Introduction

In this project, we developed a model of agents to study the impact of hate speech on their behavior. Throughout the process, we incorporated various parameters and tested different configurations to make the model as realistic as possible. The concept is based on the Spiral of Silence model.The model operates by initially assigning agents either positive or negative opinion, with some becoming haters or paragons—extreme influencers. Agents then adjust their behavior according to their confidence levels and the opinions of their neighbors. Additionally, we introduced mechanisms for banning and blocking, which can silence or isolate turtles.

To fine-tune the model with real-world data, we used NetLogo's BehaviorSearch, determining several key parameters aside from those related to intervention and real-world metrics. Finally, we employed BehaviorSpace to simulate sample data and test our hypothesis.The model features two types of agents: those with positive opinions and those with negative opinions. It includes various parameters to control the simulation. The "number of nodes" defines the total number of agents in the model. The "average node degree" determines the average number of connections each agent has with others. The "hater count" specifies the number of agents with extreme negative opinions, while the "paragon count" indicates the number of agents with extreme positive opinions. The "negative opinion ratio" determines the proportion of agents assigned to negative versus positive opinions.

The "Hater link chance" and "hater link count" are used to create new connections for active hate agents, while "paragon link chance" and "paragon link count" serve a similar purpose for paragons. The "ban request chance" simulates the likelihood of bans being requested before confidence is calculated. The "hate influence boost" increases the influence of hate agents when calculating changes in confidence based on neighbors' opinions. "Hater confidence boost" and "paragon confidence boost" enhance the confidence of haters and paragons during each step. The "block chance" determines the likelihood of removing links between positive agents and haters. The "ban success rate" aids in processing bans after the confidence update. Finally, the seed ensures the simulation can be repeated consistently.

During the setup procedure, we begin by initiating the simulation environment, starting with the representation of each individual positive or negative opinion. A user-defined random seed is set to ensure reproducibility. Next, agents are created with specific attributes, such as position, opinion, and others. Following this, the “setup-spatially-clustered-network” function generates a network by linking the turtles. Negative opinions are then assigned based on a specified percentage, with some of these designated as “haters.” Similarly, a portion of the positive opinions is assigned as “paragons,” preparing the environment to run the simulation.

The 'go' procedure begins by verifying whether the "blocking" feature is enabled. If it is, the 'block-haters' procedure is invoked to sever the connections between "haters" and agents with positive opinions before confidence is calculated. Subsequently, all turtles update their confidence levels, recalculating them based on their own confidence and the influence of their

neighbors. After this update, each turtle will decide to either remain silent or become outspoken, depending on its confidence. This step allows the 'process bans' procedure to eliminate some of the links between "haters".

We use different colors to represent the various types of agents, such as hater, paragon, silent, and others, allowing us to visualize the turtles more effectively. The procedures 'add-hater-links,' 'add-paragon-links,' and 'add-link' enhance influence by creating new connections between agents. Finally, the 'stop-check' procedure compares the current status with the previous one, and if no changes are detected, the simulation will terminate.

Our simulations developed based on three approaches. For the first hypothesis, we remove a connections between haters and other turtles based on a percentage to isolate them and reduce their influence. For the second hypothesis, we will strengthen positive opinions against hate speech by enhancing the influence of "paragons" and their connections. This approach reduces the impact of haters without directly removing them. Third, we implemented a banning procedure where hater will loose the connections in network. We believed that the best strategy out of the rest. Our simulation will focus on three hypotheses:

1. Isolating *haters* via blocking is an effective way to combat the effect of hate speech on a network and significantly reduces the silencing effect of the Spiral of Silence.
2. The effectiveness of promoting the spread of positive opinion is significantly more effective in networks with a low degree of interconnectivity (node degree) than in networks with high interconnectivity.
3. Moderation and banning as an approach are significantly more effective at reducing the impact of *haters* than the other two approaches examined in our model under otherwise equal circumstances.

Next, we calibrated the model to make real world representation. The calibration was performed separately across three different node degrees. We developed three combinations for hater count, link chance, and link count in comparison to the average node degree from the calibration. Our model demonstrated a fitness of 2.57, and we decided to proceed with it. To test our hypothesis, we conducted several experiments using the verified model. We utilized BehaviourSpace for automated experiment execution. We did various methods to evaluate the hypothesis, including the t-test, Shapiro-Wilk normality test, Mann-Whitney U test, among others.

The portfolio begins with an introduction, followed by a flowchart of the spiral of silence and an explanation of how the model operates. Next, we discuss our experience using NetLogo to develop the spiral of silence model. We then describe how we integrated hate speech into the model and developed hypotheses to address it, including the ODD protocol. Afterward, we detail the model's implementation and verification. We also explain how calibration was used to make the model as realistic as possible, along with further verification. This is followed by a discussion of the simulation and experimental results. Finally, I included a reflection on the course.

Agent-Based Modeling – Exercise 1

Group 4: Alexander Helwig, Qingbo Qiao, Prathibha Rajapaksha

Task 1: Present the program code in one (or more) clear flow charts.

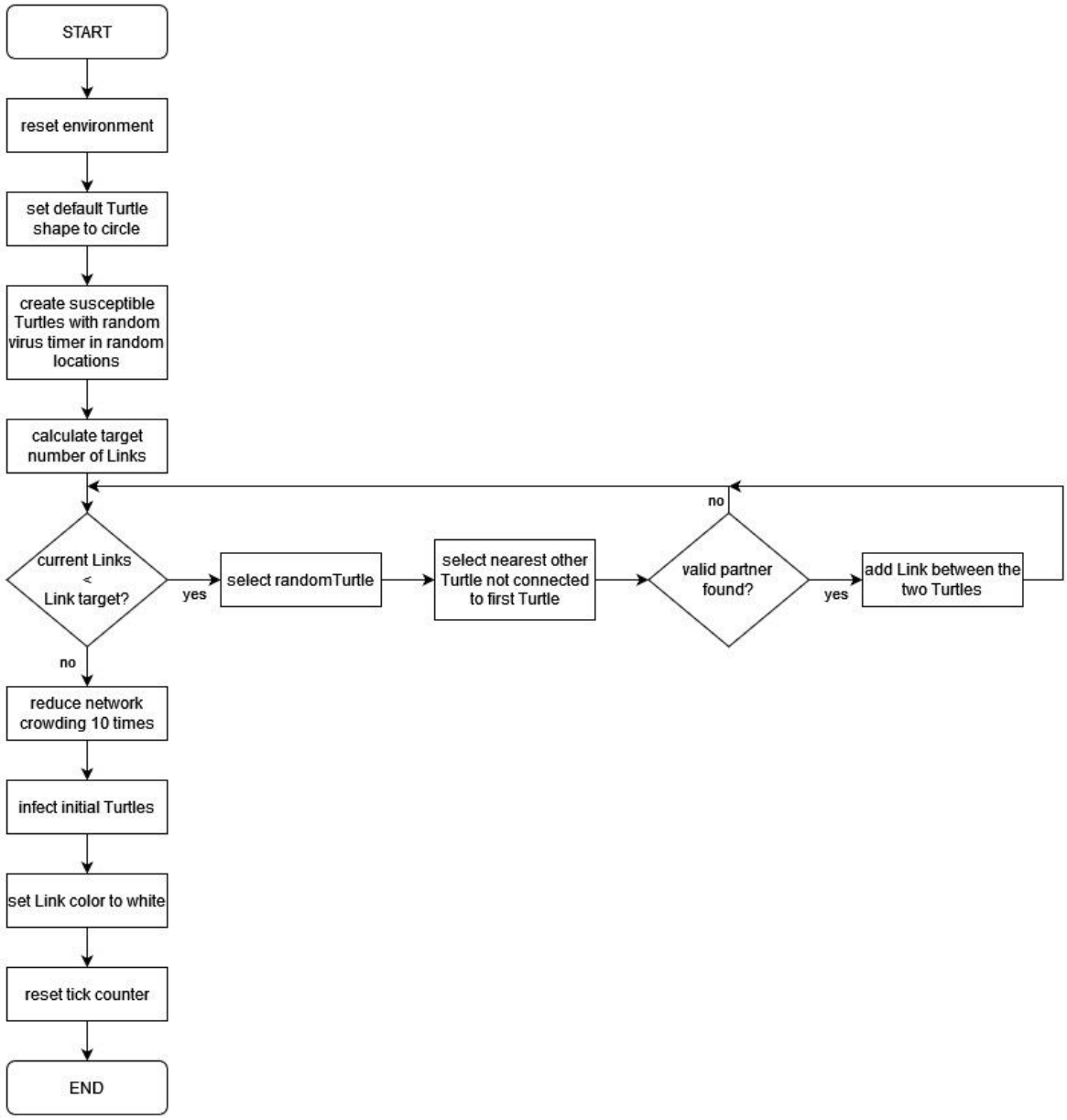


Fig. 1: Flow-Chart for the Setup-Process

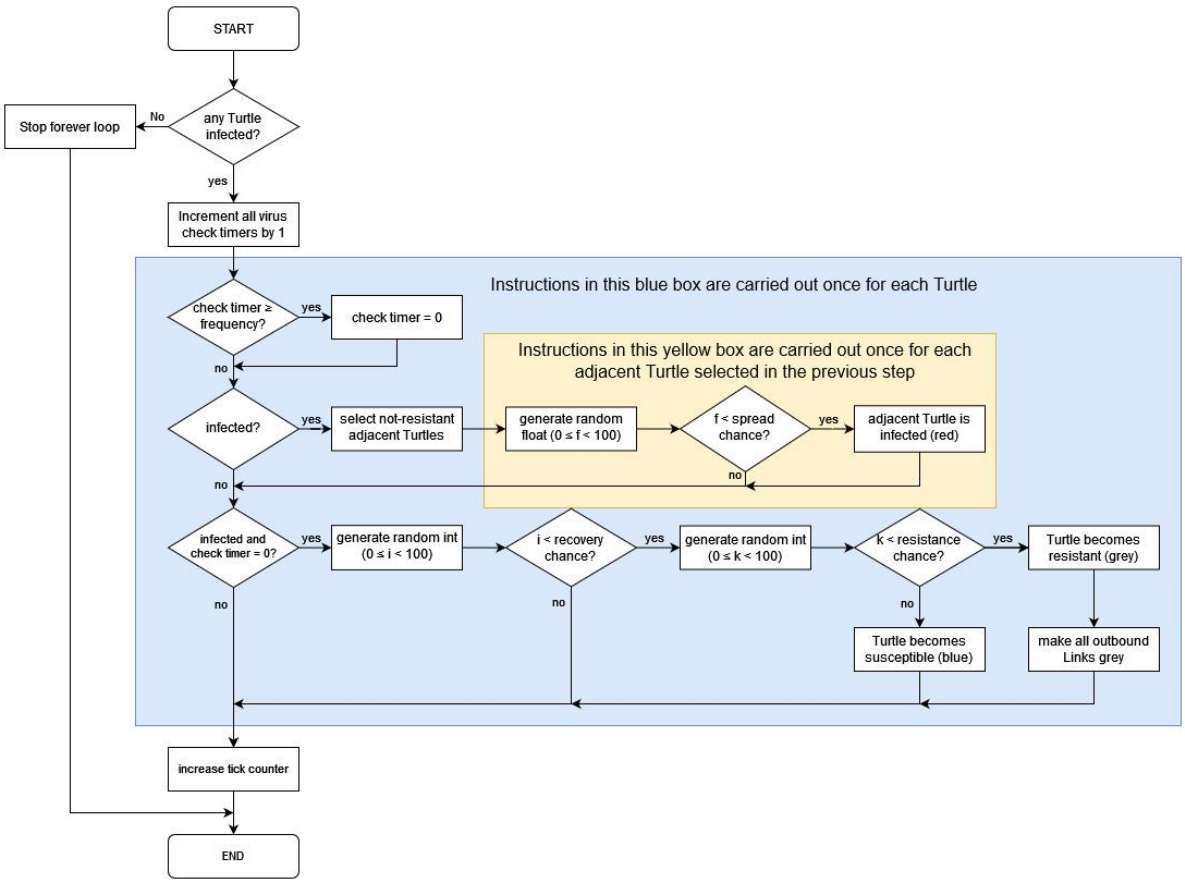


Fig. 2: Flow-Chart for the Go-Process

Task 2: Summarize briefly in your own words how the model works (max. 300 words)!

During setup, the environment is reset and the specified number of susceptible, circular Turtles are created at random coordinates and with a random check timer. Until the model contains a certain number of links, random Turtles are selected and linked to their nearest neighbor they aren't already linked to. Then, network crowding is reduced in ten passes of a modified Fruchterman-Reingold layout algorithm. Finally, the link color is set to white and the tick counter reset.

Initially during each simulation step (go), the program checks if any infected Turtles remain. If not, the program ends. Then, all check timers of Turtles are incremented by one and those whose value is equal to or greater than the check frequency afterwards are set to 0. For each infected Turtle, every non-resistant neighbor is checked to see if they become infected by comparing a random number with the spread chance. Afterwards, the program checks if any infected Turtles with a check timer of 0 recover and whether they become resistant or susceptible again if they recover. This is also done with two random numbers compared to the recovery chance and resistance chance respectively. If a Turtle becomes resistant, all outbound links are greyed out to show that they are no longer relevant for virus spread. Finally, the tick counter is incremented by one.

Agent-Based Modeling – Exercise 2

Group 4: Alexander Helwig, Qingbo Qiao, Prathibha Rajapaksha

Task 1: Briefly describe your entire model and your decisions during modeling (max. 300 words).

In this simulation, individuals experience the "spiral of silence," if they believe they're in the minority. Turtles in the model have either positive (1) or negative (-1) opinions. When a turtle finds itself surrounded by neighbors who disagree with its opinion, it tends to remain silent out of concern for social rejection. On the other hand, turtles whose opinions are the same as most of their neighbors are likely to speak up.

In the go method, turtles update their confidence in their opinion based on the opinions of their neighbors. The update-confidence procedure adjusts the turtle's confidence. It considers the delta value, which shows how much the turtle's opinion differs from its neighbors'. If most neighbors disagree, delta is negative, suggesting the turtle might be in the minority. Similarly, if delta is positive, it means the turtle's opinion matches the majority.

Turtles that are not silent affect the overall opinion atmosphere. If the combined confidence update (ĉ) and delta produce a positive outcome, indicating a stronger confidence in its opinion and agreement among neighbors, then the turtle's confidence (c) is adjusted accordingly. Otherwise, if the result is not positive, the confidence level ĉ is reset to zero before the confidence (c) is calculated.

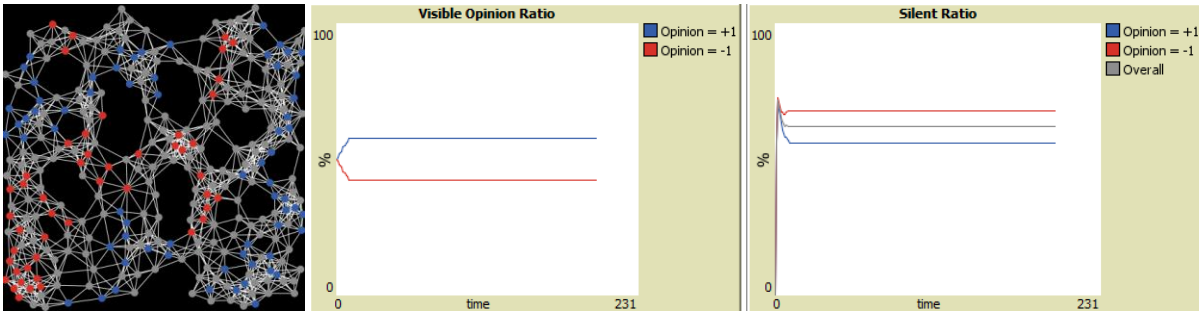
Once the confidence is updated, the turtle examines if its readiness to self-censor surpasses its confidence level. If it does, the turtle chooses to remain silent and changes its color to grey. This process shows how people decide whether to share their thoughts, based on what they sense from the people around them.

Task 2: Define three interesting parameter settings of the model concerning the input variables *number-of-nodes*, *average-node-degree* and *alternative-opinion-ratio*. Run the model with the three parameter settings and compare the results: How do both the visible opinion ratio and the silent ratio of the model change in respect to each parameter setting (max. 600 words)?

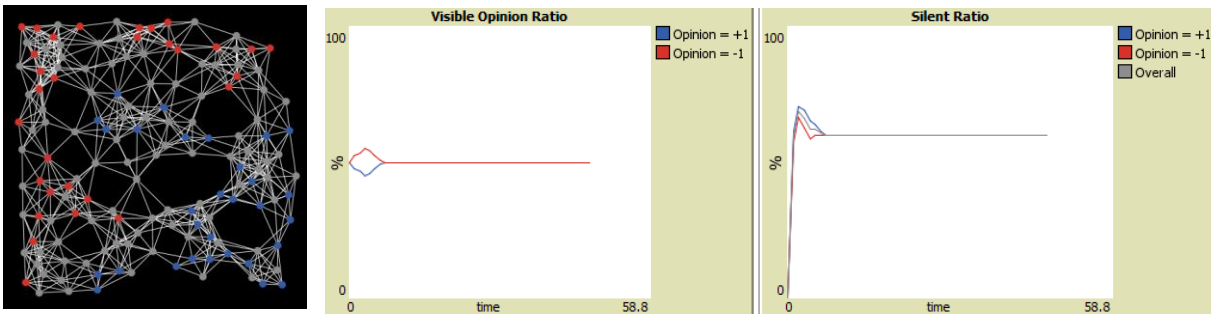
Scenario	Number-of-nodes	Average-node-degree	Alternate-option-ratio
1	300	10	0.5
2	150	10	0.5
3	200	30	0.5

Fig. 1: Scenario overview

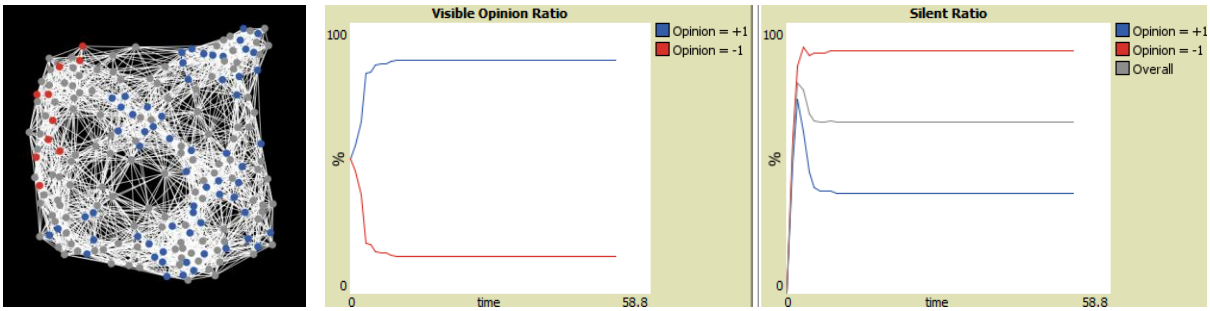
The "Visible Opinion Ratio" chart demonstrates the ratio of turtles that express positive or negative opinions compared to all non-silent turtles. Likewise, the "Silent Ratio" chart illustrates the percentage of turtles choosing to remain silent in relation to all turtles expressing their opinions.



Scenario 1: Positive opinions have a slight dominance, which is also evident in the silent ratio where negative opinions hold a slight edge. Neither opinion significantly dominates as both opinion ratios fall within the 40% to 60% range. The environment doesn't permit either opinion to prevail significantly.



Scenario 2: This is a situation where opinions are evenly balanced. In this state, the likelihood of both positive and negative opinions is 50%. The silent ratio is identical for both opinions. Since there is no dominant opinion, both opinions have an equal probability of being expressed. Unlike the runs of Scenario 1 and 3, this time the negative opinion gained a lead for a short period of time, whereas in the other simulations, the positive opinion dominated.



Scenario 3: Networking has seen a substantial increase with on average three times as many edges per node, making it difficult to even see edges in the network view. There is a noticeable disparity in loud opinion ratios, with positive opinions predominantly leading. The outspoken expression of the positive opinion has surged due to enhanced network coverage (average-node-degree increase), leading to significant social pressure influencing the turtles' opinions. Consequently, a minority of negative opinions remain silent in significant numbers. Despite the dominance of the positive opinion, it never reaches 100%. The maximum percentage reaches around 85%, while it's also apparent that negative opinions stabilize around 15%, as a group of turtles with the negative opinion has become isolated from any influence in the upper left of the graph.

All in all, the number of nodes probably has the smallest impact on visible opinion ratio and silent ratio, though the effects of the spiral of silence seem less severe in smaller groups. A high average-node-degree on the other hand significantly increases the silencing effect, causing the visible opinion ratio and silent ratio to greatly diverge between the two opinions. Finally, the alternative-opinion-ratio has such a great effect that we didn’t vary it for the three scenarios presented here, as even a relatively minor imbalance caused the minority to become silenced after merely a few ticks, with it not being uncommon for the minority opinion to become completely silenced.

Agent-Based Modeling – Exercise 3

Group 4: Alexander Helwig, Qingbo Qiao, Prathibha Rajapaksha

Task 1: Briefly describe how you want to integrate the *hater* agent type into the existing model.

The Spiral of Silence model is an ideal framework to implement the dynamics of hate speech, as agents already base the expression on their opinion on their personal confidence and beliefs as well as the perceived opinions of other agents around them. The social influence exerted by their neighbors can increase or decrease an individual's willingness to speak out publicly as individuals are more likely to speak up if they perceive their own opinion to be well represented in society. As a result, members of minority opinions tend to remain silent out of fear of social isolation or backlash, even if they hold strong opinions.

A different pattern however can emerge if some members believing in a minority opinion are exceptionally vocal and confident in expressing their own opinion, especially if they actively target the opposing opinions with hate speech. In this case, their opinion, even though it belongs to a minority, can be perceived as the dominant opinion, causing speakers of the actual majority opinion to fall silent, which exacerbates the skewed perception of the underlying opinion distribution even further towards that of those vocal actors, which will be represented in the model by the *hater* agent type.

This agent type will differ from normal agents in that its own opinion will have a stronger influence on neighboring agents to represent their vocal and extroverted behavior as well as their tendency to use more radical language. Furthermore, *haters* will be more resilient to changes of their own confidence as a model implementation of the confidence many people using hate speech have about stating their own minority opinion in a public space. This could be implemented by increasing the confidence of *haters* based on the number of agents they are in contact with, or by granting them a static bonus during confidence calculation which can be overcome if their surroundings (almost) entirely share the opposite opinion but will be difficult to overcome otherwise. We will have to see during further model development which of those two options works better for our purposes.

The impact of hate speech can be categorized into individual impact and network impact or group impact. For our implementation, we will focus exclusively on the individual impact, which allows us to closely examine how hate speech can affect personal behavior and attitude without diving into broader and far more complex network dynamics.

Task 2: Develop hypotheses that you would like to examine with your simulation study including three possible interventions to encounter Hate Speech in a network.

While there are many ways to combat the use of hate speech, our simulations will focus on the following three methods:

First, we will implement a system where connections between *haters* and other turtles will be removed based on a certain percentage to isolate the *haters* and reduce their influence in the network as their hate speech reaches fewer and fewer other agents. This approach could represent the deletion of hate speech containing content or something like defriending, muting, or blocking users on an individual basis when users don't want to see someone else's content without reporting them.

Second, we will promote the spread of the positive opinion opposing hate speech, increasing the influence of those who speak out against it by boosting their confidence and connections between those "paragons" and other (positive) agents over time. This approach could represent measures implemented by the network itself to promote positive content to drown out hate speech without directly removing hate speech containing content. This should be more effective in sparse networks where it is easier to drown out the hate speech.

Third, we will implement a ban request system where users that come into contact with hate speech can request a ban of the *hater*, which in the model will be handled by the *hater* being banned with a certain likelihood. Should an individual get banned, they will become silent and all connections to them will be removed, preventing them from influencing anyone else in the network. The key difference to the first approach is that since ban requests take time to process by moderators, the bans will only take effect after confidence is calculated instead of before this calculation. In practice, moderation and banning are an effective but costly method to combat hate speech.

Our simulation will focus on three hypotheses:

1. Isolating *haters* via blocking is an effective way to combat the effect of hate speech on a network and significantly reduces the silencing effect of the Spiral of Silence.
2. The effectiveness of promoting the spread of positive opinion is significantly more effective in networks with a low degree of interconnectivity (node degree) than in networks with high interconnectivity.
3. Moderation and banning as an approach are significantly more effective at reducing the impact of *haters* than the other two approaches examined in our model under otherwise equal circumstances.

Task 3: Identify the required model components and correlations. Present an overview and design concept according to ODD.

The model's purpose is to study the impact of three intervention methods against hate speech in a network of agents with two different opinions being present in the population based on a Spiral of Silence model. The methods examined are muting or blocking agents on a per-agent basis, increasing the reach of agents representing the opinion not spread by the haters, and implementing ban system in the network where individual agents can get banned. More specifically, we will study the effectiveness of blocking in general, the dependence of boosting the spread of positive opinions on the interconnectivity within the network, and whether moderation and banning are significantly more effective than other, less expensive solutions.

The baseline Spiral of Silence model consists of one type of agent representing an individual person in the network. Every agent has a variable representing their opinion, with -1 representing a negative and +1 a positive opinion. Each agent also has a confidence and a willingness to self-censor, both of which range between 0 and 1 with a higher value representing a higher confidence or willingness. Those two values determine whether an individual is silent or not, which is represented in the variables by a Boolean. An individual becomes silent if their confidence falls below their willingness to self-censor. Individuals can be connected by links, which have no properties of their own and merely represent a communicative connection between the two agents. We add an additional Boolean flag to show whether an agent is a *hater* or not, since the differences between *haters* and non-haters will be implemented in the procedures of the simulation rather than in a separate entity type. Only agents with negative opinion can be *haters*. Similarly, to implement the boosting of positive opinions, we will flag certain agents with positive opinion as *paragons*, who will be treated differently by the procedures as well. A simulation will stop when the number of silent agents doesn't change between two consecutive time steps or after 30 time steps in case a stable loop has been created in the model since 30 time steps are usually far more then necessary for the simulation to end or reach such a loop.

Basic principles – The model is based on the spiral of silence and a modified implementation of the virus on a network model.

Emergence – The spiral of silence and more specifically the effect of hate speech and measures against hate speech will be represented on the micro level.

Adaptation – The only decision agents can make in the baseline model is whether they become silent, which entirely depends on their confidence. The confidence is calculated based on the ratio of identical and opposite opinions among other agents with whom the agent itself is connected.

Objectives – Agents do not have a goal per se that they can try to achieve as the simulation focuses on the effects of the other agents on their behavior and most decision points are abstracted in the form of random events.

Learning – The general decision-making process will remain static over the course of the simulation with only the agent's confidence changing. The impact of confidence as well as the value it is compared to however do remain static.

Prediction – Agents do not have any prediction systems and do not plan their behavior as the simulation focuses on the impact of external forces on the individual.

Sensing – Agents perceive the opinion of any other agent they are in contact with and derive their own confidence from this information.

Interaction – Agents indirectly interact with their neighbors as they impact the confidence of their neighbors and vice versa.

Stochasticity – During the setup, randomness is used to determine the opinion of agents, which agents are connected with each other via links, which agents will be *haters* and which *paragons*, and to determine the self-censor willingness of agents. During the simulation, randomness will be used to abstract the decision of blocking a *hater*, establishment of additional connections of *paragons*, requesting a ban of a *hater*, and whether such a ban request succeeds, as the underlying processes would be far too complex to model in detail and aren't directly relevant to our study.

Collectives – Agents do not directly form groups, but the spiral of silence itself tends to create clusters of agents with the same opinion which are isolated from each other by silent agents.

Observation – The development of the number of silent agents, as well as the number of silent agents belonging to each opinion can be observed, which in turn allows the observation of the spiral of silence on the two groups, and which group is being silenced. Changes in those numbers allow the comparison of different parameters and the evaluation of the effectiveness of the three measures against hate speech that are the focus of our study.

Agent-Based Modeling – Exercise 4

Group 4: Alexander Helwig, Qingbo Qiao, Prathibha Rajapaksha

Task 1: Document your implementation process! Describe how you have transformed the concept into an executable model and which components you have implemented in which way (max. 1200 words).

At time t = 0, the model consists of n randomly placed agents connected by links which have been created by iteratively connecting a random agent with its closest not yet directly connected neighbor until the desired number of links, $\frac{n}{2} \cdot \overline{\deg(ag)}$ with $\overline{\deg(ag)}$ being the average node degree has been reached. During agent creation, their confidence is set to 1, their self-censor willingness to a random number between 0 and 1 just as in the baseline Spiral-of-Silence-model, and all of their flags, which are used to show that an agent is silent, a hater, a paragon, is subject to a ban-request, and has been banned respectively, are set to false by default. Of all agents, $n \cdot P(-)$ are assigned the opinion of -1 (negative) and the rest are assigned the opinion of $+1$ (positive). h agents with negative opinion are designated as haters and similarly p agents with positive opinion are designated as paragons, though the latter number can be zero and both numbers cannot be greater than 20% of all agents with the respective opinion. n, $\overline{\deg(ag)}$, $P(-)$, h and p are all independent variables of the model and can be chosen by the user. Agents also have a color based on their status following the following hierarchy $silent > hater, paragon > opinion \pm 1$, though this has no effect on the simulation itself and merely serves as a visualization of the process.

Each time step (tick) consists of five major steps and the incrementation of the tick counter between step four and five, since step five requires the incremented time step. Steps two and three are conducted on a per-agent basis with $ask\ turtles\ [\dots]$, and steps one and four are skipped if the corresponding intervention has not been activated.

The first step represents blocking and checks all links between an active hater and an agent with positive opinion. For each of those links, a random number between 0 and 1 is generated and, if it is lower than the block-chance, the corresponding link is removed from the network, representing the non-hater blocking the hater. This is done first so that blocked haters will not affect the calculation of confidence in the next step.

The second step consists of the confidence update for all agents in the network. A banned agent automatically is assigned a confidence of 0, while for all other turtles, the confidence change δ based on the ratio of opinions in the agent's link neighborhood is calculated first, which is done in the same way as in the original model except that hater-agents can be weighted more heavily in the calculation with each hater being counted as $1 + hater\_influence\_boost$ agents. δ is then added to the current confidence of the agent and the result, or 0, whichever is greater) is then used in the sigmoid function just like in the original model in order to calculate the new confidence. Afterwards, haters might receive a flat bonus to their confidence to represent their

greater extroversion and resilience to outward pressure. If paragons are active as an intervention, they can also receive a bonus that works the same. Both bonuses are parameters that can be set before the simulation starts. One additional procedure might be called during the second step before the opinion ratio in an agent's neighborhood is calculated, but only if banning is enabled: If the currently selected agent has a positive opinion, a random number between 0 and 1 will be generated for each of its link-neighbors that is an active hater. If that number is below the chance of a ban being requested, the ban-pending flag of the respective hater will be set to true. This represents a ban being requested by the currently selected agent.

In the third step, the newly determined confidence is compared with the self-censor willingness of the agent. If the confidence is lower than the willingness, the agent becomes silent, otherwise it remains or becomes active. If banning is active as an intervention, previously requested bans are now processed. To this end, for each agent with ban-pending = true, a random number between 0 and 1 is generated and compared to the ban-success-rate to represent a more or less lax ban-policy or moderation team independent from the actual likelihood of a user requesting a ban. If the number is below the success rate, all links of the current agent are removed, and it is set as both silent and banned, the latter of which prevents its silent status from changing. In all cases, the ban-pending flag is then set to false to prevent unnecessary function calls in the case of a successful ban or additional ban attempts from an already rejected ban request. Finally, the agent's color is updated to visualize any potential status change.

The fourth step only occurs if paragons are active as an intervention and paragon link growth has been enabled by selecting a link-chance greater than 0. The step is used to represent the reach boosting they receive from the network itself by attempting to add additional links to all paragons. Here, a random number between 0 and 1 is generated and compared to the link-chance. If it is below said chance, new links are created, if possible, in the same way as during network setup, except that all new links originate from the currently selected paragon. The number of links created can also be specified within the range of 1 and 5.

Finally, in step five, it is checked whether the simulation should terminate to prevent unnecessary calculations and to preserve legible graphs of the network development. Simulations terminate when both the number of silent agents with positive opinion and the number of silent agents with negative opinion have remained unchanged for two consecutive time steps as this shows stagnation in the network. During 75 test runs without termination, the numbers mostly remained static after a single unchanged step and never started changing again after two consecutive steps without change, hence this criterion. Sometimes however, a stable loop emerges where the number of silent agents oscillates between two values. For this reason, the simulation also terminates after time step 30, as during no test run changes occurred after time step 20, aside from the aforementioned oscillations.

Overview of all parameters, their value ranges and step sizes (parameters that activate an intervention in **bold**):

Parameter	Minimum	Maximum	Step Size
Number of nodes n	50	300	50
Average node degree $\overline{\deg(\mathrm{ag})}$	5	15	5
Negative opinion ratio $P(-)$	0.1	0.5	0.05
Hater count h	1	$\frac{n \cdot P(-)}{5}$	1
Hater confidence boost	0.1	0.7	0.1
Hater influence boost	0.5	2	0.5
Paragon count p	0	h	1
Paragon confidence boost	0.1	0.7	0.1
Paragon link chance	0	1	0.2
Paragon link count	1	5	1
Blocking chance	0	0.4	0.05
Ban request chance	0	0.2	0.05
Ban success chance	0.3	1	0.1

Task 2: Perform a complete verification of your model. Create a documentation during the verification (max. 800 words)!

The goal of validating the NetLogo simulation is to ensure that the model behaves as expected and adjusting the implementation if required.

Test Case 1: Basic Network Initialization

Parameters: $n = 50$

$\overline{\deg(\text{ag})} = 5$

$P(-) = 0.1$

Expected Result: Network with 50 nodes with average node degree of 5 created.

Actual Result: Network initialized correctly.

Test Case 2: Hater Influence

Parameters: $n = 100$

$\overline{\deg(\text{ag})} = 10$

$P(-) = 0.3$

$h = 6$

$hater\ influence\ boost = 0.7$

Expected Result: Haters influence opinion dynamics significantly.

Actual Result: Haters influenced opinions as expected.

Test Case 3: Paragon Influence

Parameters: $n = 200$

$\overline{\deg(\text{ag})} = 10$

$P(-) = 0.4$

$h = 16$

$p = 4$

$paragon\ influence\ boost = 0.7$

Expected Result: Paragons foster positive opinions effectively.

Actual Result: Paragons fostered opinions as expected.

Test Case 4: Banning

Parameters: $n = 150$

$\overline{\deg(\text{ag})} = 7$

$P(-) = 0.2$

$h = 6$

$ban\ request\ chance = 0.1$

$ban\ success\ chance = 0.5$

Expected Result: Ban mechanisms function correctly.

Actual Result: Banning worked as expected.

In all cases, the simulation did not exceed 30 ticks as planned hence the simulation terminates as we expected. All these verification tests confirmed that our model works as intended. These tests evaluated different aspects and no adjustments were necessary.

Agent-Based Modeling – Exercise 5

Group 4: Alexander Helwig, Qingbo Qiao, Prathibha Rajapaksha

Task 1: Calibration

Our model has thirteen parameters, of which seven are part of the implemented interventions and therefore not subject to the calibration. This leaves the number of nodes, the average node degree, the percentage of agents with a negative opinion, the number of haters, the bonus haters get during their own confidence calculation and the additional weight haters get during the confidence calculation of other agents.

The node degree and negative opinion ratio are both part of our experimentation and therefore also not subject to the calibration. During the calibration, we used the same node degrees as during our actual tests (5, 10 and 15), and used example ratios of 50%, 33% and 25%. For the number of nodes, we opted for a static 300 as this was recommended as the default in the lectures.

We initially tried to calibrate the number of haters and their confidence and influence boosts based on the strength of the spiral of silence effect but couldn't find any real world data suitable for this optimization. We therefore decided to change our approach by adding the same link-adding method used by paragons in one of our interventions and applying it haters as well to simulate their growing reach and the increasing attention they get. This added two new parameters, the chance that a hater gets additional links in any given tick and the number of links added per successful tick.

This allowed us to use the number of agents that have come into contact with hate speech as our optimization goal to determine those two new values as well as the number of haters. We used 50% as our optimization goal as multiple sources report values between 48% and 54% (cf. Emmer et al. 2021, Schaetz et al. 2020, Thomas et al. 2021).

We limited the number of haters to a minimum of one, as the model is supposed to simulate haters which requires the presence of at least one, and a maximum of fifteen as haters are generally quite active posters and data suggests that only 10% of users participate at least moderately in online discussions (Palekar et al. 2015). We decided to half this to 5% because not every moderately active user is a hater. We limited the number of links to be added to the range from 1 to 5 (both inclusive) as those seemed reasonable for our given network size and node degrees. Finally, we limited the link chance to a minimum of 5% to make sure that the calibration doesn't deactivate the adding process.

We calibrated for each of the node degrees separately, in each case averaging the results over the three negative opinion ratios. We used both the general genetic algorithm (GA) and the simulated annealing (SA) algorithms provided by BehaviorSearch.

The results for the number of links to be added were very conclusive, as only three searches – one in each node degree – yielded a value other than 1. The results for the link chance and hater count on the other hand were less conclusive as both showed significant spread and a comparatively high standard deviation.

Therefore, we decided to once more calibrate within each node degree, but this time with the number of links set to 1 and the range for the hater count and link chance limited to the interval centered around their mean during the first calibration $\pm$ the respective standard deviation. As the SA-algorithm had performed significantly better and was less resource intensive, we only used this algorithm for the second round of calibration.

The calibration resulted in the following values being chosen:

Node Degree	Hater Count	Link Chance	Link Count
5	12	0.21	1
10	13	0.11	1
15	11	0.06	1

With no way to calibrate the confidence boost and influence boost for haters coming to mind as their impact on the spread of hate-speech is negligible at best, we had to choose arbitrary values for those two parameters, selecting 0.2 for the confidence boost and 1 for the influence boost, effectively making every hater count as two non-hater agents with negative opinion.

As our model showed good fitness over all three node degrees (average fitness: 2,57), we concluded that no further changes were necessary.

Data Sources:

Emmer, M., Leißner, L., Porten-Cheé, P., & Schaetz, N. (2021). Weizenbaum Report 2021: Politische Partizipation in Deutschland.

Palekar, S., Atapattu, M. R., Sedera, D., & Lokuge, S. (2018). Exploring spiral of silence in digital social networking spaces. In *International Conference on Information Systems (ICIS 2015): Exploring the Information Frontier*. Association for Information Systems (AIS).

Schaetz, N., Leißner, L., Porten-Cheé, P., Emmer, M., & Strippel, C. (2020). Politische Partizipation in Deutschland 2019.

Thomas, K., Akhawe, D., Bailey, M., Boneh, D., Bursztein, E., Consolvo, S., Dell, N., Durumeric, Z., Kelley, P. G., Kumar, D., McCoy, D., Meiklejohn, S., Ristenpart, T., & Stringhini, G. (2021). SoK: Hate, harassment, and the changing landscape of online abuse. *2021 IEEE Symposium on Security and Privacy (SP)*. https://doi.org/10.1109/sp40001.2021.00028

Task 2: Validation

First, we performed face validation on the structural level for all node degrees and several negative opinion ratios with the settings derived from the calibration. We kept all interventions off as they are not part of the validation process in this case.

We started with an animation assessment which showed that the model performs as expected: Over the course of the simulation, clusters of connected agents with the same opinion that remain outspoken emerge, whereas comparatively isolated or surrounded agents in fall silent. The number and size of positive clusters increases with a decreasing negative opinion ratio, as positive agents are more likely to find positive agents in their vicinity. Similarly, the number of negative clusters decreases. This effect is reinforced by a higher node degree. Haters are more resilient to being silenced and can keep negative groups active or even revitalize them if a new connection is created between them and a previously silent agent that then becomes active again. Overall, this matches our expectations regarding real world behavior.

Next, we performed an output assessment, in which we found that in most cases, the prevalence of hate speech reaches around 50% as desired during calibration. The spiral of silence effect also occurs as desired, as hater activity can shift the perceived majority opinion causing the actually dominant positive opinion to fall increasingly silent. A small sized immersive assessment of five agents during different model runs shows that agents correctly decide when to become or remain silent and when to become or remain active.

Second, we performed empirical validation on the behavioral level by bootstrapping with random seeds. For this validation, we considered each node degree separately as they were calibrated separately as well.

For each node degree, we ran 30 simulation runs for the validation – 10 for each of the opinion ratios used during calibration – and gathered the percentage of agents that had been in contact with hate speech as with the calibration. We then calculated the mean, standard deviation and mean squared error (MSE) for our gathered data (see table below). We also calculated the 99% and 99.9% confidence intervals for each of the three node degrees. In all six cases, the 50% target lies within the confidence interval (see table below), therefore we consider our validation successful and deem no changes necessary.

Node Degree	5	10	15
Mean	46.633	50.911	50.711
Standard Deviation	9.002	6.51	6.352
MSE	92.367	43.207	40.859
99% Interval	42.59–50.68	47.99–53.84	47.86–53.57
99.9% Interval	41.05–52.21	46.87–54.95	46.77–54.65

Agent-Based Modeling – Exercise 6

Group 4: Alexander Helwig, Qingbo Qiao, Prathibha Rajapaksha

Task 1: In order to test your hypotheses, perform experiments with your validated model and create a documentation (max. 1800 words)!

All three of our experimental setups have many parameter settings in common: All of them use three different node degrees, 5, 10 and 15, to simulate networks of different interconnectivity and to limit the effect of this factor on our results. We chose two different settings for the ratio of agents with a negative opinion, 0.3 and 0.45, to reduce the impact of this value as well. The number of haters, chance of haters gaining new connections, and number of new connections per time step were determined during calibration. The number of agents was set at 300 as recommended during the lectures. The confidence boost and influence boost for haters were set at 0.2 and 1 respectively to represent the difference in behavior between haters and non-hater agents. To ensure otherwise equal circumstances for hypothesis 3, which compares all three of our interventions, we used the same seeds for each combination of parameter settings. To this end, we used the seeds from 400 to 429, which were chosen arbitrarily except that we ensured that no seed used during calibration was used again. In each simulation run at most one intervention was active, with the others being disabled by setting the blocking chance, paragon count or ban request chance to 0 respectively. The parameters for the interventions themselves are described below. 30 repetitions were run with NetLogo's BehaviorSpace tool for each distinct combination of values.

Hypothesis 1: *Isolating Haters via blocking is an effective way to combat the effect of hate speech on a network and significantly reduces the silencing effect of the Spiral of Silence*.

As blocking is controlled by only one parameter, the chance that a positive agent blocks a hater it is connected to after interacting with it, this naturally was the only value to vary during experimentation. We chose three different values to represent this, 10%, 20% and 30% as real world numbers for actual blocking vary significantly, though we opted to choose the lower end of the spectrum so as not to overrepresent the effect of blocking. We also included a blank run with a 0% block chance in order to be able to determine if blocking actually has an effect on the visible opinion ratio.

After obtaining the data from BehaviorSpace tool, we analyzed the data with the help of the statistics software RStudio. We used this program to check for significant differences in the means of the different simulation runs with the same settings for node degree and opinion ratio but different blocking chances, always comparing with the 0% baseline. If possible, we used an unpaired Student's t-Test after testing for both of its requirements, the normal distribution of data with a Shapiro-Wilk test and the equality of variances with an F test. If equality of variances was violated, we used the Welch approximation to the degrees of freedom. If the criterion of normal distribution was violated, we resorted to a non-parametric Wilcoxon test. To maintain

comparability within a series of simulations differentiated by blocking chance only, we used the same test for all comparisons, choosing the best test applicable to all three comparisons.

Test	ND05_R30_B00	ND05_R30_B10	ND05_R30_B20	ND05_R30_B30
Mean	0,2524	0,2384	0,2249	0,2147
Median	0,2572	0,2297	0,2137	0,2046
Deviation	0,0834	0,0864	0,0823	0,0786
Wilkes-P	0,6616	0,8851	0,9526	0,9667
Var-P	Not Applicable	0,8516	0,9422	0,7469
t-Test-P	Not Applicable	0,5258	0,2036	0,0766

Test	ND05_R45_B00	ND05_R45_B10	ND05_R45_B20	ND05_R45_B30
Mean	0,6237	0,6160	0,5965	0,5784
Median	0,6402	0,6368	0,6227	0,5895
Deviation	0,1023	0,0933	0,0953	0,0980
Wilkes-P	0,6713	0,0538	0,0606	0,5426
Var-P	Not Applicable	0,6217	0,7057	0,8205
t-Test-P	Not Applicable	0,7609	0,2905	0,0853

Test	ND10_R30_B00	ND10_R30_B10	ND10_R30_B20	ND10_R30_B30
Mean	0,1501	0,1241	0,1178	0,1099
Median	0,1330	0,1103	0,1175	0,1107
Deviation	0,0847	0,0639	0,0523	0,0561
Wilkes-P	0,3210	0,2646	0,8768	0,4086
Var-P	Not Applicable	0,1341	**0,0116**	**0,0303**
Welch-P	Not Applicable	0,1853	0,0820	**0,0351**

Test	ND10_R45_B00	ND10_R45_B10	ND10_R45_B20	ND10_R45_B30
Mean	0,4850	0,4603	0,4563	0,4427
Median	0,4764	0,4459	0,4435	0,4248
Deviation	0,1181	0,1190	0,1154	0,1099
Wilkes-P	0,3247	0,6330	0,7127	0,5729
Var-P	Not Applicable	0,9705	0,8993	0,7001
t-Test-P	Not Applicable	0,4241	0,3449	0,1570

Test	ND15_R30_B00	ND15_R30_B10	ND15_R30_B20	ND15_R30_B30
Mean	0,0653	0,0555	0,0528	0,0501
Median	0,0651	0,0578	0,0545	0,0504
Deviation	0,0447	0,0357	0,0358	0,0322
Wilkes-P	0,1724	0,3406	0,1260	0,1748
Var-P	Not Applicable	0,2364	0,2402	0,0840
t-Test-P	Not Applicable	0,3552	0,2368	0,1355

Test	ND15_R45_B00	ND15_R45_B10	ND15_R45_B20	ND15_R45_B30
Mean	0,4422	0,4265	0,4088	0,3882
Median	0,4411	0,4193	0,4084	0,3686
Deviation	0,1137	0,1186	0,1131	0,1110
Wilkes-P	0,6144	0,1217	0,3086	**0,0238**
Var-P	Not Applicable	/	/	/
Wilcox-P	Not Applicable	0,5298	0,2643	**0,0476**

Fig 1: Results of the statistical evaluation. Test results leading to the rejection the H_0-hypothesis shown in bold. ND refers to the node degree, R to the opinion ratio and B to the blocking chance.

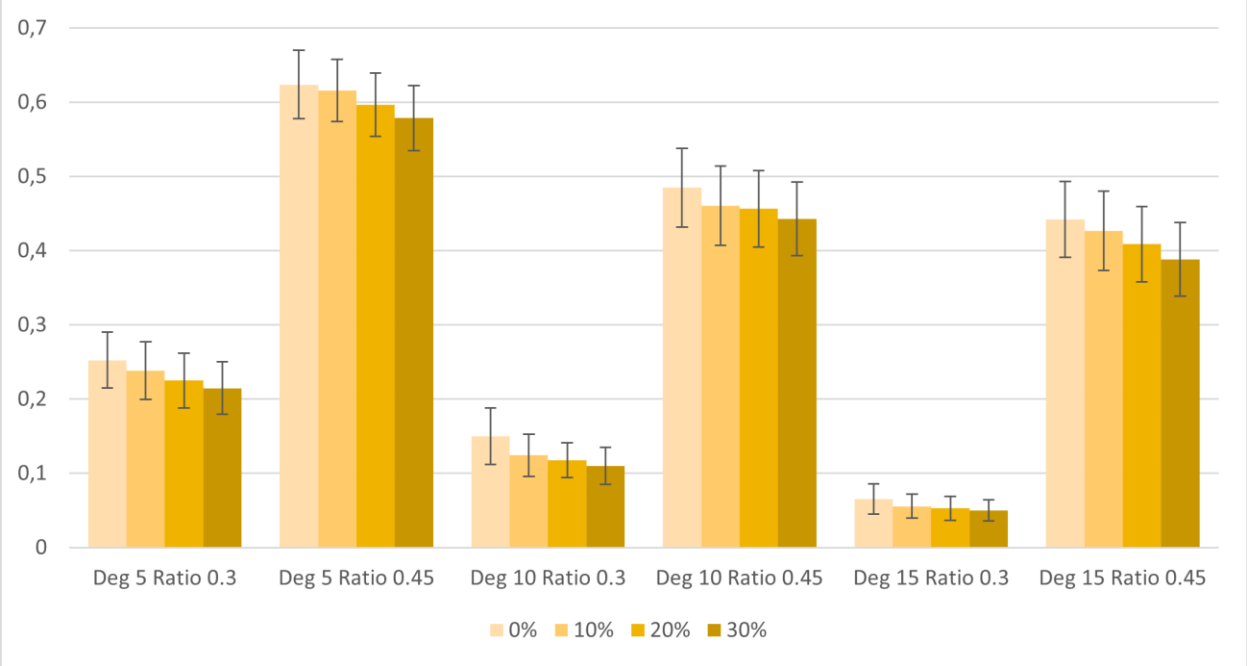


Fig 2: Bar graph showing the means of the simulation runs grouped by parameter settings. Whiskers indicate a 99% confidence interval based on standard deviation.

The graphical representation of our results shows several effects (see Fig. 2). The visible opinion ratio depends on the starting opinion ratio which is logical since more negative agents are present and therefore less likely to become silent. In fact, all runs with a starting ratio of 0.3 result in a

visible opinion ratio below 0.3, meaning that the negative agents were silenced by the majority. The stronger negative minority with opinion ratio 0.45 however is frequently overrepresented, reaching above 0.6 in two cases, demonstrating the Spiral of Silence effect. The average node degree also plays a great role, as the visible opinion ratio decreases with an increase in node degree, as isolated positive agents surrounded by a local majority of negative agents become rarer in the overall network with increased interconnectivity.

The statistical analysis with RStudio shows that only two runs significantly ($\alpha = 5\%$) differ from the blank runs with 0%, in both cases runs with a block chance of 30%. While this suggests that blocking is ineffective as a method to combat the spread of hate speech, it is notable that there is an overall downward trend with increasing block chance. This could mean that blocking is effective, but our chosen values simply were too low to demonstrate this effect, which in turn means that a high number of actively blocking users would be required in the real world to have a meaningful impact, perhaps too high to be realistic. Overall, we have to accept the h_0-hypothesis that blocking is not effective at stopping the Spiral of Silence.

It should be mentioned that in the scope of our experiments, blocking was limited to hater agents only, which means that normal negative agents are unimpeded in their influence. Therefore, it is a possibility that blocking the most active users alone isn't sufficient and that one should also block those who merely agree with them, though this definitely requires further testing and simulation.

Hypothesis 2: *The effectiveness of promoting the spread of positive opinion is significantly more effective in networks with a low degree of interconnectivity (node degree) than in networks with high interconnectivity.*

The experimental setup for this hypothesis has three additional parameters, the number of paragons, the chance for paragons to gain additional links and the number of links gained per time step. For the latter two, we selected the same values that haters used depending on the node degree and determined during calibration. For the number of paragons, we chose two different values for each set up, the same number as there are haters or half as many paragons as there are haters, rounded up.

While we ran the simulation for all three node degrees, only degrees 5 and 15 were considered for this hypothesis as the difference between node degrees is the focus of the experiment and those two have the largest difference between them.

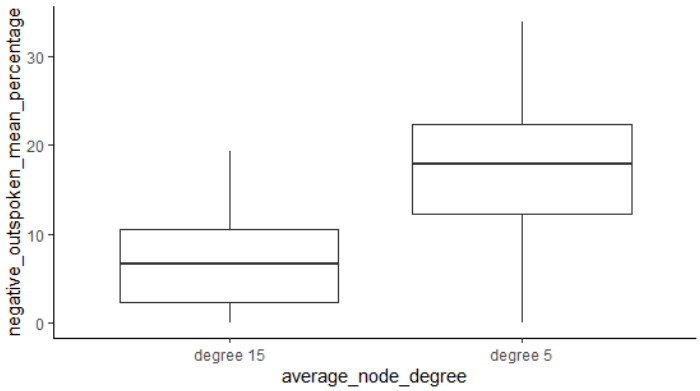


Fig. 3: Boxplot of the visible opinion ratio based on the node degree with opinion ratio 0.3

Figure 3 shows that there is a significant difference with node degree 15 having a mean of 7, whereas the node degree of 5 has a mean of 17. Similarly to hypothesis 1, we conducted a two-sided unpaired t-Test after confirming normal distribution with the Shapiro-Wilk test. The t-Test yielded a p-value of 0.000001358, vastly below the threshold of 0.05. Therefore, we reject the h_0-hypothesis and can infer that the true means are most likely different.

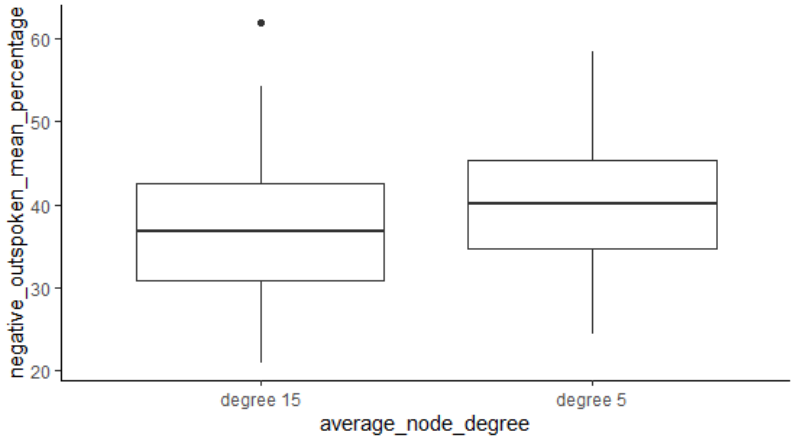


Fig. 4: Boxplot of the visible opinion ratio based on node degree with opinion ratio 0.45. Note different y-axis scale compared to Fig. 3

With a higher starting opinion ratio of 0.45, the differences between the two node degrees were far less pronounced. Node degree 15 had a mean of 37, whereas node degree 5 has a mean of 40. We conducted the same tests as with opinion ratio 0.3 and calculated a p-value of 0.1372, above the α of 0.05. Hence we accept the h_0-hypothesis, meaning there is no significant difference between the two true means.

Overall, the results from this experiment are inconclusive and necessitate more detailed testing. A visual comparison with the results from the first experiment (Fig. 2) implies that having paragons can reduce the impact of haters as the means from this experiment are almost always lower than the one from the first experiment, with the sole exception being the combination of node degree 15 and opinion ratio 0.3, where the odds might have been stacked against the

negative opinion so much in the general premise that the presence of paragons cannot realistically reduce the spread of hate speech in a meaningful amount. Though this also requires future testing.

Hypothesis 3: *Moderation and banning as an approach are significantly more effective at reducing the impact of haters than the other two approaches examined in our model under otherwise equal circumstances.*

To test this hypothesis, we used the results from the previous two experiments and compared them to the results from our third intervention, banning. This mechanism has two parameters that we varied, the chance of a user to request a ban, for which we chose the values 0.05 and 0.1 as requesting a ban is likely to be performed by fewer users than blocking since the effort is far greater, and the chance of a ban request to be acted upon to represent different effectiveness levels of the moderation team in the network. Here we chose 0.5 and 0.8 as values to get a reasonable spread of effectiveness. Note that all ban requests made in a single time step are treated as a single ban request for the determination of ban success or failure.

Statistically, we once conducted the same tests and pre-tests as in experiment 1 to check for a significant difference in means. Additionally, we calculated the directionality of the difference, as our hypothesis not only states that the two are different, but also that banning performs better than the other two methods. Comparisons were made only between simulations series that had the same parameters aside from the intervention settings.

Comparing blocking and banning, we found that all tests yield p-values far below the threshold, with the lowest being 4.12E-18 and the highest being 0.000377. We therefore reject the h_0-hypothesis that the means are equal. In addition, in all cases the means of the banning series were lower than the ones from blocking, showing that banning is more effective than blocking.

Similarly, when comparing paragons and banning, all tests yielded very low p-values, with the lowest here being 1.33E-9 and the highest being 0.00246. Once again, we reject the null hypothesis and just as with blocking and banning, the means of the banning series were lower than those of the paragon series. We can thus conclude that banning is also more effective than paragons.

All in all, we were able to determine that blocking, at least at realistically low rates, is not sufficient on its own to stop haters. Our results for paragons were less conclusive and need further testing to properly evaluate the effectiveness of paragons. Finally, banning as an intervention is far more effective than the other two methods, though given their ineffectiveness, additional testing is necessary to ensure that banning in and of itself is effective, though based on mere visual data inspection the results seem promising.

Reflection:

We create a model for hate speech behavior based on the features and logic of spiral of silence. Model display how agents behave on each other when hate speech encounter. In order make it realistic as possible we made 15 parameters. With those parameters we try model few simulations on reducing or eliminating the impact of the hate speech. We try to catch most of the psychological aspect of the real world.

NetLogo is a highly user-friendly software that facilitates programming, calibration, validation, simulation, visualization, and other functionalities. This tool has been instrumental in helping me grasp the concept of agent-based modeling and how to create near-real-world agents.

Following the model execution, we moved on to statistical analysis. Although not typically part of agent-based modeling, this step was included in the final assignment to provide a comprehensive understanding of tackling real-world problems. We learned how to conduct statistical and mathematical analyses for a data-driven experiment, for the initially defined three hypotheses.

There were a few aspects of the model that could have been improved. For instance, agents who were blocked could be allowed to unblock over time, or the model could differentiate between an agent's public opinion and their actual opinion.

I appreciated the way that course was designed, where we first created a flowchart in the initial assignment, followed by an introduction to the spiral of silence network. After that, we explored hate speech before being guided on implementing a conceptual model, calibration, and verification, ultimately leading to model validation and simulation. Receiving feedback from the lecturer and fellow students throughout the course was incredibly helpful in developing the model.

Finally, I would like to express my gratitude to the teaching unit of the Agent-Based Modeling course, especially to Christian Lohr, whose valuable advice helped improve the model and guided us throughout the course. I also want to thank my group mate, Alexander Helwig, for enhancing the quality of our solution and model.