for the acquisition of news, the circulation of rumours, fake news and urban legends have become prevalent, leading to an *unwise* society.

We decide whether or not a society is wise explicitly on two metrics: (i): *misinformation* and (ii) *polarization*. This is based on the proposition made by [4], which states that an undivided society (information can flow freely) without any deliberate spreaders of false beliefs will eventually reach a wise state in which polarization and misinformation levels converge to zero. Misinformation generally refers to fabricated or misperceived facts which are then spread with or without deliberate intention [48]. Whereas polarization measures how much the society is divided. In contrast, a highly polarized society holding beliefs drastically different from each other can exacerbate misinformation even further [17], A Phenomenon which is referred to as affective polarization [45] indicates that different members of political parties with conflicting interests or ideologies may oppose each other in a hostile way merely due to their difference in stances. In the context of learning the truth, it adds serious friction to the spread of truth or correction of misbeliefs.

However, this proposition lies heavily on the assumption of the no presence of such "spreaders of false beliefs" within society. Although it is natural for people to exchange ideas and opinions with acquaintances throughout their every day lives, subsequently forming and updating their views which may or may not be correct, we instead divert our attention to the "unnatural" part, focusing on the deliberate manipulation of opinions stemming from all sorts of exploits and is intentions. At our centre of attention, we focus on the specific subject, "bot", referring to the type of entity existing within social media, which serves the particular purpose of truth distortion for private gains, hence causing society to become unwise. Individuals who are biased or inattentive could be misled into believing such information. Furthermore, even if some people find ways to shield themselves from misinformation using rationale and critical thinking, there is no guarantee that their families and friends do not become advocates of misbeliefs, unknowingly acting as spreaders of fake news [4]. Such "network externality effect [4]" leaves even the most rational people vulnerable to misinformation, as it is extremely difficult to trace the source of a piece of information shared by members within their social circles.

Real-life interventions deployed by social media companies to address false information comes either in the form of misinformation correction [28, 5], relying on explicit labelling and removal of inappropriate content by human experts and computer algorithms (commonly referred to as *debunking*). Or mitigating strategies like social campaigns aimed at people against any future exposures of misinformation [35, 36, 28] (commonly referred to as *prebunking*). However, their efficacies can be limited at times, as something like fact-checking social media posts can be seen as policing over what is right or wrong for certain users and provoke their psychological resistance.

At the centre of our research, there lie two questions:

RQ1 *How effective are misinformation interventions against bots on social media?*

RQ2 *Do differences in bot-related properties affect interventions' effectiveness?*

Most intervention strategy evaluations are grounded in real life through the methods of user surveys[43].

Our motivation for RQ1 hence is an alternative, computer-simulated approach, which not only allows for more freedom of control over the properties of the population, such as size or composition but, more importantly, the ability to predict evolutions of the population through time, which can not be done in a real-world simulation.

To fulfil the requirements of RQ1, we must find a way to capture and simulate the behaviour of a complex system like social media and the various interactions between different entities. This inspired the adoption of an *agent-based model* (ABM) in tackling this problem, which provides us with a heuristic way of modelling such a complex system through decomposition [6]. We then design and implement two specific intervention methods and assess their abilities to counteract the negative impacts of bots, making society converge to a truthful state in the long run.

The motivation for RQ2 comes from the varying efficacies of the intervention approaches. The exact explanations for this vary and are often explained by the psychological resistance of human nature, but rarely has this been analysed from the standpoint of behaviours on a societal level, which is to say how specific types of entities and their behaviours could potentially hinder the intended purpose of interventions. We focus our attention on bots, specifically on how variations in bot properties affect efficacies. The specific variations here refer to three behavioural properties of the bot agents: (i) How fast do the bots become extreme overtime (A_b), (ii) how many bots exist within the ABM (N_b), (iii) How many active followers/believers of bots exist within the ABM (P_{bf}).

The rest of the project is structured in the following way: Chapter 2 explains background terminologies and metrics relevant to our research and previous attempts. Chapter 3 explains the components of the social network model, describing each agent type and their interaction behaviours, in addition to the algorithmic approach we used to synthesise our social network structure. In Chapter 4 we give implementation details of the two intervention methods proposed. In Chapter 5, we first validate our model against the original model, which we replicate from. conduct a series of controlled experiments investigating the causalities between different properties of bots and their impact on the effectiveness of the intervention. Chapter 6 provides reflections in terms of possible improvements as well as directions for future works, whereas Chapter 7 summarizes our findings.

Chapter 2

Background

2.1 Metrics for Determining a Wise Society

2.1.1 Misinformation

Formally, the term *misinformation* is used as "an umbrella term to include all false or inaccurate information that is spread on social media" [48], this includes a list of sub-terms such as "spam" or "conspiracy theory" which all share similar meanings to circulating information which is away from the truth.

The presence of widespread misinformation on social media comes from the platform's unrestrained nature [48]. The premise of social media is built on the idea of ordinary users being able to share anything and anytime with few restrictions applied, making it the perfect breeding ground where misinformation can hatch.

Under most circumstances, an abundance of misinformation results in diminished efficiency and loss of welfare for the entire society. For example, degradation in the quality of healthcare caused by the dissemination of false beliefs about possible side effects of vaccination uptakes [47], making the population vulnerable to epidemic diseases. Another possible impact is related to societal unrest, this can be exemplified by the PizzaGate conspiracy theory. The theory suggested links between human trafficking and a pizzeria located in Washington, D.C., which was then widely spread online forums [27] during the 2016 US presidential campaign, and eventually led to a gun-firing incident inside the restaurant [19].

2.1.2 Polarization

Similarly, polarization can be defined as when people with similar traits (opinions etc.) are grouped together into the same cluster and between different clusters, there often lies substantial differences [10]. These differences would often generate tensions between groups and lead to civil unrest [10].

The reasons for such phenomenons of inter-group divergence and intra-group-convergence of opinions to occur are manyfold. For instance, the former phenomenon can be a cause of the assimilation effect of Social Judgement Theory, where if people receive information that falls within their latitude of acceptance, their respective opinions would be drawn closer together [15]. Whereas the latter can be explained through a manifestation of polarization, affective polarisation. This type of polarisation describes the animosity against the opposing political party members, causes societal segregation and an increased level of difficulty in coordinating the whole population in achieving objectives of collective benefits [45], such as raising the national minimum wage for workers. Hence the aggression induced by such polarisation drives people further apart.

Link recommendation algorithms used by social media with the intention of connecting users with people they may know, inadvertently exacerbates polarization even further [38] through the creation of so called "echo chambers", described as environments that reinforce beliefs through limiting people's exposure to only those who hold similar opinions to each other [8].

2.2 Interventions

In the context of this particular study, we confine the word "intervention" to specifically refer to methods that specifically target misinformation rather than polarisation on social media. This is mostly for the practical purposes of our experiment implementation, misinformation intervention targeting social media

is a well-established field of study where plenty of previous work exists [5, 28, 35, 36, 39, 14], and relevant methods, such as fact-checkers, are easy to implement algorithmically. On the contrary, although the research into polarisation interventions (depolarisation) are present, they are often confined within the specific case studies, and lack the general applicability of social media platforms. For instance, one study [2] exemplifies that depolarisation could be achieved for members within one party by telling the commonalities they share with the people from an opposing party. Such measures, in theory could be constructed for the simulation of social dynamics, but the complications involved in designing it goes against the KISS (keep it simple, stupid) principle proposed by Robert Axelrod [3] when designing systems of simulations. Instead of conducting user surveys with human participants, our experiments rely entirely upon designing high level interaction rules for the simulation process, which we then apply inductive analysis of the results produced.

Most existing social-media-related intervention strategies can be generalized into two categories: *debunking* and *prebunking* [28]. Where debunking takes the active approach by attempting to rectify already-existing misinformation in the system through means such as (i) providing expert/user ratings of information source or content [22]; (ii) removal of inappropriate information; (iii) banning of misinformation-spreading accounts [5].

On the other hand, Prebunking refers to the idea of preemptive education of the general public to recognise misinformation, making them more resistant to later exposures [36], which is also known as *social inoculation*. Based on inoculation theory [25], social inoculation originates from the field of medicine, where just like medical inoculation would immunize people against incoming diseases, social inoculation equivalently would help people to build up resistance against misinformation [36].

However, there has been a fair share of criticism of the effectiveness of both categories of interventions. Some can view debunking methods such as expert labelling as the platform policing over what is right or wrong [37], which may trigger some users' adversary emotions and deem the labelling itself as untrustworthy. Providing warnings for news headlines can produce an implied truth effect [32, 13], where people automatically assume unflagged headlines to be reliable. Removal of posts is suggested to be not practical in containing misinformation due to the sheer amount of information the platform needs to process [5]. Account suspension is proven to be more effective in comparison [5], at the cost of depriving users of communication channels, which is particularly detrimental to the elderly and disabled [29]. Whereas prebunking methods can be hindered by the confirmation bias [28], referring to the people's tendency to maintain their current views or opinions and oppose anything that contradicts them [23].

The above-mentioned work acts as our main source of inspiration when designing alternative form of implementation of the intervention methods while still keeping the approach succinct enough to be strictly within the scope of our study rather than trying to model every single detail [40].

2.3 The Applications of ABM and Previous Attempts

The three most prominent features of an ABM are (i) *agents* possessing different behaviour properties, (ii) pre-determined rules defining the possible interactions between agents (iii) and the environment they reside in [6]. The most prevalent use of the approach is in modelling human behaviours, although previous research has also been conducted with subjects such as animals [26] and infectious diseases [33]. In the context of this research, we focus primarily on humans, where ABM's usage lies mostly in the production of systems of economics and social sciences disciplines [7].

As has been stated [7], "The crucial idea that is at the heart of these approaches is to use computing as an aid to the development of theories of human behaviour", meaning that the primary goal of ABM is to provide explanations for the phenomenons observed as the model evolves, instead of how it will evolve [6]. This heuristic approach under most circumstances provides a close enough alternative to the optimal solution to the problem, especially when the problem itself is related to the modelling of a large-scale, dynamic system such as social media [6].

We must also acknowledge a few of ABM's limitations, "Whereas the purpose of induction is to find patterns in data and that of the deduction is to find consequences of assumptions, the purpose of agent-based modelling starts with assumptions' " [7]. We can not use it to prove theorems like a deduction would, but only for inductive analysis with simulated data [6]. Another downside of an ABM approach is the computational cost due to constant communications between agents [6].

The specific usages of ABM in the context of modelling misinformation intervention is fairly limited [14]. Among those limited few [39] creates the ABM simulating the process of rumour creation and sharing, and evaluating the use of two rumour control strategies, along with the possible variations. The first one limits the ability of influential agents to spread rumours by dropping their messages, whereas

the second takes a more gentle approach through the dissemination of anti-rumour messages. One shortcoming of [39], as have been pointed out by [14] is the absence of validation through the use of real-world data. On the other hand, [14] proposes a different ABM also tries to capture the concept of rumour spreading and looks into three intervention strategies: (i) rule-based policies, (ii) social inoculation and (iii) accuracy flags, evaluates their abilities in containing the spread of rumour against the case in which no intervention is applied, as well as against the authentic user data gathered from Twitter.

Chapter 3

ABM Implementation for Social Media

An accurate yet succinct high-level definition of the subject under research is key to acquiring accurate results. Our research revolves around the topic of opinion exchanges within the social network and the study of possible intervention implementation and evaluation, hence the need to establish a base model of social simulation. The paper by Azzimonti and Fernandes [4] proposes the idea of using the Monte-Carlo Simulation to replicate the process of opinion evolution of people within a social media network over a set amount of iteration periods. Our model builds on the structures of the Monte-Carlo Simulation in an ABM way (We refer to the original model as M_{mcs} and our replicated model as M_{abm} henceforth). The M_{abm} has also provided the following extensions from the original: (i) The graph that is used to represent a network of social media connections has been algorithmically synthesised rather than selectively chosen from a part of the real-world dataset as in M_{mcs}. (ii) We have designed methods of interventions for misinformation and polarization with regard to the model. These are also implemented as rules of interactions that modify certain agent groups' behaviour.

3.1 Components of Social Network

Following the definition from [6], at the heart of our model implementation, there lie three essential components: (i) The social network *environment* that the agents (users) exchange information in, as well as other higher-level concepts like the *truth*. (ii) three different types of *agents*, of which one is named intuitively as *bot*, which we identify as the main source of misinformation spread within social media. As a result of the bot agent being the centre of focus, the other two types of agents, *regular agents* and *bot followers*, have their properties solely revolving around the bot agent (we will explain this in detail below). (ii) predefined *communication rules* defining the process of opinion exchange and update in each iteration.

3.1.1 Environment

The exact definition of the social network environment is two-fold: Conceptually, the environment activates each agent randomly in each iteration, and the activation process is strictly sequential in the sense that in the update process of one agent, no other agent will be updated. Physically, a social network refers to the construct of a directed graph, $G(V, E)$. To our intuitive understanding, G represents a "web" of social relationships between all network users (V), where such a relationship ($e \in E$) can be formed through something like the "follow" feature of Twitter. Each e' directed nature represents the unilateral influence on opinion from one agent to another. For example, an edge $e_{i,j}$ from agent i to j would represent i, paying attention to j. This is a design decision made by [4] to capture the most common way of information acquisition on Twitter of reading a feed from someone a user follows, which, unlike offline face-to-face communication, usually is considered to be a one-way process the way and for the most part asynchronous, which does not require the presence of the other party (the author or sharer of the post) to complete the information exchange process. This allows better information coverage, but at the cost of dissemination of misinformation to a wider audience.

Truth

Truth acts as the single source of information which spreads to an unbiased view on a subject matter, θ ($\theta \in \Theta = [0, 1]$ [4]) to all other agents in work. Alternatively, this can be perceived as a single node to which every other node in the graph directly connects. One way of interpreting the value of θ is the extent to which an approach needs to be for the optimal outcome, where 0 would mean do nothing and 1 would mean an all-out approach. We define $\theta = 0.5$ in the context of this research (the optimal approach needs to be moderate).

The process of acquiring the truth is represented by a single draw $s_{i,t}$ by agent i at time t from a Bernoulli distribution represented as follows:

$$s_{i,t} \sim Bernoulli(P_{truth} = 0.5) \tag{3.1}$$

$s_{i,t}$ is equivalent to a single draw by the agent from a Bernoulli distribution of which the probability of learning the truth, $P_{truth} = 0.5$, which can be perceived as the amount of attention paid to an authentic, truth-conforming authority or news source. We can deduce that the number of people receiving at iteration period t is $N_{truth} \approx N \cdot P_{truth}$, meaning almost half of the population learns the truth each time.

3.1.2 Agents

Agents are social media users who hold an *opinion* on a certain topic of discussion worldwide. However, the range of this topic should exclude a person's subjective view, such as affection or antagonism, and should instead be confinable into a linear scale of $op = [0, 1]$. More formally, this can be defined as the *optimal* way to conduct or intervene in a matter of concern, such as how ethical it is for social media platforms to moderate user content without their consent. In a more political context, the value can be considered to characterize an individual's political spectrum, where a value relatively closer to 0 is viewed as radical (left-leaning) and a value closer to 1 is viewed as conservative (right-leaning) [11].

The specific value of the opinion is modelled by the beta distribution of the form: $Beta(\alpha_{i,t}, \beta_{i,t})$, where $\alpha_{i,t}$ and $\beta_{i,t}$ decide the skewness of the distribution. A higher α would make the distribution skew to the right, and a higher β would skew it to the left. We calculate $op_{i,t}$, the opinion of agent i, at time t, by calculating the expected value of the given beta distribution[4]:

$$op_{i,t} = \frac{a_{i,t}}{a_{i,t} + b_{i,t}} \tag{3.2}$$

All agents share the common ability to update their opinions via receiving information from different sources. Agents also differ from each other (i) in their relative locations in the graph and (ii) in the different sources of information they are exposed to. This difference is a cause of the random initialisation process at the start of each simulation (which we will discuss later), where each node $v \in V$ within network G is assigned corresponding agent identities, all three of which will be explained below.

Bots

Two out of three agent types rely on a clear conceptualisation of the bot agent, hence the need to address it first. The idea of such a deliberate spreader of misinformation is, in fact, derived by [4] from an even earlier concept, referred to as forceful (stubborn) agent [1]. Both of these share the common property of completely ignoring or resisting outside influences from other agents. Similar to sinks that have no outgoing edges, the set of agents that a bot agent i pays attention to (its out-neighbourhood) is *always* represented as:

$$NB_i^{out} = \{\} \tag{3.3}$$

Their difference lies in the fact that the bot agent plays a more aggressive role by additionally converging its own opinions to the extreme as time passes to affect others ($NB_i^{in} \neq \{\}$, some agents still pay attention to bots). This update process of bot agent a_i at time t is defined mathematically as:

$$a_{i,t+1} = a_{i,t} + A_b \qquad \text{(R-bot only)} \tag{3.4}$$

$$b_{i,t+1} = b_{i,t} + A_b \qquad \text{(L-bot only)} \tag{3.5}$$

Note the difference between a left-leaning bot (L-bot) and a right-leaning bot (R-bot) in this case, they represent a sub-category of the bot agent type and intuitively can be understood as promoters of left-wing or right-wing political views. A_b can be interpreted as the aggressiveness level of bots, otherwise referred to as flooding capacity [4]. The scale of A_b determines how fast a bot's opinion gets extreme over time (converging to 0 or 1). This also implies the fact that bot agents start out by disguising themselves as innocent people, with an arbitrary value of opinion assigned to it as the other agents do during the initialisation process.

Intuitively, bots serve as the main sources of misinformation spreading "whose roles are to exert disproportional influence" [4] over other agents in an attempt to manipulate public opinions to their own favour, similar to the sanctioning process defined in a socio-technical system [30] (STS henceforth). Without it, society converges to a truthful state with the help of a trustworthy news source, provided the network is strongly connected [4].

Bot followers

As mentioned earlier, the exposure to information for agents of different types is different. Agents under the category do not have the capability to distinguish bots from others and thus be affected by misinformation often at the very early stages of the dissemination process. The set of agents a bot follower agent i pays attention to is thus:

$$NB_i^{out} = \{i|(i,j) \in E\} \tag{3.6}$$

The NB_i^{out} this instance must include at least one bot agent, and just like bot agents can be further categorized into L-bot and R-bot, bot followers can equivalently be split into L-bot followers and R-bot followers. The information they receive, as a result, is often skewed away from the truth towards the left or right, making misinformation diffuse out into more parts of society, unknowingly making them become spreaders of misinformation.

the update process for bot followers is two-fold: (i) incorporation of opinions from out-neighbours (ii) incorporation of opinion from an unbiased source. Mathematically, this is defined as:

$$a_{i,t+1} = \begin{cases} (1-w_{i,t})(a_{i,t}+s_{i,t}) + w_{i,t}\sum_{j \in NB_{i,t}^{out}} \frac{a_{j,t}}{|NB_{i,t}^{out}|} & \text{, if } |NB_{i,t}^{out}| > 0 \\ a_{i,t} + s_{i,t} & \text{, otherwise} \end{cases} \tag{3.7}$$

Equivalently, the update rule for β shares the same procedure. There are a few notable things from Formula 3.7: firstly, the incorporation of opinions first is commenced on the out neighbours of agent i ($NB_{i,t}^{out}$), subsequently the unbiased source. Secondly, $w_{i,t}$ (assumed to be 0.5 by [4]) represents the amount of attention paid to his out-neighbours (friends and families), with an equal amount of weighting attributed to each one of them. The exception is the case in which $|NB_{i,t}^{out}| = 0$, for which $w_{i,t} = 0$ and all agent attention is given to the unbiased source (as shown by the otherwise case of the formula).

Regular agents

These are rational individuals with the ability to differentiate between a bot and authentic users. This shields them from exposure to misinformation, but only partially. One key assumption made by [4] is that no people can back out the source of information. This is to say that even though regular agents can tell bots and humans apart, they are still vulnerable to misinformation due to the fact that they could be affected otherwise, possibly through neighbouring bot follower agents or even other regular agents surrounded by misinformed agents who become skewed over time.

Regular agents and bot followers share the same update process, although there exists one difference between them, that is the agents they pay attention to at time t:

$$NB_i^{out} = \{i|(i,j) \in E, j \neq \text{bot}\} \tag{3.8}$$

Essentially any regular agent filters out any possible agent connections it might have during each update process.

3.1.3 Simulation Process

At the start of each simulation process, the initial opinions of each agent are evenly spread out across $K = 7$ groups over the linear scale of $[0, 1]$. We assume there exist seven different groups of people existing

within the society/social media, where each group holds a different kind of belief kinds of ideologies/views within the simulated society, of which each view holds an equal proportion of people. The α and β component of each opinion is calculated through the following formula[1]:

$$a = -\frac{\mu(\sigma^2 + \mu^2 - \mu)}{\sigma^2} \tag{3.9}$$

$$b = \frac{(\sigma^2 + \mu^2 - \mu)(\mu - 1)}{\sigma^2} \tag{3.10}$$

We randomly assign 15% of the population, that is, all nodes in the set of $\{v \in V | G(V, E)\}$ to be bot followers of each type ($P_{bf} = 0.15$), which is then further divided into two random parts, one part set to follow the L-Bot and the other follow the R-bot. The rest (85%) of the population is then assigned to be regular agents. Finally, we add two extra nodes as bot agents (one L-bot and one R-bot) and form edge connections between the corresponding bot followers and bots.

We set the total simulation periods to be $T = 1000$ for each experiment we are going to conduct. By observation of some preliminary results produced from running the simulation model, we can see that metrics like misinformation or polarization tend to stabilize at this point. Hence, there is no need to run the experiment further. At each simulation period t ($1 <= t <= 1000$), all agents within the system (regular agents, bot followers and bots) will be activated once sequentially in an arbitrary order. Once an agent is activated, it will conduct its update process conforming to a pre-defined set of behaviours, summarized as (i) receiving opinions from the unbias source, as well as (ii) receiving opinions from NB_i^{out}, its out-neighbours, which may vary depending on agent types.

One notable aspect of [4]'s implementation is the stochasticity involved in information acquisition. The stochastic disabling and enabling of edge connections come from choices made by previous implementations [1, 4]. The exact logic behind this is stated as the modelling of natural attention shifts that occur in any human being by its inability to conceive every bit of information it is exposed to. The implementations of the update process rely on a simple process, which is either a draw from the Poisson distribution [1], or a draw from the Bernoulli distribution [4]. For our implementation, we adopt the latter approach where the chance of an agent communication at time t, is modelled as i.i.d variables drawn from the Bernoulli distribution in the form of:

$$op_t \sim \text{Bernoulli}(P_{comm} = 0.5) \tag{3.11}$$

The value of P_{comm} is a subjective choice decided by [4], the higher the P_{comm}, the higher the rate of communication.

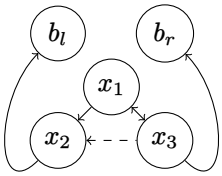

Figure 3.1: A high-level, minimal representation of the graph structure as well as agent relationships, where x_1 represents a regular agent, x_2, x_3 represents bot followers and b_l, l_r represent bots. The dash edges from x_2 to x_3 represent the stochastic nature of the communication process.

3.2 Network Synthesization

Graph G, which is the representation for modelling social network relationships is synthesised algorithmically rather than making use of existing datasets like the ego-Twitter dataset. Real-life social media networks often have a large number of nodes and edges (The ego-Twitter network from the SNAP dataset has 81,306 nodes and 1,768,149 edges). In comparison, a network (graph) generation algorithm gives more control to users over the properties of the generated network. For instance, the size of our generated

[1]The equations are themselves derived from $\mu = \frac{a_{i,t}}{a_{i,t}+b_{i,t}}$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, which are mean and variance measures of any beta distribution accordingly ($\sigma^2 = 0.03$ as defined in [4])

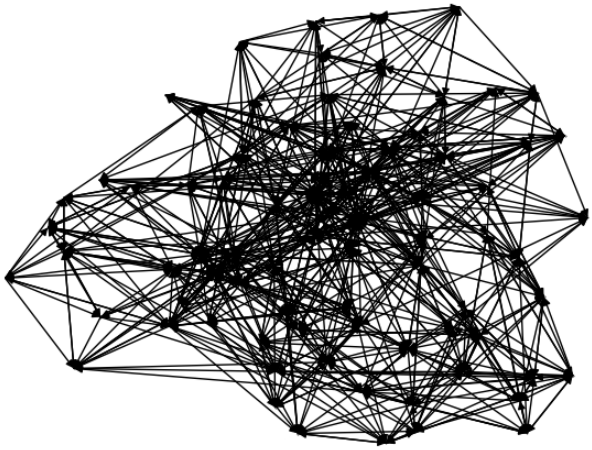

Figure 3.2: A network generated through the "ego-centric" algorithm. The network shown has 80 nodes and 767 edges.

network is only a fraction of its size as shown in table 3.1. As a result of our network's compact representation, in particular its short average path length, opinions can be rapidly spread throughout the network in a short amount of simulation periods. In the case of misinformation spreading, the rate of dissemination is significantly faster than in a real-life situation, where an argument can be made that if any misinformation intervention is proven to be effective, it most likely will work in real life [40]. Additionally, a smaller graph also saves research time spent running each simulation, as the long-term results generally converge faster.

	Ego-Twitter	Our generated network
Average clustering coefficient (Transitivity)	0.57	0.55
Reciprocity	0.48	0.46
Average in-degree	21.75	14.22
Number of nodes	81306	1067
Number of edges	1768149	15178
Average degree	43.49	28.45
Diameter	15	5
Average path length	4.91	3.40

Table 3.1: A comparison between the properties of the ego-Twtter network our algorithmically generate network, all metrics are rounded to 2 decimal places.

In order to model a realistic social network whilst keeping run time reasonable, we decide the follow the same key constraints used in [4], which are average clustering coefficient (AVC), average in-degree, and reciprocity in our synthesis process. There exist several different measures of AVC, including local clustering coefficient and global clustering coefficient [31]. We choose to use the (undirected) global clustering coefficient also known as transitivity in this case, which is defined as the number of closed triplets overall open and triplets within a graph [31]. We see that the measure of transitivity produces exactly the same result on AVC as the one shown in ego-Twitter's dataset statistics (0.5653).

The metric of transitivity based on the proportion of closed triplets can be seen as the essential unit of measurement for a real-world social network, as in real life, there is a high chance for close contacts of a person to form connections as well [46]. Several graph generation or modification methods have been proposed that provide a tunable transitivity parameter [46]. One such method we experimented with is based on rewiring edges between a set of randomly selected nodes for an existing graph until a desirable level of transitivity is reached [16]. However, this is too slow in practice and can not apply to a large network with thousands of nodes. As a result, we have chosen an alternative more suitable for large networks [41], where an algorithm of large-scale network generation is implemented using an

"ego-centric" approach. The phrase "ego" here refers to the creation of so-called "ego networks", which are networks that resembles a social media user's social circle, formed by a central node (ego) and all nodes that are directly connected to such node (alters), as well as the edges that exist between them[24]. The procedure proposed by [41] starts off with a single ego network based on an Erdos-Renyi graph of the desired transitivity. Then the network is expanded by randomly selecting a different node as ego, and then forming another similar ego network to the initial one by generating the required amount of nodes and edges. Throughout the generation process, the amount of nodes generated diminishes, and it eventually reaches a stage where only edges are formed between existing nodes. Through testing, we are able to easily generate large-scale network structures in a relatively short amount of time.

The aforementioned ego-centric approach only provides a method for synthesizing undirected networks, this does not conform with the requirements proposed by [4] for its model, where people's attention is modelled as directed edges. We dealt with this naively by randomly removing edges from bidirectional connections of the graph until the reciprocity reaches a suitable amount.

A comparison between the network statistics of both ego-Twitter and our generated network is shown in Table 3.1, where the AVCs and reciprocities of the two networks are extremely close, although the in-degree for our generate network is lower than ego-Twitter, it is still reasonably accurate due to the generated network's much smaller scale.

Chapter 4

Intervention Design

When designing agents for the ABM, even small adjustments at a micro level can have a notable impact and produce intriguing results at the macro level [7]. This ideology of design applies equivalently in the context of intervention design, where we follow a simplistic design process by abstracting away the technical details behind the methods of intervention, only preserving the key notions and then transforming them into the applicable mechanism for our ABM. We have chosen the method of account banning and social inoculation as the primary focus for their effectiveness in reducing misinformation.

4.1 Social Inoculation

Social inoculation is applied empirically through the provision of weakened versions of fake information to people in a preventative manner, in an attempt to help trigger their protective responses when actual misinformation arrives [9]. We abstract this ideology away by mainly focusing on the intended outcome of the strategy, which is training people to be more resistant to misinformation. In other words, they become more capable of differentiating between fake and real news. We draw inspiration from the Social Judgment Theory [15], specifically the term defined as the *latitude of acceptance*, which represents the range of beliefs that a person considers to be reasonable. In the context of our implementation, the set of agents that inoculated agent i's receive information from at t becomes:

$$NB_{i,t}^{out} = \{j | j \in NB_{i,t}^{out}, op_{j,t} \in R_{si}\} \tag{4.1}$$

The percentage of agents who are inoculated ($P_{si} = 20\%$ for the base case) can recognise other agents who hold beliefs that are distant from the truth, in this case, outside of the inoculation range, and ignore them when updating their beliefs[1]. The wider the range of R_{si}, the more tolerant we allow the inoculated population to be in the sense that they are more likely to receive and incorporate beliefs that are away from the truth value.

4.2 Account Banning

Account banning is generally viewed as the go-to method to contain the spread of misinformation, as has been proven to be more effective when compared to a series of other methods such as nudging or content removal [5]. We implement account banning by monitoring at each simulation period, the level of opinion possessed by each agent, specifically those who become extreme. The definition of extreme in this context simply means opinions far away from the truth ($\theta = 0.5$). Once an extreme agent emerges, it is immediately shut down and put into a disabled state, in which it does not update its own opinion. The pseudocode implementation for the intervention is described by 4.1. Additionally, these disabled agents are also ignored throughout the update process by other agents, specifically those that are neighbours to them. Mathematically, the set of agents an active agent i, update their opinions from at simulation period t becomes:

$$NB_{i,t}^{out} = \{j | j \in NB_{i,t}^{out}, op_{j,t} \in R_{si}\} \tag{4.2}$$

[1]This implementation, at first glance, can be viewed as quite similar to the one of a regular agent. However, they are fundamentally different as regular agents are only able to recognise and reject agents that are pre-assigned to be bots, our definition of inoculated individuals reject inaccurate opinions spread by any agent type, whether it is a bot or not

Algorithm 4.1: At the start of the simulation, each agent is assigned an initial value of T_{ban}, representing the number of iterations an agent can be disabled for. At simulation period t, before each agent updates its opinion, its current opinion, $op_{i,t}$ is checked against R_{ab} to see if it falls within or not (In other words, R_{ab} determines the extent to which our imaginary social media moderator deems an agent as extreme). If it does, then the agent updates its opinion normally. If it does not, an active agent would be disabled and its opinion would not be updated. Whereas an already disabled agent would be checked for its current value of c, if it has not reached zero, we reduce it by one. Otherwise, if it reaches zero, then it would be "woken up", returning back to active, assigned a new value within an acceptable range and resetting every other relevant setting back to normal.

```
if op_{i,t} ∉ R_{ab} then
    if Agent_i = banned then
        if c ≠ 0 then
            c = T_{ban}
            Agent_i = notBanned
            op_{i,t+1} ← draw(R_{ab})
        else
            c ← c − 1
    else
        Agent_i = banned
        c ← c − 1
else
    op_{i,t+1} = update(op_{i,t})
```

Our implementation of account banning shares a fair degree of similarity to the rule-based policies introduced by [14], where agents are blocked from interacting with others if it receives a given amount of complaints. However, one distinction from our approach is the fact that the disabled state is not permanent. They get woken up after a set number of iterations and get arbitrarily assigned new *op* values, and continue on to their normal behaviours. The reason behind this process of waking agents up is two-fold, firstly for the assimilation of the dynamic nature of any social network. In a real-life scenario, there are constantly new users joining the network, even in the likely event of mass account banning, the population should still stay at a healthy amount, our mechanism simulates this exact process of new users replacing old users in a social media network.

Chapter 5

Evaluation

5.1 Evaluation Metrics

5.1.1 Misinformation

The metric for misinformation is a variance measure between the individual opinion levels and the truth [4]. This can be defined mathematically as:

$$MI_t = \frac{1}{n}\sum_{i=1}^{N}(op_{i,t} - \theta)^2 \tag{5.1}$$

It is a general summary of how misinformed society is, but it does not capture the variances between people's opinions.

5.1.2 Polarization

The measure for polarization [4] is defined as follows, which is an adaptation of the original definition from Esteban and Ray [10]:

$$POL_t = \sum_{k=1}^{K}\sum_{k=1}^{K} P_{k,t}^{1+\omega} P_{l,t} |\overline{op}_{k,t} - \overline{op}_{l,t}| \tag{5.2}$$

K is the number of pre-assigned groups for initial opinion initialisation. $P_{k,t}$ is the percentage of group k at time t out of the whole population. $\overline{op}_{k,t}$ is the average opinion of group k at period t. (The t can not be omitted since agents within each group k or l change throughout the iteration period). ω represents inter-group heterogeneity in reference to [10], which is within the range $[0, 1.6]$. This representation of polarization is comparable to the measure of the Gini coefficient, where different groups of the population are compared pair-wise for their differences.

The metric captures the level of disagreement amongst the population.

5.2 Model Validation

To conduct subsequent research, we must first evaluate the validity of our replicated model. Specifically, we would like to compare the results of opinion evolutions through fixed simulation periods ($T = 1000$) and see if the findings from both M_{mcs} and M_{abm} conform to each other. We run each model 10 times with 10 different sets of initialisation parameters (different initial opinion assignments, agent distributions) under base case settings as defined in 5.1. Following the evaluation metrics used in the benchmark case of [4], we collect the corresponding information of each agent type (all agents, L-bot followers, R-bot followers) for every period of the simulation for both models. Comparisons of results for both models are shown in figure 5.1.

From the left panel of Figure 5.1, We observe a fairly similar pattern as per described in [4]. There is an initial divergence in average opinion levels for L-bot and R-bot followers. This increase in opinion dispersion comes to a halt after roughly $t = 40$, where the average opinion of both groups gets pulled towards $\theta = 0.5$ by the influence of the unbiased source and eventually stabilizes around $t = 600$. Despite this, the average opinion level of all agents is shown to be deviating further and further away from θ,

Parameters	Symbol	Value
Attention to friends	w	0.5
Truth	θ	0.5
Probability of agent communication	P_{comm}	0.5
Probability of learning truth	P_{truth}	0.5
Number of agents	N_a	1069
Level of bot aggressiveness	A_b	25 (base case), 5, 1
Percentage of bot followers	P_{bf}	0.15 (base case), 0.3, 0.6
Number of bots	N_b	2 (base case), 10
Percentage of inoculated agents (social inoculation)	P_{si}	0.2
Range of inoculation (social inoculation)	R_{si}	[0.2, 0.8]
Range of not banning (account banning)	R_{ab}	[0.1, 0.9]
Time of sleep (account banning)	T_s	10
Number of simulations	M	10
Number of groups for opinion initialisation	K	7

Table 5.1: This table contains a list of parameters used for the experiments conducted. The term "base case" in bracket refers to the base-line settings that are used in the validation experiment as well as the bot behaviour experiments for the cases of no intervention.

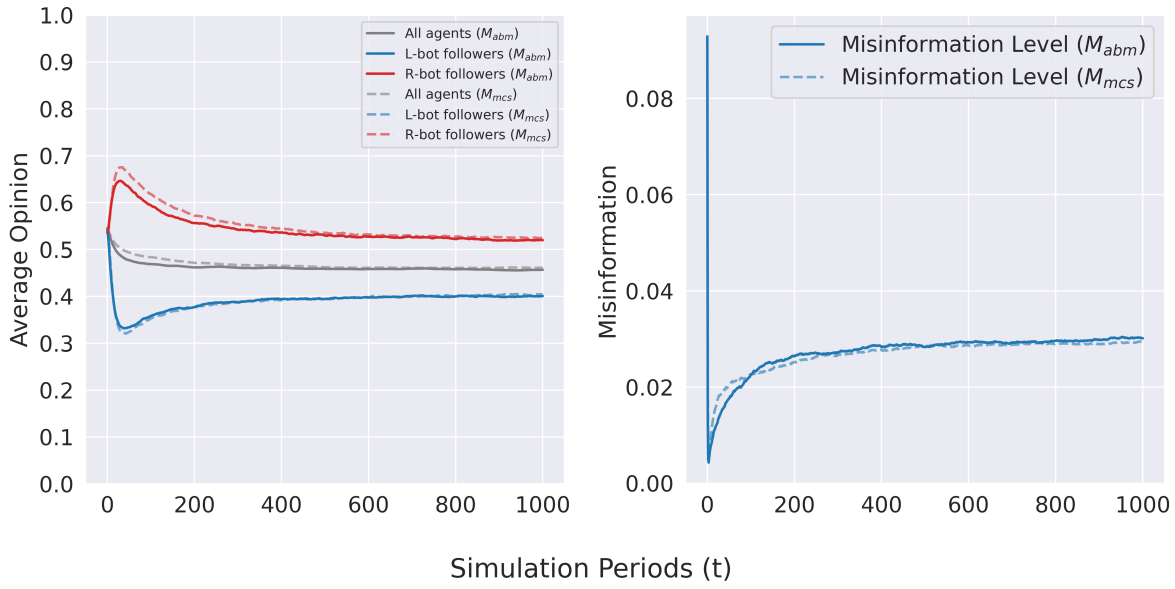

Figure 5.1: In the left panel, we compare the average opinion evolutions between M_{abm} and M_{mcs}. The data is collected for three different agent categories, specifically, the agents that follow the left-wing bot or right-wing bot as well as all agents combined. For the right panel, an overall level of misinformation within the society is measured at each iteration for 1000 iterations.

which can be shown more clearly by the right panel of Figure 5.1. We explain the reason behind this observation through the behaviour of bots. Due to their property of posing as their real agents at first, their influence slowly built up and eventually out-weighing the moderation effect brought by the unbiased source. Hence, there still exists a considerable amount of misinformation within society ($MI_{1000} \approx 0.03$). Another note-worthy thing is the slump in misinformation level at the very start of the simulation, at a magnitude of roughly 0.08. This is no surprise given the distribution of initial agent opinions is sparse, scattered across $K = 7$ groups, which makes the moderation effect significant at first.

In order to obtain a quantitative result for the validity of this replication, M_{mcs}, we choose the metric of root mean square error (RMSE), and comparing in a pair-wise fashion, the corresponding regression line of time-series data (average opinions of all agents, L-bot followers and R-bot followers) of the two models. However, RMSE is a scale-dependent measure [18] which fails to provides meaningful information

on the evaluation of success, hence the need to transform it into its scale-free counterpart, normalised root mean square error (NRMSE), as calculated as follows:

$$NRMSE = \frac{RMSE}{op_{avgmax} - op_{avgmin}} \tag{5.3}$$

Where op_{avgmax} is the maximum average opinions of M_{mcs}'s 10 simulations across 1000 iterations, whereas op_{avgmin} is the minimum. The results produced from comparing the NRMSE difference between the two models are shown in Table 5.2, for which all three metric comparisons through $NRMSE$ result in percentage values that are < 0.2. Thus we can determine M_{abm} to be valid.

	Average opinion (L-bot follower)	Average opinion (R-bot follower)	Average opinion (all agents)
NRMSE	0.022	0.072	0.086

Table 5.2: NRMSE is calculated between the corresponding regression lines (average opinion level against simulation period) from the left panel of Figure 5.3. For each case, we decide that a percentage difference under 20% ($NRMSE < 0.2$) has similar patterns of opinion evolution.

5.3 Intervention Evaluation

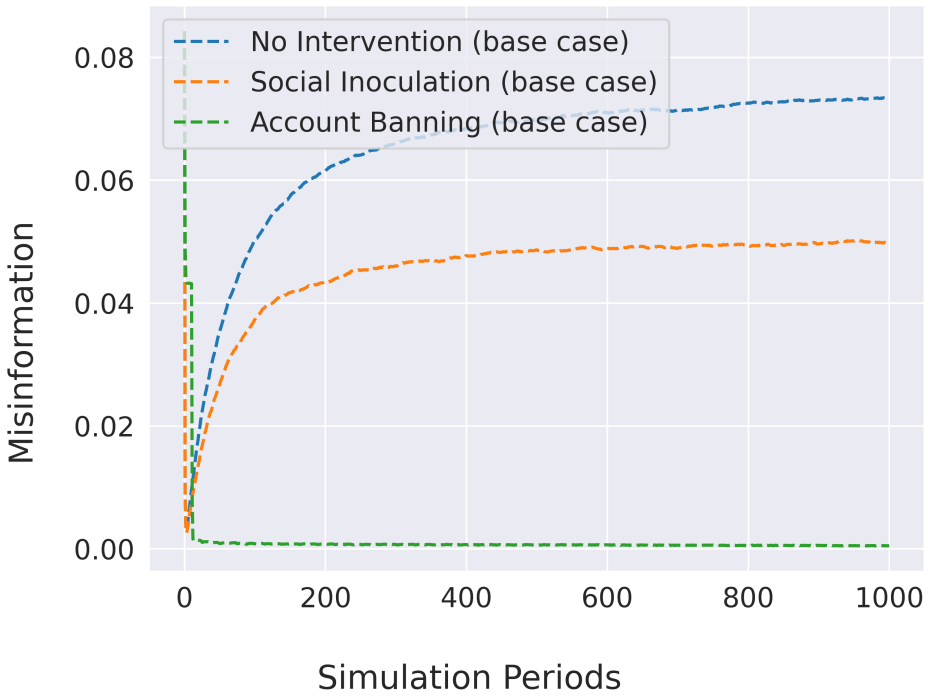

Figure 5.2: Misinformation level evolution applied under no intervention, social inoculation and account banning

We evaluate the relative effectiveness of social inoculation and account banning by conducting the same simulation procedure as in the model validation case, but this time under three different settings: without any intervention applied, social inoculation being applied, and account banning being applied. We record the misinformation and polarization level changes that occur over time for each simulation and additionally collect the final opinion distribution of individual agents. We also conduct each experiment 10 times to reduce bias. From the results observed in Figure 5.2, we can observe that both social inoculation and account banning have achieved various degrees of success in curbing misinformation.

Quantitatively, the level of misinformation that exists within the system starts out quite high (around 0.08), this is no surprise considering our initial opinion distribution is quite scattered. We then immediately observe a sharp decrease in misinformation level for all three curves as agents start receiving truth.

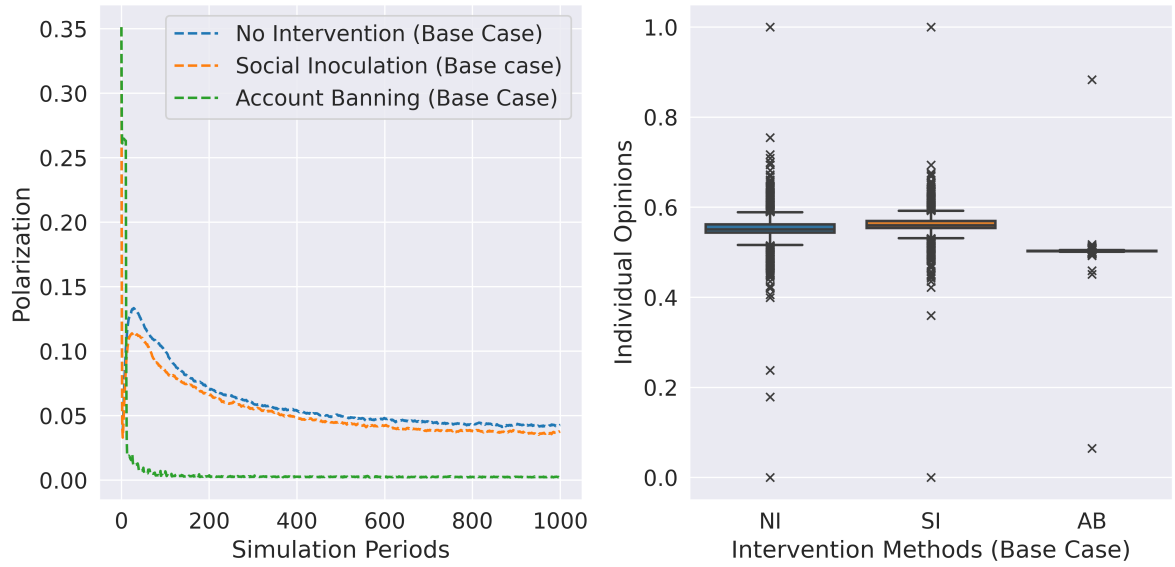

Figure 5.3: The left panel shows polarization level evolution applied under no intervention, social inoculation and account banning , and the right panel shows the corresponding final opinion distributions for these three approaches.

However, without any intervention applied, this quickly gets overshadowed by the influence of bots, as they get extreme over time, causing a steady increase in misinformation. Eventually, the misinformation level reaches approximately 0.07 at the end of 1000 periods of simulation and continues to show a trend of increase. In contrast, social inoculation quickly stabilises misinformation after 200 periods, converging to a value of 0.03. Account banning as expected completely outperforms social inoculation as the level of misinformation plummets down to almost zero at the start of the simulation.

We also look into reasons for the efficiency behind account banning, This is somewhat expected behaviour for two reasons. Firstly, in the context of our model structure, shutting down one agent is extremely effective in stopping its opinion from spreading, as it not only puts it goes into a disabled state, during which the agent does not update its own opinion for a certain period of time, but also makes the neighbouring nodes ignore it. Due to the particular behaviour exhibited by a bot agent (rapidly converging to the extreme), they instantly get spotted and shut down before making an impact on the system. Secondly, as have been pointed out, in a real-life scenario, the misinformation spreader does not always get detected or responded to in time, causing the "toxin" to diffuse thoroughly long before any form of intervention can take place. However, the fact that we set our intervention to take place right from the beginning and for every single iteration, has made the emergence of extreme agents almost impossible, and likely opinion deviations people might have would be immediately adjusted back through receiving the truth.

Despite the aforementioned issues brought by our simplistic approach to modelling accounts banning impose limitations on our ability to capture real life, we still are able to derive something interesting from the measurements. For instance, we observed that the population banned in each iteration, let alone the first iteration (initially opinions are equally distributed along $[0, 1]$ in 7 groups), are quite small. This implies that a timely response to the emergence of misinformation would stamp it out quickly without the need for involvement later.

The left panel of Figure 5.3 shows the polarization level shifts over 1000 interations under the base case set. Social inoculation is observed to slightly decreases polarization by a slight margin. On the other hand, Account banning reduces the polarization level down to zero, this is an expected outcome since almost all agents converge to the truth value θ and there exists little disagreement amongst the population.

In addition, the right panel of Figure 5.3 shows the distribution of individual opinions at 1000 iterations. The median values of individual opinions under no intervention and social inoculation sit close to 0.6, Whereas for account banning it is close to 0.5, our ground truth value. All three medians sit at the centre of their interquartile range (IQR henceforth), meaning the data are not skewed. There is a small

amount of dispersion of opinions for no intervention and social inoculation, as indicated by the small IQR range, whereas for account banning the dispersion is almost non-existent. Another observation is that the top and bottom-most outliers are opinions that belong to bot agents, since their opinions reach the extremes (0 and 1 respectively) overtime at a fixed rate and can not be rectified due to their sink nature (no outgoing connections to other agents).

5.4 Bot Behaviour Analysis

5.4.1 Relative Effectiveness Comparison on Misinformation

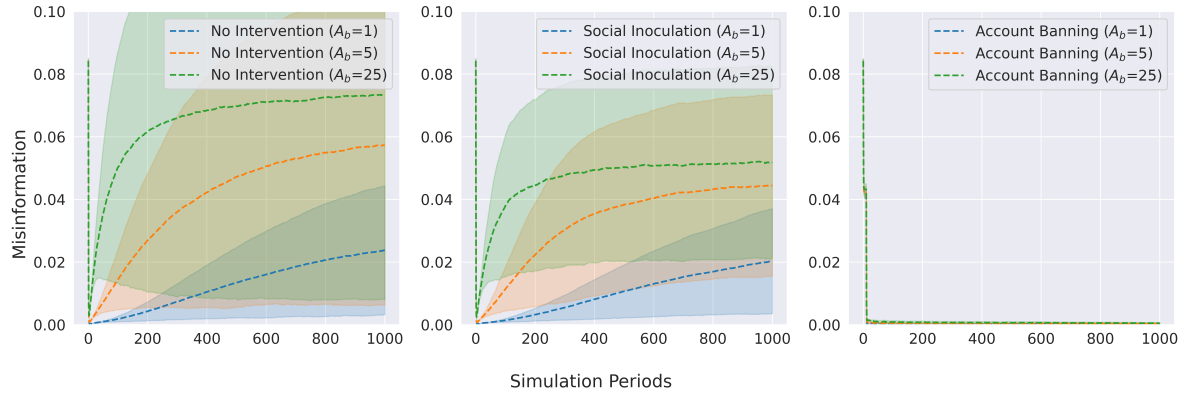

Figure 5.4: Misinformation level evolutions of varying levels of bot aggressiveness ($A_b = 1, 5, 25$) applied under no intervention, social inoculation and account banning (averaged across 10 simulations, the shaded area represents 1 S.D.).

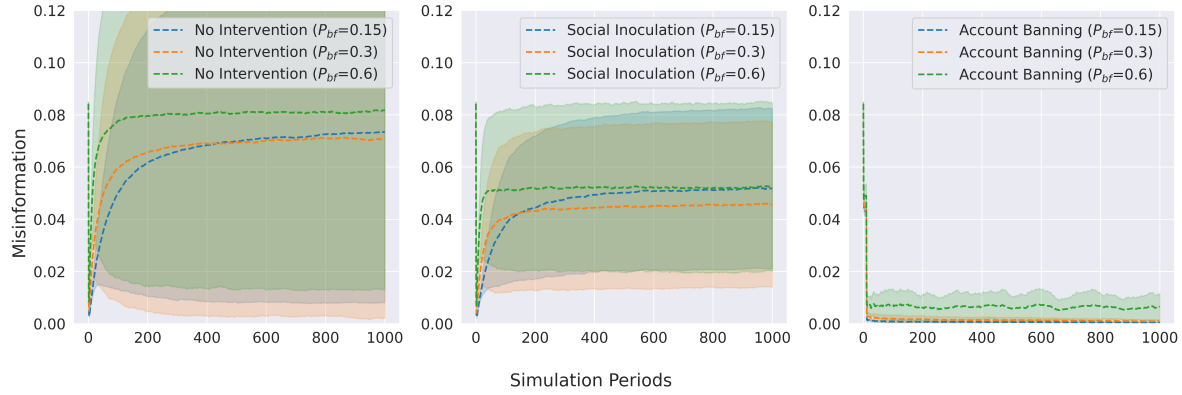

Figure 5.5: Misinformation level evolution of varying percentages of bot followers within the society ($P_{bf} = 0.15, 0.3, 0.6$) applied under no intervention, social inoculation and account banning (averaged across 10 simulations, the shaded area represents 1 S.D.).

Alternative Bot Aggressiveness

Bot Aggressiveness (A_b), also dubbed as flooding capacity, "...measures the ability of bots to spread fake news at a different rate ..." [4]. Our base case is conducted under the setting of $A_b = 25$. By observation of bot opinion changes in that specific circumstances, the bots converge to the extreme (0 for L-bot and 1 for R-bot) within very few iterations. We hence propose an alternative set of experiments, in which we attempt to slow down this convergence by lowering the value of A_b (specifically $A_b = 5$ and $A_b = 1$). In other words, we investigate the impactfulness of interventions for societies with less

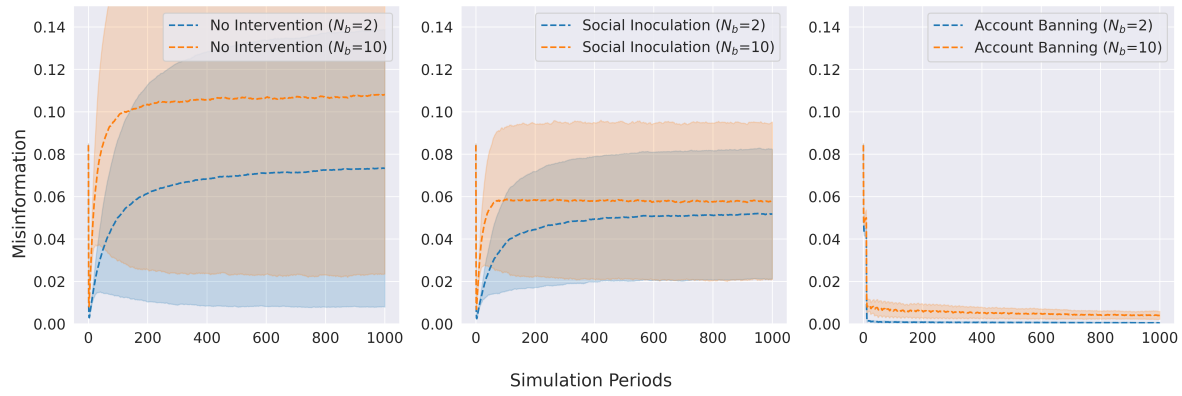

Figure 5.6: Misinformation level evolution of varying numbers of bots within the society ($N_b = 2, 10$) applied under no intervention, social inoculation and account banning (averaged across 10 simulations, the shaded area represents 1 S.D.).

aggressive misinformation spreaders. The other properties of the system act as control variables and stay constant throughout the experiments.

We start by analysing Figure 5.4, under no intervention, misinformation level correlate positively to bot aggressiveness, meaning the more aggressive the bots are the more the misinformation spreads. The significant size of standard deviation exhibited by the shaded regions of $A_b = 5$ and $A_b = 1$ has the implication that bots are a lot dependent on network structures to succeed in swaying public opinion.

Social inoculation performs best with high A_b values, at $A_b = 25$, the presence of the intervention provides a reduction of approximately 0.04 in misinformation level. For lower values of A_b, the reduction is not as significant in comparison, providing little improvement at $A_b = 5$ and almost no impact is observed at $A_b = 1$. A likely reason for this diminishing is due to how social inoculation is implemented, as each inoculated agent is controlled by the size of R_{si} in order to determine whether the information is rejected or not, the effect it will have on less aggressive bots is minimal. These bots lack the strength to divert the opinions a lot away from the truth, which means most people often end up within the R_{si}, where the inoculation rule would not apply.

Account banning on the other hand out-performs social inoculation across all levels of A_b, eliminating misinformation close to zero. This has the implication that account banning is unlikely to be affected no matter how extreme the bots get.

Alternative Bot Follower Sizes

Under base case settings, the percentage of bot followers (P_{bf}) is set to be 0.15, meaning 15% of the total population is under the influence of bots. How much will the interventions be affected if applied within a more biased society? We investigate this by conducting experiments with higher levels of bot follower percentages, specifically 0.3 and 0.6.

As presented in the left-most panel of figure 5.5, for which there is no intervention applied, we observe that for different levels of P_{bf}, the societies all converge to very similar levels of misinformation where the difference that lies between them is less than 0.01. This is an interesting observation in the sense that even if most of society is loyal followers of bots, as long as they are also receiving the truth, the society will not become too misinformed in the long run.

Social inoculation is about equally as effective to all three levels of P_{bf}, reducing misinformation by levels of $0.02 - 0.03$. Whereas the effectiveness of account banning is surprisingly held back to quite some extent as P_{bf} rises, which is observed most clearly for the case $P_{bf} = 0.6$. Our specific implementation of account banning focuses its attention solely upon extreme agents ($op_i \notin [0.1, 0.9]$), while neglecting those that fall within. This implies there still remains a substantial amount of "semi-extreme" agents in the system, and if that amount reaches a certain threshold as in the case of a large percentage of bot followers, bots will be able to overcome the moderation effect brought by the unbiased source.

Alternative Bot Sizes

In the final experiment, we focus on the effect brought by an increase in bot counts from 2 to 10, in which 5 are left wing, and the other 5 are right wing. Quoting from [4], this alters bot behaviour by "increasing the attention span they receive from their followers".

Populating the system with 5 times the number of bots increases misinformation by 0.04 under no intervention. This is quite significant, considering it is more than a 50% increase from the initial misinformation level. In the case of social inoculation, we observe reductions of approximately 0.04 and 0.05 accordingly for cases of $N_b = 2$ and $N_b = 10$ when compared against no intervention. Account banning fails to reduce misinformation completely down to zero with 10 bots present in the network, although it is not too far from achieving it.

5.4.2 Relative Effectiveness Comparison on Polarization

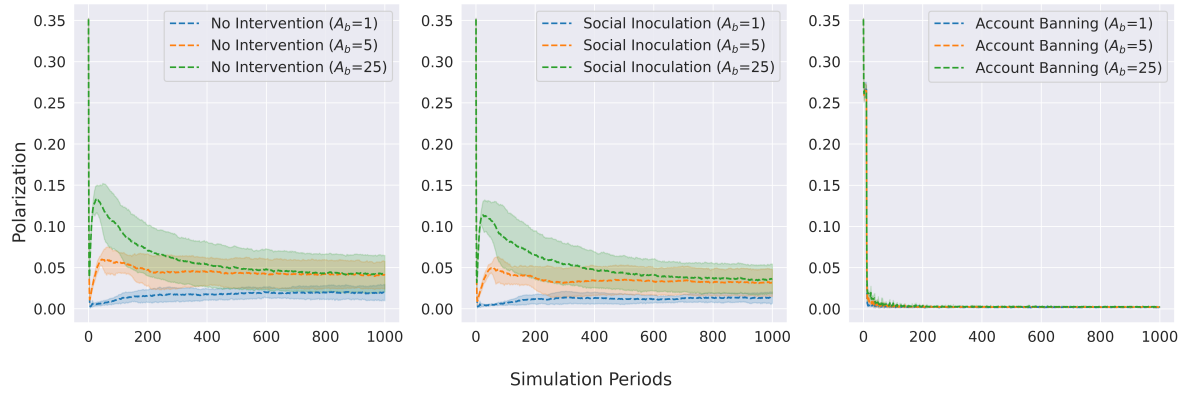

Figure 5.7: Polarization level evolutions of varying levels of bot aggressiveness ($A_b = 1, 5, 25$) applied under no intervention, social inoculation and account banning (averaged across 10 simulations, the shaded area represents 1 S.D.).

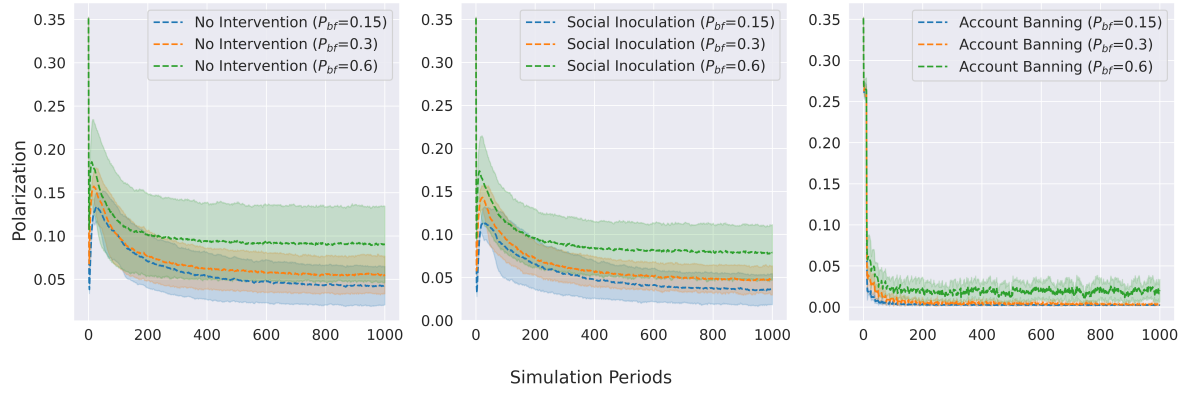

Figure 5.8: Polarization level evolution of varying percentages of bot followers within the society ($P_{bf} = 0.15, 0.3, 0.6$) applied under no intervention, social inoculation and account banning (averaged across 10 simulations, the shaded area represents 1 S.D.).

The subject of misinformation and polarization is closely interrelated. Intuitively, an idealized world in which every single person conforms to the truth, where in that case both misinformation and polarization reach zero. Hence we may argue that even though the intervention methods in Chapter 4 is designed with the intention of stopping misinformation, by either removing extreme agents (account banning) or immunizing agents from receiving extreme opinions (social inoculation), opinions should converge towards the value of the truth, meaning ideally the level of polarization should in turn be reduced as well.

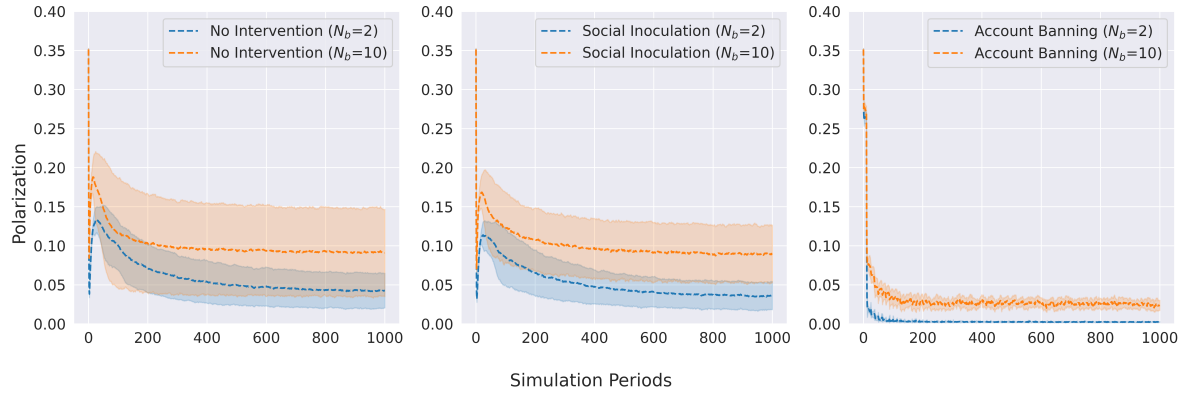

Figure 5.9: Polarization level evolution of varying numbers of bots within the society ($N_b = 2, 10$) applied under no intervention, social inoculation and account banning (averaged across 10 simulations, the shaded area represents 1 S.D.).

In order to verify this postulation, we again conduct three sets of experiments by varying A_b, P_{bf}, N_b, but this time with polarization as the primary metric. The results are presented in Figure 5.7, 5.8, and 5.8. By just eyeballing the results produced, we can clearly observe from under the common setting of no intervention, society's level of polarization across all three experiments is positively correlated to the value of A_b, P_{bf} and N_b. This is somewhat expected due to all three factors we vary relate positively to the ability of bots to spread misinformation, and strengthing any one of them, has the potential to drive society towards being more extreme. With that said the finding is not trivial, as the presence of more "powerful" bots could also imply that more people would be influenced and cluster towards one end of the extreme, and as a result, polarization could in fact decrease. We do not observe this, however, possibly due to the experiment having a balanced proportion of bots, causing society to diverge in two different directions.

Eyeballing quickly becomes inadequate when we try to analyse the magnitude of difference between cases with and without intervention. Especially for the cases of social inoculation, the patterns observed for polarization changes are almost visually identical to the cases of no intervention. We try to address this limitation by conducting a statistical analysis of variance between the final distributions of opinions between experiments under different settings. As has been proven by [4], the measure of opinion distribution variance measures correlates highly to polarization, hence the decision to use them interchangeably.

Statistical Analysis

We examine whether the results produced are significant by comparing the statistical difference between different relevant properties of the bot (bot aggressiveness, bot size, bot follower percentage) and the corresponding final distribution of agent opinions. The following null hypothesis is proposed:

Null Hypothesis H_0: *There is no significant difference between the final opinion distributions of agents with or without an intervention strategy applied.*

We start with conducting a normality analysis on each distribution using the Shapiro-Wilk test to decide whether the subsequent statistical tests should be parametric or not. From the results, we see that none of the opinion distributions falls under the category of normal distribution. For any data that are not normally distributed, we conduct non-parametric tests for the task of significance testing. The Kruskal-Wallis test is chosen for pair-by-pair comparisons between distributions[1], to decide whether H_0 should be rejected or not.

From the results found in Table 5.4, we observe that the Kruskall-Wallis tests for all but one pair of distributions have resulted in p values above 0.01, hence for those below 0.01, we can safely assume that their differences are significant enough to reject H_0. The one exception is the test between no intervention and social inoculation under $P_{bf} = 0.30$. The p-value is quite significant at 0.44, showing a high degree of similarity between the two opinion distributions after indicating the fact that social inoculation lacks the impact to influence polarization within a society in which 30% of the people are followers of bots.

[1] In this case the comparison is done pair-wise. Hence we can also refer to it as Mann Whitney U test

ϵ^2	**Interpretation**
0.00 <0.01	Negligible
0.01 <0.04	Weak
0.04 <0.16	Moderate
0.16 <0.36	Relatively strong
0.36 <0.64	Strong
0.64 <= 1.00	Very strong

Table 5.3: This interpretation of ϵ^2 results is inspired by how the correlation coefficient is measured, additionally squaring each bin's lower and upper boundaries to accommodate the squared nature of the metric [34].

Bot Property	Property Value	Interv. 1	Interv. 2	H Statistic	p-value	Effect Size
A_b	1.00	No Intervention	Account Banning	1231.05	< 0.01	0.58
	1.00	No Intervention	Social Inoculation	16.11	< 0.01	0.01
	5.00	No Intervention	Account Banning	1189.26	< 0.01	0.56
	5.00	No Intervention	Social Inoculation	193.32	< 0.01	0.09
	25.00	No Intervention	Account Banning	1198.93	< 0.01	0.56
	25.00	No Intervention	Social Inoculation	156.16	< 0.01	0.07
P_{bf}	0.15	No Intervention	Account Banning	1198.93	< 0.01	0.56
	0.15	No Intervention	Social Inoculation	156.16	< 0.01	0.07
	0.30	No Intervention	Account Banning	1122.22	< 0.01	0.53
	0.30	No Intervention	Social Inoculation	0.60	≈ 0.44	0.00
	0.60	No Intervention	Account Banning	754.15	< 0.01	0.35
	0.60	No Intervention	Social Inoculation	7.02	< 0.01	0.00
N_b	2.00	No Intervention	Account Banning	1198.93	< 0.01	0.56
	2.00	No Intervention	Social Inoculation	156.16	< 0.01	0.07
	10.00	No Intervention	Account Banning	975.31	< 0.01	0.46
	10.00	No Intervention	Social Inoculation	86.24	< 0.01	0.04

Table 5.4: A_b, N_b and P_{bf} refer to bot aggressiveness, bot follower percentage and bot size. H statistic corresponds to the test statistics from the Kruskall-Wallis tests conducted, and p dictates whether H_0 should be rejected or not (The significance level, in this case, is pre-defined to be 0.05). The effect size is computed using an interpretation of epsilon square [44].

.

However, not only do p-values fail to convey the strength of a difference between distributions, their expressiveness can be significantly held back by a large sample size [42]. We must introduce another metric, effect size, which quantifies such differences using a percentage value to solve these limitations. We calculate effect sizes using epsilon square (ϵ^2) [20]. To draw meaningful conclusions behind the exact values of ϵ^2, we follow the approach used by [34] to evaluate the meanings behind the effect sizes measured, as illustrated by Table 5.3. cite The findings we found by analysing the effect size results largely conform to the patterns observed in Figures 5.7, 5.8 and 5.9. Across most levels of A_b, N_b and P_{bf}, account banning performs equally as well in reducing polarization, as shown by the strong effect sizes. The only case where its effectiveness is seen to be held back is when $P_{bf} = 0.6$, in which case the effect size falls to relatively strong.

In contrast, social inoculation performance in reducing polarization is less than ideal. Judging from the effect size measures, under cases of $A_b = 5, 25$, $P_{bf} = 0.15$ and $N_b = 2, 10$, there are moderate differences between the distribution pairs, whereas for $A_b = 1$ and $P_{bf} = 0.3, 0.6$ then the difference is negligible.

The findings above are somewhat surprising, considering although social inoculation's effectiveness can be outshined by account banning, as is in the case of misinformation, their performance difference is not as drastic as in the case of polarization mitigation. In other words, the ability to filter out information outside of R_{si} given to some agents is not influential enough to keep society away from polarization.

Chapter 6

Reflections and Future Works

There obviously is a cost to accuracy attached when designing intervention strategies through the "KISS" approach. The effect produced from the base implementation of account banning fails to produce interesting results, Even worse, due to the extreme "effectiveness" exhibited by these assumptions, we fail to deduce the proper upper bound of misinformation reduction as most of them converges to zero.

Future work could attempt to remedy this by incorporating more real-world observations that downplay the effectiveness. For instance, we have observed from real life that there is a general sense of delay before any intervention occurs (Algorithms tend to outperform humans, but even they require processing time). Someone could model this belief by simply delaying the time at which the account-banning mechanism is applied.

ABM's ideology of modelling poses significant limitations for us when deriving induction on the derived phenomenon. Specifically the fact that whether the observations from our experiments come as results of the algorithmic implementation, or it is something more meaningful that might have real-world implications. This possible variability is based solely on the accuracy of the implementation contradicts the ideology of the "KISS" approach when designing an ABM, and further research could be conducted on possible ways to mitigating the impact it caused and ways to improve the future of ABM design.

An observation made from Figure 5.4, 5.5 and 5.6 by measuring their standard deviations across different simulations is that the spread of the misinformation level is very large, and we can not definitively tell that whether the patterns we observed is replicable, or just purely by coincidence.

Regarding extensions, many alternative approaches can be experimented with based on our model. Our model's update process is unilateral in that each interaction only affects the reader of the post (the source node) but not the author (the destination node). [30] proposed an alternative information diffusion process, which has considered the sanctioning process that usually occurs in STS. It has been suggested that participants within such STS, may provide sanctions in the form of likes, dislikes, and comments in an attempt to drive the STS in the direction that aligns with their own preferences [30]. This bilateral update process could be an interesting future extension to research such as a social media network.

More variations in agent behaviours are another point of extension. For example, in a real-life scenario, bots do not always intend to go for the extreme but rather promote a view that would benefit a certain party. This could lie anywhere along the range from 0 to 1 rather than 0 or 1. Future research could look into the impact of having such "non-extreme seeking bots" on society.

There is also a lot of variability within the system that we have not investigated fully, such as the scenario under which the truth value regarding the specific topic is not 0.5 as we assumed, and the impact that might have on different interventions.

Chapter 7

Conclusion

By [4]'s definition on a "wise" society, the level of misinformation as well as polarization in the the long run should all converge to zero. We investigate the impact of two intervention methods that are commonly applied these days to investigate their effectiveness in moving society towards a wise state, as well as testing their resilience to variances in bot properties.

In order to address RQ1, we followed the implementation of [4] on the simulation of (mis) information diffusion on social media and implemented it in an ABM fashion. We synthesized algorithmically our own graph for the representation of social network structure. We designed and implemented the methods of social inoculation and account banning as ways to intervene against malicious activities by bots spreading misinformation. We then applied the two interventions within the model and ran the model for 1000 simulations and ten times iterations with different initialisation of agent settings. We found out that account banning out-performs social inoculation by a large margin. This is mostly a cause of our implementation decisions and does not necessarily reflect the same outcome in real life. However, one conclusion we can draw from this is that if given an optimal version of the account banning (detecting misinformation constantly and being able to identify the misinformation spreader quickly), society could converge to a wise state.

To address RQ2, we decided to vary bot properties based on the bot aggressiveness, bot follower sizes and bot sizes and measure and analyse the different efficacies of the interventions reflected by misinformation and polarization levels through several sets of controlled experiments, and additionally, conduct an extra set of non-parametric statistical analyses in the form of Kruskal-Wallis testing to find out quantitatively the difference made with or without intervention on the level of polarization. Our findings are as follows:

- On the subject of reducing misinformation, We have found that social inoculation performs worse when bots are less aggressive and lurk around longer, making them undetected by intervention methods in the short term. It performs better in a highly misinformed society than in a less misinformed one. In comparison, account banning can be slightly held back when society is highly misinformed with a lot of followers of bots, or with bots that occupy most of the people's attention span, but mostly still outperforms social inoculation under the same parameter setting.
- Regarding reducing polarization, social inoculation performs poorly across most parameter settings where we see no substantial improvement in polarization level. On the contrary, polarization performs equally as well across all parameter settings, although being slightly held back when the percentage of bot followers is high.

We conclude by stating that all intervention strategies have varied impacts on different aspects of social welfare. There is no panacea offered to address them all at once. The different efficacies of the intervention strategies under varying circumstances provides inspiration for future intervention designs.

Bibliography

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

[2] Douglas J Ahler and Gaurav Sood. The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*, 80(3):964–981, 2018.

[3] Robert Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, 1997.

[4] Marina Azzimonti and Marcos Fernandes. Social media networks, fake news, and polarization. *European Journal of Political Economy*, page 102256, 2022.

[5] Joseph Bak-Coleman, Jevin West, and Lisa Friedland. Modest interventions complement each other in reducing misinformation, 2022.

[6] Maria Barbati, Giuseppe Bruno, and Andrea Genovese. Applications of agent-based models for optimization problems: A literature review. *Expert Systems with Applications*, 39(5):6020–6028, 2012.

[7] Francesco C Billari, Thomas Fent, Alexia Prskawetz, and Jürgen Scheffran. Agent-based computational modelling: an introduction. *Agent-based computational modelling: applications in demography, social, economic and environmental sciences*, pages 1–16, 2006.

[8] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

[9] Josh Compton. Inoculation theory. *The SAGE handbook of persuasion: Developments in theory and practice*, 2:220–237, 2013.

[10] Joan-Maria Esteban and Debraj Ray. On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, pages 819–851, 1994.

[11] Hans Jurgen Eysenck. Sense and nonsense in psychology. 1957.

[12] Anna Fleck and Felix Richter. Infographic: Americans turn to social media for news, despite lower trust, Feb 2023.

[13] R Kelly Garrett and Shannon Poulsen. Flagging facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. *Journal of Computer-Mediated Communication*, 24(5):240–258, 2019.

[14] Anna Gausen, Wayne Luk, and Ce Guo. Can we stop fake news? using agent-based modelling to evaluate countermeasures for misinformation on social media. In *International AAAI Conference on Web and Social Media (ICWSM). https://doi. org/10.36190*, 2021.

[15] Emory A Griffin. *A First Look at Communication Theory*. McGraw-Hill, New York, 2011.

[16] Weisen Guo and Steven B Kraines. A random network generator with finely tunable clustering coefficient for small-world social networks. In *2009 International Conference on Computational Aspects of Social Networks*, pages 10–17. IEEE, 2009.

[17] Michael Hameleers and Toni GLA Van der Meer. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, 47(2):227–250, 2020.

[18] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[19] Cecilia Kang and Adam Goldman. In washington pizzeria attack, fake news brought real guns, Dec 2016.

[20] Truman L Kelley. An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, 21(9):554–559, 1935.

[21] Will Kenton. What is web 2.0? definition, impact, and examples, Dec 2022.

[22] Antino Kim, Patricia L Moravec, and Alan R Dennis. Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, 36(3):931–968, 2019.

[23] Joshua Klayman. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418, 1995.

[24] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.

[25] William J McGuire and Demetrios Papageorgis. The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *The Journal of Abnormal and Social Psychology*, 62(2):327, 1961.

[26] Adam J McLane, Christina Semeniuk, Gregory J McDermid, and Danielle J Marceau. The role of agent-based models in wildlife ecology and management. *Ecological modelling*, 222(8):1544–1556, 2011.

[27] Wendling Mike. The saga of 'pizzagate': The fake story that shows how conspiracy theories spread, Dec 2016.

[28] Sadiq Muhammed T and Saji K Mathew. The disaster of misinformation: a review of research in social media. *International journal of data science and analytics*, 13(4):271–285, 2022.

[29] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, 2018.

[30] Luis G Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K Kalia, Jaime S Sichman, and Munindar P Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *The Knowledge Engineering Review*, 31(2):142–166, 2016.

[31] Tore Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social networks*, 35(2):159–167, 2013.

[32] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11):4944–4957, 2020.

[33] Liliana Perez and Suzana Dragicevic. An agent-based approach for modeling dynamics of contagious disease spread. *International journal of health geographics*, 8(1):1–17, 2009.

[34] Louis M Rea and Richard A Parker. *Designing and conducting survey research: A comprehensive guide*. John Wiley & Sons, 2014.

[35] Jon Roozenbeek and Sander Van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1):1–10, 2019.

[36] Jon Roozenbeek, Sander Van Der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. Psychological inoculation improves resilience against misinformation on social media. *Science advances*, 8(34):eabo6254, 2022.

[37] Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School (HKS) Misinformation Review*, 2(5), 2021.

[38] Fernando P Santos, Yphtach Lelkes, and Simon A Levin. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50):e2102141118, 2021.

[39] Arun V Sathanur, Miao Sui, and Vikram Jandhyala. Assessing strategies for controlling viral rumor propagation on social media-a simulation approach. In *2015 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–6. IEEE, 2015.

[40] Emilio Serrano and Carlos A Iglesias. Validating viral marketing strategies in twitter via agent-based social simulation. *Expert Systems with Applications*, 50:140–150, 2016.

[41] Hans-Peter Stricker. An ego-centric approach to synthesize more realistic social networks. 2021.

[42] Gail M Sullivan and Richard Feinn. Using effect size—or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282, 2012.

[43] Li Qian Tay, Mark J Hurlstone, Tim Kurz, and Ullrich KH Ecker. A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*, 113(3):591–607, 2022.

[44] Maciej Tomczak and Ewa Tomczak. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in sport sciences*, 21(1), 2014.

[45] Vítor V. Vasconcelos, Sara M. Constantino, Astrid Dannenberg, Marcel Lumkowsky, Elke Weber, and Simon Levin. Segregation and clustering of preferences erode socially beneficial coordination. *Proceedings of the National Academy of Sciences*, 118(50):e2102153118, 2021.

[46] Cheng Wang, Omar Lizardo, and David Hachen. Algorithms for generating large-scale clustered random graphs. *Network Science*, 2(3):403–415, 2014.

[47] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Soc. Sci. Med.*, 240(112552):112552, November 2019.

[48] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90, 2019.