LVA-Modeling and Simulation

Fake News

Giorgio Bertone, 12402238 - UE 066 645 *
Klodian Dervishi, 12402532 - UE 066 645 †
Bingliang Song, 51831953 - UE 066 453 ‡
Torres Antonio, - §

February 9, 2025
Supervisor: Daniele Giannandrea

Belief formation in networked societies is shaped by convincing power, initial distribution of beliefs, troll activity, and social connectivity. Our agent-based simulation reveals that higher convincing power accelerates consensus, while trolls induce polarization and misinformation dominance. Broader social interactions reduce fragmentation, but localized negative beliefs persist if unchallenged. Effective misinformation control requires trusted sources, shadowbanning trolls, and community-driven fact-checking. These insights emphasize the need for multilayered strategies to enhance belief stability and resilience to misinformation.

*Implemented the full code for Task 1, Task 2 and Task 4 and wrote Chapter 1, 2.1, 2.2, 2.3, 3.1, 3.2. Oversaw Chapter 2.4. Made plots of Chapter 3.3.2. Contributed to Result interpretation

†Implemented the full code for Task 3. Wrote Chapter 3.3. Contributed to Result interpretation

‡Analyzed and checked the results and code of Task 4, wrote the abstract, Chapter 2.4, 3.4, summary and conclusion in Chapter 4. Contributed to Result interpretation

§Wrote Outlook in Chapter 4. Made presentation layout

Contents

1. Introduction **3**

2. Model **3**

2.1. Modelling . . . 3
2.2. Parametrization . . . 4
2.3. Implementation . . . 5
2.4. Troll Countermeasures . . . 5
2.4.1. Trusted Agents . . . 5
2.4.2. Reporting and Shadow Banning . . . 6
2.4.3. Community Notes . . . 6

3. Simulation Results **6**

3.1. Basic Model . . . 6
3.2. Modeling Troll Impact . . . 7
3.3. Parameter and Neighborhood Variations . . . 8
3.3.1. Increasing the Number of Neighbors . . . 8
3.3.2. Different values for c_{max} and p_+ . . . 9
3.3.3. Concentration of Negative Beliefs . . . 11
3.4. Troll Countermeasures . . . 12

4. Discussion **14**

A. Appendix **17**

A.1. Acknowledgments . . . 17
A.2. Code Snippets . . . 17
A.3. Results for Changed Initial Parameters . . . 19
A.3.1. Without Trolls . . . 19
A.3.2. With Trolls . . . 19

1. Introduction

Motivation. The rapid dissemination of false information, particularly the intentional distribution of fake news — such as through troll farms — has become a major issue in the digital era. The ability of misinformation to shape public perception and influence societal discourse highlights the urgent need to understand its spread.

Introduction to the Topic. Opinion dynamics, which examines how individuals adjust their views based on interactions with others, provides a useful approach to modeling the propagation of fake news. While such models are often simplified representations of reality, they help analyze how misinformation circulates within networks, whether on social media platforms or in society at large.

Reated Work. Several studies have explored agent-based modeling (ABM) as a tool for understanding and mitigating the spread of misinformation in online and offline networks. Gausen et al. (2021) developed an ABM to simulate information dissemination on Twitter, using real data to validate the model and assess countermeasures before implementation. Tseng and Son Nguyen (2020) applied ABM with Social Impact Theory to analyze workplace rumor propagation, finding that improving work environments and management counseling can reduce misinformation. Alassad et al. (2023) combined ABM with organizational cybernetics to monitor misinformation spread in online and offline networks, testing the approach on Twitter data from a COVID-19 protest. These studies demonstrate ABM's effectiveness in modeling misinformation dynamics and evaluating intervention strategies.

Aim. This report explores how agent-based modeling can be applied to study the spread of fake news. By simulating interactions between individuals in a network, we aim to gain insights into the mechanisms driving misinformation dissemination and potential strategies to mitigate its impact.

2. Model

2.1. Modelling

The proposed agent-based model (ABM) simulates the dynamics of belief propagation and the spread of misinformation within a population of N individuals (agents). These agents are positioned on a rectangular grid, where each agent is connected to a set of neighbours through a neighbourhood structure, such as Moore or von Neumann, and the grid employs periodic boundary conditions.

Each agent i holds a belief $b_i \in [-1, 1]$, where a belief of -1 signifies acceptance of a fake news story, while 1 represents the acceptance of the fact-checked, reliable version of the story. Over time, an agent's belief can change based on the influence of their neighbours, reflecting social interactions and information sharing.

In addition to beliefs, each agent has a convincing power $c_i \in [0, c_{\max}]$, where $c_{\max} \in [0, 1]$ is the maximum persuasive influence that an agent can have over others. The value of c_i remains constant throughout the simulation, and higher values of c_i indicate that an agent has a greater capacity to influence the beliefs of others in their neighbourhood.

Initially, a fraction p_+ of the population holds beliefs drawn from the range $[0, 1]$, indicating a tendency to support the fact-checked information. The remaining agents have beliefs in the range $[-1, 0]$, indicating a predisposition to believe fake news.

At each timestep, 10% of the population is randomly selected, and their beliefs are updated based on the influence of their neighbouring agents. The belief update rule is given by:

$$b_i \to b_i + \frac{1}{n} \sum_{j \in N_i} c_j b_j,$$

where N_i is the set of neighbours of agent i, and n is the number of neighbours. The updated belief is then truncated to remain within the bounds of $[-1, 1]$, ensuring that beliefs do not exceed the extreme values. In addition to regular agents, also trolls are introduced into the model. Trolls are agents that have a fixed belief of $b_i = -1$ and this belief does not change over time. Trolls actively spread misinformation, influencing other agents within their neighborhood.

This simple model captures the main dynamics of belief change through social influence, allowing for the study of how misinformation spreads and how individuals' beliefs evolve in response to interactions with others.

2.2. Parametrization

The ABM is driven by several key parameters, which determine the dynamics of belief propagation and misinformation spread.

Parameter	**Description**
Population size (N)	Total number of agents in the simulation. A larger N allows for more complex interactions, but requires higher computational cost.
Convincing power (c_i)	Measures an agent's ability to influence others. Higher values indicate stronger persuasion, enabling the modeling of heterogeneous agents (e.g., influencers, news sources).
Neighborhood structure	Defines interaction scope and affects the diffusion of beliefs across the population.
Initial belief fraction (p_+)	Fraction of agents initially holding fact-checked beliefs ($b_i \in [0, 1]$). Determines the starting distribution of opinions.
Troll fraction (f_{troll})	Proportion of agents with a fixed negative belief ($b_i = -1$). A high fraction can disrupt consensus if unchecked.

Table 1: Description of model parameters.

By adjusting these parameters, the model can be used to explore various scenarios and the impact of different factors on the spread of fake news.

The baseline model employs the following parametrization:

Parameter	**Value**
Population size (N)	10,000
Maximum convincing power ($c_{\max}$)	0.5
Initial positive belief fraction (p_+)	0.8
Troll fraction (f_{troll})	0
Neighborhood type	von Neumann (order 1)

Table 2: Model parameters for the base simulation.

2.3. Implementation

The agent-based model is implemented in Python using object-oriented programming principles, where each agent is represented by a class with attributes defining their position, belief, influence, and role (Appendix 1). Agents interact with their neighbors in a von Neumann neighborhood, and each agent has the potential to update their belief over time based on the influence of their neighbors.

To track the dynamics of the population and the evolution of beliefs over time, we calculate the following observables:

Average Belief. The average belief across all agents at each timestep, indicating the overall trend of the population.

$$\text{Average Belief} = \frac{1}{N}\sum_{i=1}^{N} b_i$$

Fraction of Positive Belief. The proportion of agents holding positive beliefs (i.e., those who believe in fact-checked news).

$$\text{Fraction of Positive Belief} = \frac{\text{Number of agents with } b_i > 0}{N}$$

Belief Variance (Polarization). The variance of beliefs in the population, which indicates the degree of polarization.

$$\text{Belief Variance} = \frac{1}{N}\sum_{i=1}^{N}(b_i - \bar{b})^2$$

These observables were computed at each timestep and used to visualize and analyze how the population's beliefs evolve in response to the interactions between agents.

2.4. Troll Countermeasures

To counter the spread of misinformation by trolls, we have developed a series of countermeasures that progressively enhance the simulation's ability to regulate user interactions and reinforce credible sources. These countermeasures involve the introduction of trusted agents, a reporting and shadowbanning system, and a community note mechanism that collectively work to mitigate the influence of trolls.

2.4.1. Trusted Agents

The first mechanism in the simulation is the introduction of trusted agents (Appendix 2), which represent credible news sources or verified profiles. These agents serve as authoritative voices in the network, always maintaining a belief value of 1, signifying their strong alignment with factual information. Additionally, trusted agents possess a higher convincing power (a value drawn from a uniform random distribution in the range $[0, c_{max} + \frac{c_{max}}{2}]$), making them more influential than ordinary agents. This enhancement ensures that factual content has a stronger presence in the belief network, counterbalancing the influence of trolls who aim to spread misinformation.

In the initial version of the model, all agents followed the same belief update mechanisms without distinguishing between credible sources and unverified individuals. The improved version introduces the `is_trusted` attribute, which explicitly identifies trusted agents. We tested two fractions of trusted agents: 1% and 5%.

2.4.2. Reporting and Shadow Banning

While the introduction of trusted agents strengthens the credibility of information sources, it does not directly address the presence of trolls. To mitigate the impact of trolls spreading misinformation, we introduced a reporting and shadow banning system (Appendix 3). This mechanism allows users to report trolls when they detect misinformation. When an agent accumulates a certain amount of reports, they become shadowbanned, significantly reducing their influence in the network.

Shadowbanning is implemented by adding a `reports` attribute to each agent, which keeps track of the number of times they have been reported. At every timestep a fraction of users, determined by the parameter `active_fraction`, is sampled. We call these users *interactors*. The interactors report trolls with a 70% probability, and once the threshold of five reports is reached, the agent is shadowbanned. When shadowbanned, the agent's convincing power is reduced by half, making them less effective at spreading their beliefs. This mechanism ensures that the system remains self-regulating, as users collectively reduce the impact of trolls by flagging misinformation when encountered.

2.4.3. Community Notes

While shadowbanning trolls reduces their influence, it does not actively correct misinformation that may have already spread within the network. To address this, we implemented community notes (Appendix 4), which act as a fact-checking mechanism driven by collective disagreement among agents. If an agent believes in fake news (i.e., holds a negative belief), but the majority of its neighbors disagree, a community note is triggered. This mechanism encourages the agent to shift their belief toward the accurate information.

With the introduction of community notes, an additional mechanism is implemented: if an agent's belief is negative (fake news) and it contradicts its neighbors, the system actively nudges their belief toward positive values (fact-checked news) by adding a small positive amount to the agent current belief. This models the real-world effect of community notes, where exposure to fact-checked information helps counter misinformation. As a result, the spread of false narratives is gradually suppressed, making it more difficult for misinformation to persist within the network

In the final enhancement (Appendix 5), we made the community notes system more aggressive against trolls. While in the previous implementation, community notes only affected non-troll agents, in this modified version, trolls now lose an additional 30% of their convincing power whenever they are flagged by a community note. This should simulate the a limitation of visiblity or spread that could be applied by a social network to fake-news posts that received a community note. As a result, trolls gradually lose their ability to influence others, diminishing their impact over time. This final adjustment reinforces an adaptive, self-correcting information ecosystem, making it increasingly difficult for false narratives to persist.

3. Simulation Results

3.1. Basic Model

For the simulation, we employ a Monte Carlo method, running the base model 10 times, with each simulation consisting of 1000 iterations. The results obtained from these simulations are presented in Figure 1.

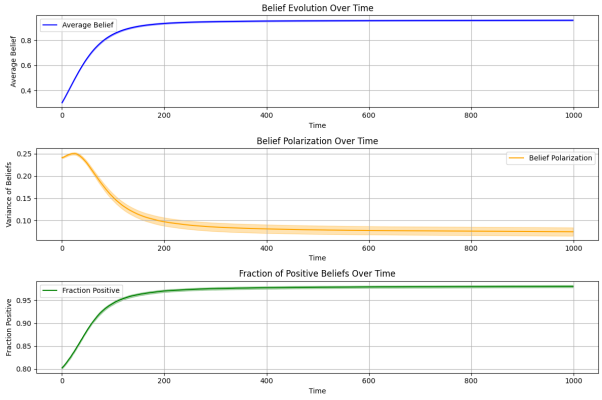


Figure 1: Global observables for base model

As we can see from the Figure 1, the average belief and fraction of positive belief quickly converge towards 1, indicating a spread in belief in fact-checked news. Also the belief polarization suggests the same, converging towards zero.

Additionally, we display the distributions of belief values at the initial and final timesteps (Figure 2).

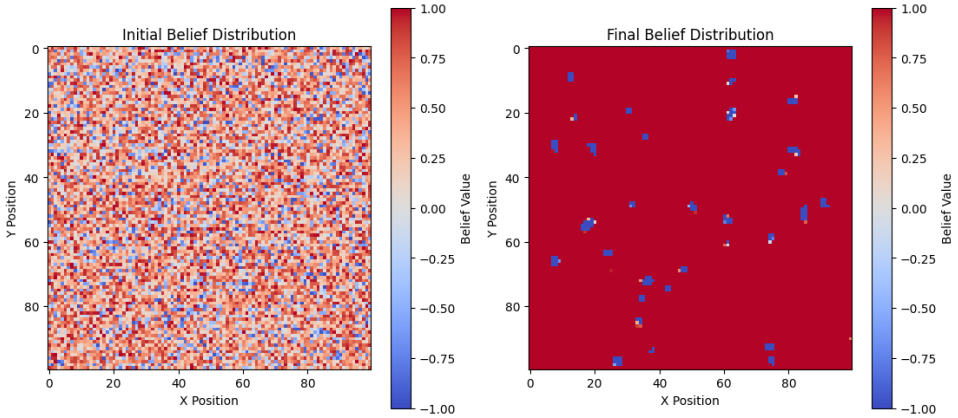


Figure 2: Distribution of initial beliefs (left) and final beliefs (right)

The initial beliefs are uniformly distributed and it is clear the impact of p_+ as the proportion of agents with positive belief appears greater. In the final belief distribution, we observe a noticeable shift toward fact-checked beliefs, indicating the spread of reliable information. However, agents with negative beliefs tend to remain clustered in small groups.

3.2. Modeling Troll Impact

To investigate the role of trolls in manipulating beliefs, we determine the critical fraction of trolls f_{troll} needed to ensure that more than half of the non-troll population (normal agents) holds a belief $b_i < 0$ after 3000 timesteps. We employ binary search to pinpoint this threshold. To ensure statistical robustness we run each simulation 10 times per value of f_{troll}. Additionally, we require the upper bound of the confidence interval to be below 0.5. Our results identify

this critical threshold at $f_{troll} \approx 0.207$ (Figure 3), beyond which negative beliefs dominate the system, preventing the spread of positive beliefs.

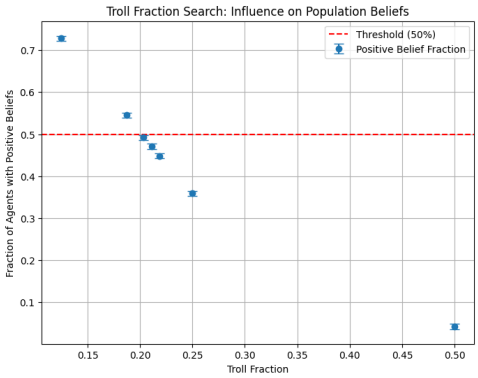


Figure 3: Fraction of Agents with Positive Beliefs as function of Troll Fraction

Next, we can run the simulation with the new found fraction and look at the results which show indeed a final fraction of positive beliefs below 0.5. (Figure 4).

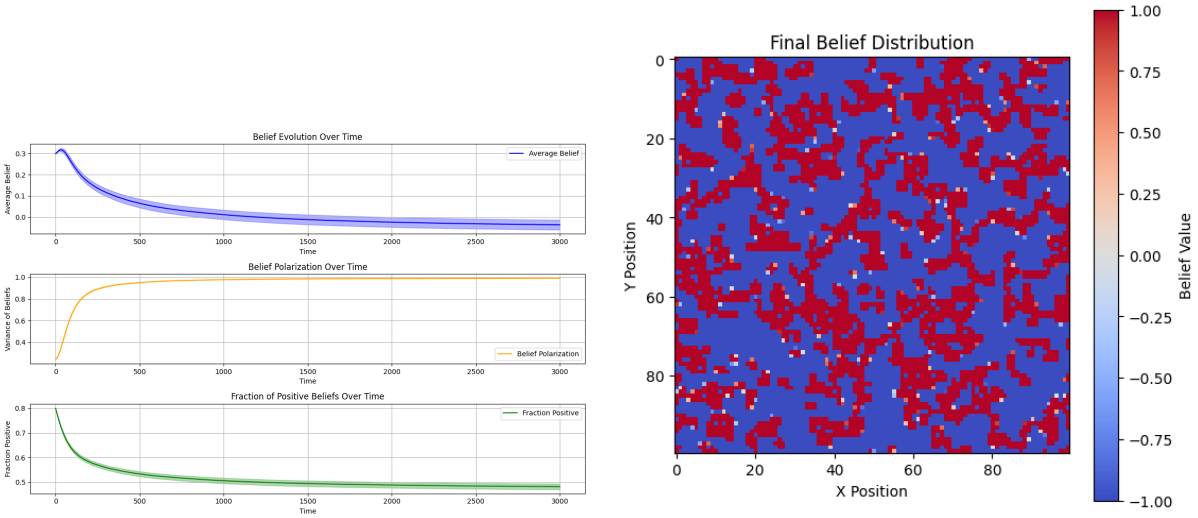


Figure 4: Results for Simulation with Critical Troll Fraction

3.3. Parameter and Neighborhood Variations

To gain a deeper understanding of the model and observe different outcomes, multiple experiments have been conducted using various implementations and parameter settings.

3.3.1. Increasing the Number of Neighbors

We analyzed how the dynamics change when increasing the number of neighbors. Specifically, we considered a Von Neumann neighborhood of order 2. This was examined in two scenarios: first, without the presence of trolls, and second, with a troll fraction of 0.1.

First, we plot the behavior of the three observables after 1000 steps, using the initial settings for $c_{\max}$ and p_+, while considering a Von Neumann neighborhood of order 2. This is done first

without the presence of trolls in the system.

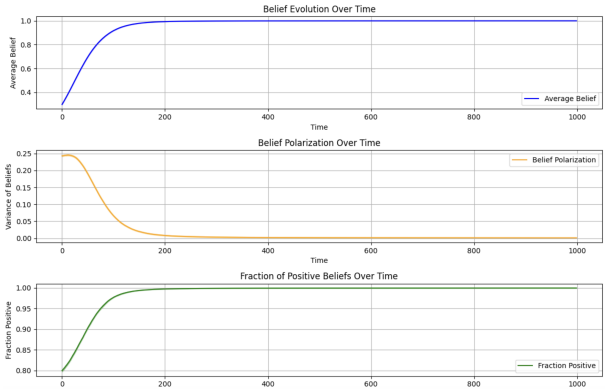


Figure 5: Behavior of the three observables after 1000 steps with initial settings for $c_{\max}$ and p_+ and a Von Neumann neighborhood of order 2 (without trolls).

Next, we present the results with the presence of trolls in the system.

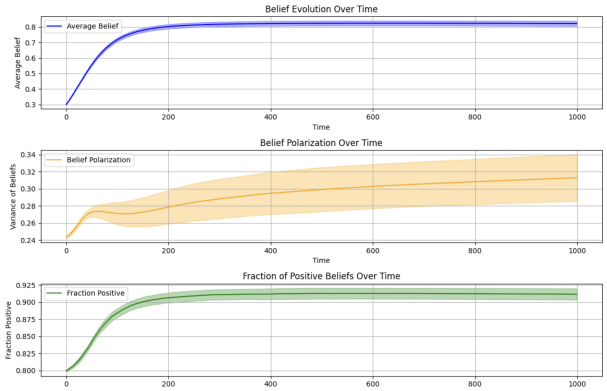


Figure 6: Behavior of the three observables after 1000 steps with initial settings for $c_{\max}$ and p_{plus} and a Von Neumann neighborhood of order 2 (with trolls).

From Figure 6, it is evident that increasing the number of neighbors has a significant effect in countering the spread of fake news. However, it is important to note that this conclusion is drawn specifically for the current parameter setting.

3.3.2. Different values for c_{max} and p_+

We conducted multiple experiments with different initial parameter settings for c_{max} and p_+ to observe their effects on the system's behavior. The values used in these experiments are shown in the following table:

c_max	0.1	0.325	0.55	0.775	1.0
p_plus	0.1	0.3	0.5	0.7	0.9

Table 3: Parameter settings for p_+ and c_{max}

Again, we performed the experiments in two stages: first, without the presence of trolls, and second, with a troll fraction of 0.1 included.

We present a summary of the observable values after 1000 time steps for various c_{max} and p_+ values, **without** the presence of trolls, in the tables provided in the appendix: the average belief table (Table 5), the polarization table (Table 6), and the positive belief fraction table (Table 7). We also display these values in Figure 7.

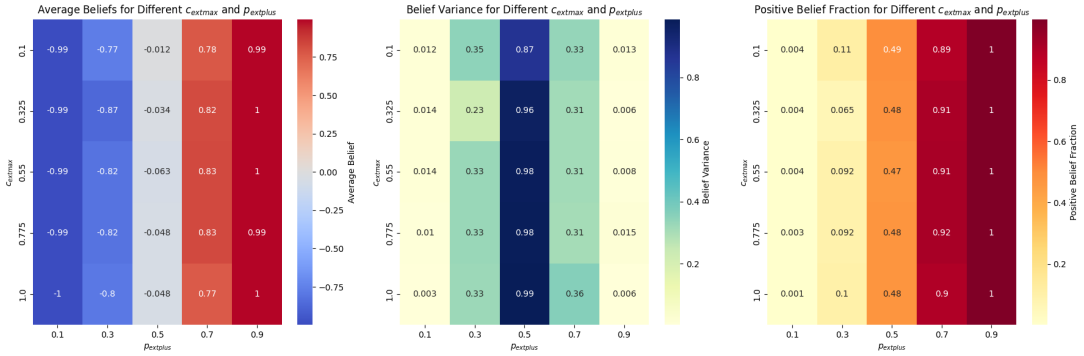


Figure 7: Heatmaps for different combinations of c_{max} and p_+ (without trolls)

From these results, it is evident that the average beliefs and the fraction of positive beliefs are highly sensitive to the initial value of p_+. Specifically, as p_+ increases, the value tends to rise as well. On the other side we observe the highest polarization, or the greatest degree of disagreement in belief, when the initial p_+ is close to 0.5 . The convincing power seems to have the most significant impact on the level of polarization, especially when the initial p_+ is near 0.5, where belief differences were already evident from the beginning.

Next, we also provide a summary of the observable values after 1000 time steps for various c_{max} and p_+ values, **with** the presence of trolls, in another set of tables provided in the appendix: the average belief table (Table 8), the polarization table (Table 9), and the positive belief fraction table (Table 10). As before we also display these values in Figure 8. The observed results follow a similar pattern as before, where the average belief and positive fraction increase as p_+ increases. However, due to the presence of trolls, both values are on average smaller. On the other hand, the variance reaches its highest values for middle values of p_+. A particularly interesting result occurs at $p_+ = 0.7$, in fact, even though the majority of agents initially have a positive belief, they struggle with the presence of trolls, which leads to an increase in polarization.

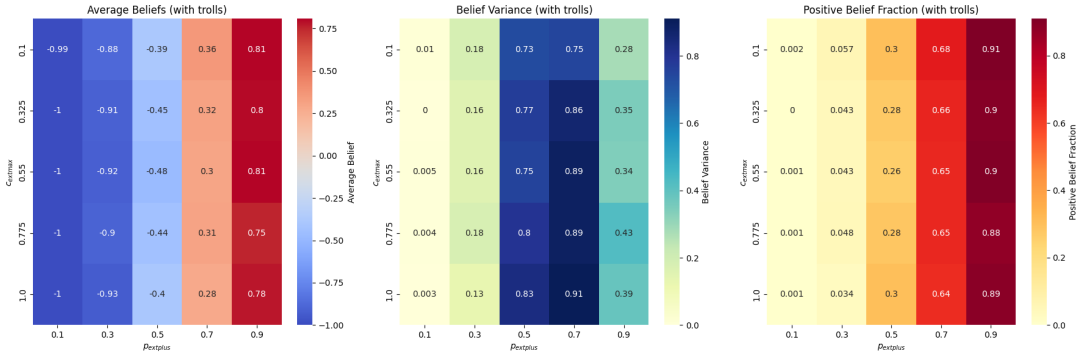


Figure 8: Heatmaps for different combinations of c_{max} and p_+ (with trolls)

3.3.3. Concentration of Negative Beliefs

We ran a final experiment where a portion of the population with negative beliefs was concentrated in a specific region of the grid (Figure 9).

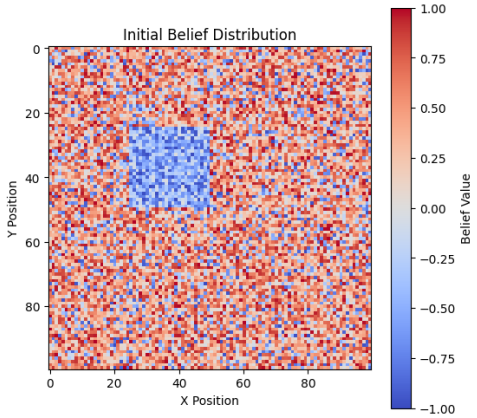


Figure 9: Heatmap of initial beliefs when negative beliefs are concentrated in a specific region

Here, we depict the behavior of the three observables and the final distribution of belief when the portion of the population with a negative belief is concentrated in a specific region, with $c_{max} = 0.5$ and $p_+ = 0.8$, without presence of trolls. From Figure 10, we can observe that the values of average belief and positive fraction increase without reaching their maximum value. On the other hand, polarization increases quickly in the first time steps, then decreases and stabilizes at a value slightly higher than 0.3. From these results, we can deduce that being surrounded by other agents who likely share the same belief makes it less likely for an agent to change its belief.

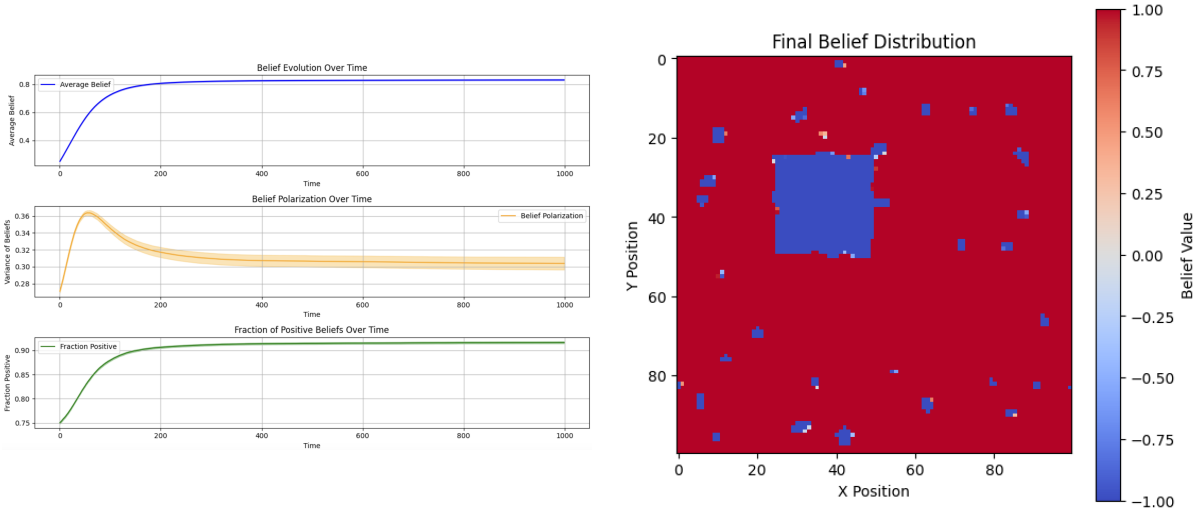


Figure 10: Dynamics (left) and final distribution (right) of beliefs when negative beliefs are concentrated in a specific region.

3.4. Troll Countermeasures

The simulation results shows the progressive impact of different countermeasures against misinformation. For each countermeasure we plot the behavior of the three observables after 3000 steps, and recorded the final simulation results after each countermeasure was applied (see in Table 4).

Model	Average Belief	Belief Polarization	Fraction of Positive
Baseline Model	-0.0360 ± 0.0234	0.9916 ± 0.0022	0.4822 ± 0.0116
Trusted Agents			
Fraction of trusted agents (1%)	-0.0079 ± 0.0283	0.9939 ± 0.0008	0.4963 ± 0.0141
Fraction of trusted agents (5%)	0.0689 ± 0.0301	0.9888 ± 0.0042	0.5349 ± 0.0150
Reports & Shadowbanning	0.2767 ± 0.0325	0.9169 ± 0.0178	0.6380 ± 0.0164
Community Notes	0.4115 ± 0.0367	0.8226 ± 0.0295	0.7063 ± 0.0187
Troll Influence Reduction	0.6691 ± 0.0193	0.5461 ± 0.0257	0.8349 ± 0.0096

Table 4: Final Simulation Results for Different Countermeasures

In the baseline model (see in Figure 4), where no interventions exist, misinformation spreads freely, leading to a steady decline in average belief (-0.0360 ± 0.0234), high polarization (0.9916 ± 0.0022), and a decreasing fraction of positive beliefs (0.4822 ± 0.0116).

When trusted agents are introduced (see in Figure 11), the decline in average belief slows down, reaching -0.0079 ± 0.0283 (1% trusted agents) and 0.0689 ± 0.0301 (5% trusted agents). However, polarization remains high, at 0.9939 ± 0.0008 and 0.9888 ± 0.0042, respectively. The fraction of positive beliefs improves slightly to 0.4963 ± 0.0141 and 0.5349 ± 0.0150.

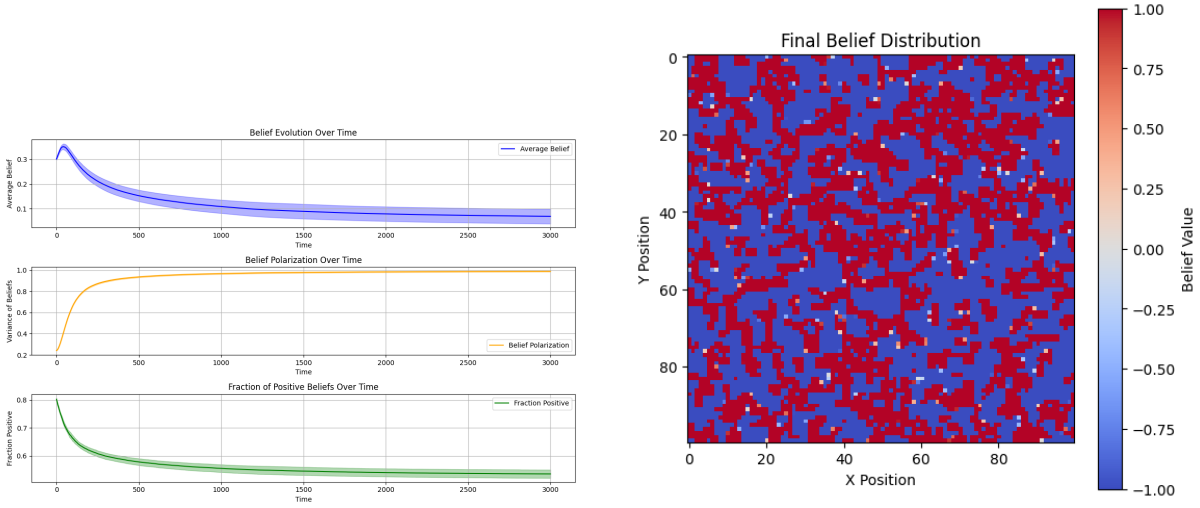


Figure 11: Impact of Trusted Agents

With the addition of reporting and shadow banning (see in Figure 12), the average belief stabilizes at a much higher level (0.2767 ± 0.0325). The belief variance decreases to 0.9169 ± 0.0178, and the fraction of positive beliefs increases to 0.6380 ± 0.0164.

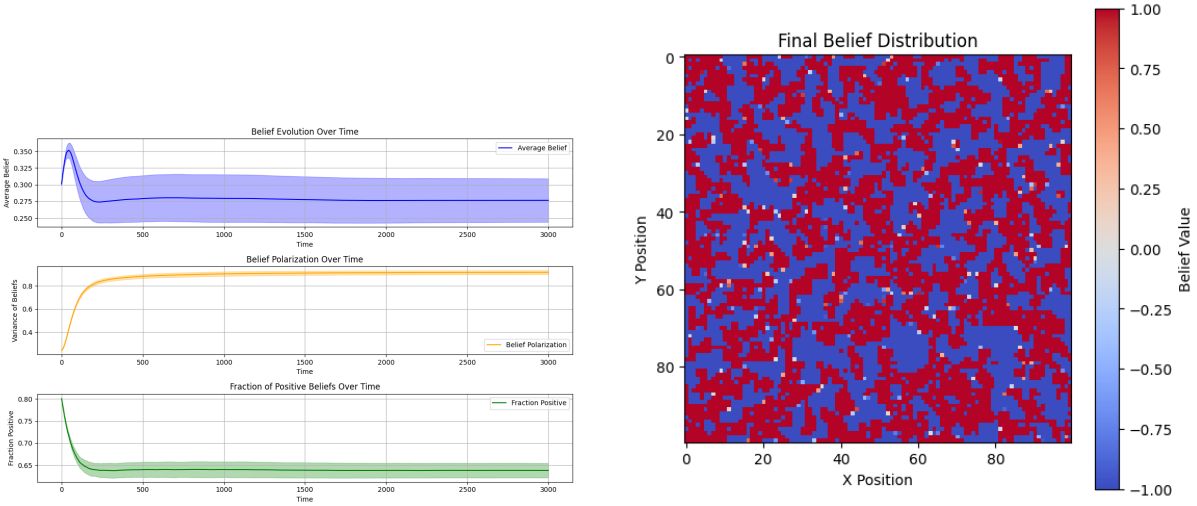
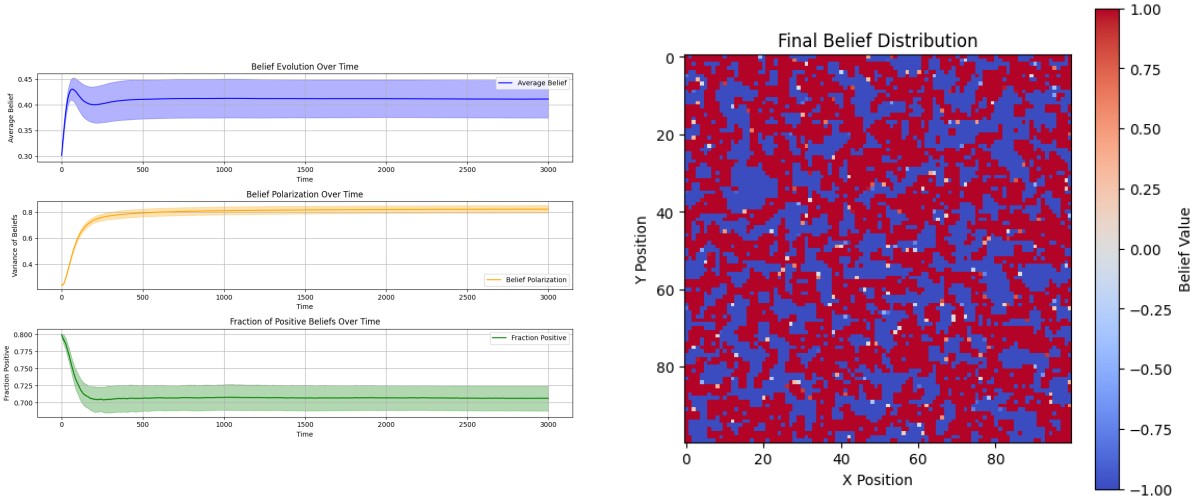
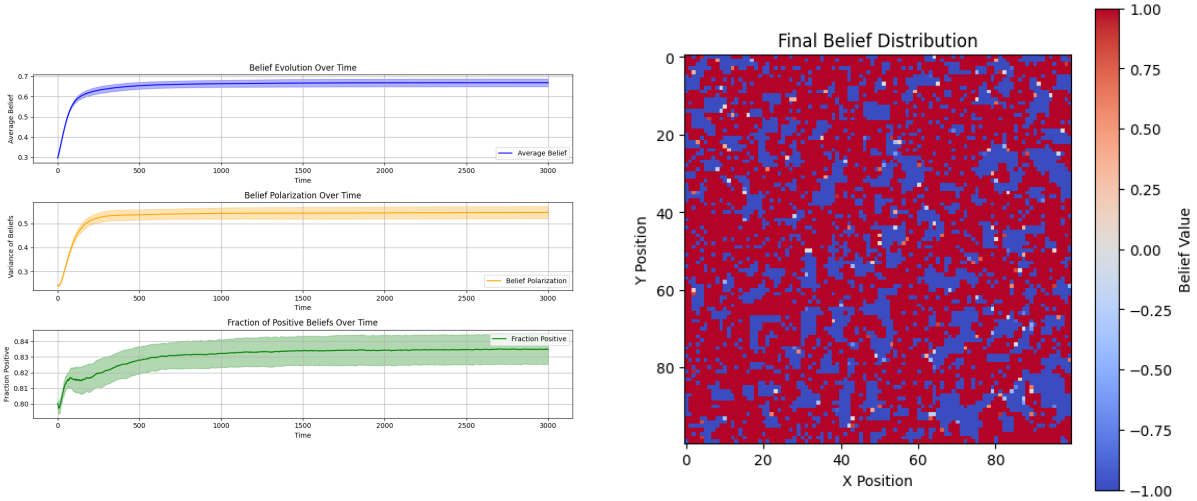


Figure 12: Effects of Reports & Shadowbanning

The introduction of community notes (see in Figure 13a) further enhances belief stabilization. As a result, the average belief rises to 0.4115 ± 0.0367, polarization drops to 0.8226 ± 0.0295, and the fraction of positive beliefs increases to 0.7063 ± 0.0187. When trolls also lose convincing power after being contradicted by community notes (see in Figure 13b), the average belief reaches its highest stable level (0.6691 ± 0.0193), polarization is significantly reduced to 0.5461 ± 0.0257, and the fraction of positive beliefs increases to 0.8349 ± 0.0096.



(a) Application of Community Notes



(b) Troll Influence Reduction via Community Notes

4. Discussion

Summary. The project investigated the spread of fake news using an Agent-Based Model on a population grid. Each individual (agent) has a belief ranging from -1 (indicating belief in fake news) to 1 (indicating trust in fact-checked information). Individuals can influence each other through a von Neumann neighborhood. Important parameters are given and key variables such as average belief, fraction of positive belief and belief variance are tracked over time. Next trolls, agents with a fixed belief of -1, were introduced to see their influence on the population and determine what percentage was needed to make half the population belief in fake news. Different initial parameters combinations and conditions were explored to see how the important variables changed. Finally, different counter measures were implemented to reduce the impact of trolls.

Result Interpretation Examining the final distribution of beliefs (Figure 2) in Section 3.1, we observe that agents with negative beliefs tend to remain clustered in small groups. This may

indicate the formation of echo chambers, where misinformation persists even in the presence of broader exposure to fact-checked content. Moreover, we noticed that a Von Neumann neighborhood of order 2 doesn't affect much the spread of the beliefs when trolls are not present. Indeed, from Figure 5 we observe that the dynamics of the three observables do not change significantly compared to the initial implementation using a Von Neumann neighborhood of order 1 (Figure 1). However, with the introduction of trolls, we noticed that that the average belief and the fraction of positive beliefs when using a Von Neumann neighborhood of order 2 (Figure 6) are higher compared to the Von Neumann of order 1 (Figure 4), while the variance, and therefore the polarization, is lower. In a more general context, we could suggest that having more neighbors might actually facilitate the spread of the initial majority belief.

The simulation results of troll countermeasures demonstrate that a multi-layered approach is essential for mitigating misinformation. In the baseline model, misinformation spreads freely, leading to a decline in average belief, high polarization, and a decreasing fraction of positive beliefs, showing trolls' strong influence. Trusted agents slow this decline, but polarization remains high (Figure 11), indicating that credible sources alone are insufficient to prevent misinformation. Reporting and shadow banning significantly reduce troll influence, leading to higher belief stability, lower polarization, and greater resilience against misinformation(Figure 12). Community notes further enhance correction, ensuring that misinformation is actively challenged and belief divergence is minimized(Figure 13a). When trolls lose additional convincing power after being contradicted, the system achieves the highest belief stability, lowest polarization, and strongest misinformation control(Figure 13b).

Conclusion Our simulation highlights the complex evolution of belief in a networked society, influenced by convincing power, initial belief distribution, troll activity, and social connectivity. Higher convincing power (c_{max}) accelerates consensus formation, allowing dominant beliefs to spread more effectively while reducing long-term polarization. Similarly, a higher initial probability of positive beliefs (p_+) increases the likelihood of a positive belief-dominated system, even in the presence of negative influences. However, troll activity significantly disrupts this process, sustaining polarization and slowing convergence. Our results reveal a critical troll fraction ($f_{troll} \approx 0.207$), beyond which negative beliefs become dominant, preventing the spread of factual news.

The structure of social interactions also plays a crucial role in shaping belief dynamics. When agents interact with a broader set of neighbors (Von Neumann order 2) instead of just immediate neighbors, belief fragmentation decreases, and convergence toward a dominant belief becomes more stable. Conversely, localized clusters of negative beliefs can persist if left unchallenged, particularly when convincing power is low, forming misinformation pockets that resist broader consensus.

To counteract the effects of trolls and misinformation, our results suggest that a multi-layered intervention strategy is most effective. Trusted sources alone slow the spread of misinformation, but cannot fully prevent polarization. Shadowbanning trolls significantly reduces their influence, stabilizing belief evolution, while community-driven fact-checking actively corrects misinformation, reducing long-term belief divergence. Furthermore, weakening trolls' influence when they are contradicted leads to the most stable belief environment, minimizing polarization and misinformation spread.

Overall, our findings emphasize the importance of social influence mechanisms in belief dynamics and the need for proactive misinformation control strategies. By implementing a combination of trusted information sources, reduced exposure to disruptive agents, and active misinformation correction mechanisms, belief stability can be maintained, polarization minimized, and social

networks made more resilient against the spread of false information. These insights have direct implications for designing robust online information ecosystems that foster constructive discourse and prevent harmful belief fragmentation.

Outlook While the implemented model provides valuable insights into the spread of fake news and the influence of trolls on public opinion, it also has limitations that could be improved upon future work. One of them is a lattice-induced bias. Working with a von Neumann neighborhood, interactions only directly propagate vertically and horizontally. Another shortcoming of this neighborhood is the rigidity. An agent only has a set amount of neighbors and doesn't directly interact with anyone else. People frequently change their social circles and engage with different communities. Online environments, in particular, exhibit highly fluid network structures. A model that allows changes in neighborhoods for agents could reflect changes in both offline and online contexts. Despite these limitations, the model serves as a useful testbed for evaluating potential countermeasures against misinformation. Several effective strategies have already been identified, including shadowbanning, trusted agents, and community notes. In the future these strategies could be refined and tested further in different environments. One possible extensions the introduction of algorithmic and cognitive biases, where agents more likely to interacted and be susceptible to like-minded individuals. Another could be giving the agents a way to have their past interactions impact their susceptibility, a memory of sorts. Overall, this work lays a foundation for studying misinformation dynamics in controlled settings and offers a framework for testing interventions that could mitigate the harmful impact of trolls and fake news in society.

References

Alassad, M., Hussain, M. N., and Agarwal, N. (2023). Developing an agent-based model to minimize spreading of malicious information in dynamic social networks. *Computational Mathematics and Organization Theory*, 29:487–502.

Gausen, A., Luk, W., and Guo, C. (2021). Can we stop fake news? using agent-based modelling to evaluate countermeasures for misinformation on social media. In *Proceedings of the Department of Computing, Imperial College London*, UK. Imperial College London.

Tseng, S.-H. and Son Nguyen, T. (2020). Agent-based modeling of rumor propagation using expected integrated mean squared error optimal design. *Applied System Innovation*, 3(4).

A. Appendix

A.1. Acknowledgments

We acknowledge the use of AI tools for assistance in improving syntax, grammar and clarity of the writing of this report. Prompts such as "Improve the following text for clarity and grammar" were used to refine the text.

A.2. Code Snippets

Agent Class is implemented by:

```
class Agent:
    def __init__(self, i, j, is_troll, belief, convincing_power):
        self.i = i  # Grid x-coordinate
        self.j = j  # Grid y-coordinate
        self.is_troll = is_troll  # Boolean indicating if the agent is a troll
        self.belief = belief  # Belief value between [-1, 1]
        self.convincing_power = convincing_power  # Influence strength [0, cmax]
```

Listing 1: Agent Class

The implementation of trusted agents is achieved by:

```
...
                if is_trusted:
                    belief = 1.0
            ...
            convincing_power = (
            np.random.uniform(0, cmax) + np.random.uniform(0, cmax/2)
                if is_trusted
```

Listing 2: Introducing Trusted Agents

The implementation of reporting and shadow banning is achieved by:

```
...
active_fraction = 0.05  # Fraction of active agents that report
report_threshold = 5  # Reports required to shadowban a user
    ...
```

```
        """Return the convincing power, halved if the agent is shadowbanned."""
        return self.convincing_power / 2 if self.shadowbanned
    ...
        # Simulate content flagging
        if a1.is_troll and not a2.is_troll:
            if random.random() < 0.7: a1.reports += 1
        elif a2.is_troll and not a1.is_troll:
            if random.random() < 0.7: a2.reports += 1
        ...
            if agent.reports >= report_threshold: agent.shadowbanned = True
```

Listing 3: Introducing Reporting and Shadow Banning

The implementation of community notes is achieved by:

```
...
community_note_threshold = 2 # Number of disagremments required
    ...
        if agent.belief < 0:
            # Track the number of disagreements with neighbors
            disagreements = sum(
                1 for nb in neighbors if not nb.is_troll and np.sign(nb.belief)
                    != np.sign(agent.belief)
            )
            # Apply community notes effect if disagreements exceed threshold
            if disagreements > community_note_threshold:
                agent.belief += 0.1  # Increase belief in true news
```

Listing 4: Introducing Community Notes

An improvement in the effect of community notes is achieved by:

```
...
if agent.belief < 0:
            ...
                if not agent.is_troll:
                    ...
                # For troll agents, decrease convincing power (also trusted
                    agents can lose convincing power)
                else:
                    agent.convincing_power *= 0.7  # Reduce convincing power by
                        30%
```

Listing 5: Introducing Community Notes

A.3. Results for Changed Initial Parameters

A.3.1. Without Trolls

c_max	**p_plus**				
	0.1	0.3	0.5	0.7	0.9
0.100	-0.992	-0.768	-0.012	0.777	0.991
0.325	-0.992	-0.872	-0.034	0.820	0.997
0.550	-0.993	-0.817	-0.063	0.826	0.996
0.775	-0.995	-0.816	-0.048	0.829	0.993
1.000	-0.999	-0.796	-0.048	0.769	0.997

Table 5: Average Beliefs for Different $c_{\max}$ and p_{plus} values (without trolls)

c_max	**p_plus**				
	0.1	0.3	0.5	0.7	0.9
0.100	0.012	0.348	0.868	0.334	0.013
0.325	0.014	0.230	0.965	0.313	0.006
0.550	0.014	0.328	0.979	0.312	0.008
0.775	0.010	0.328	0.983	0.307	0.015
1.000	0.003	0.326	0.988	0.364	0.006

Table 6: Belief Variance for Different $c_{\max}$ and p_{plus} values (without trolls)

c_max	**p_plus**				
	0.1	0.3	0.5	0.7	0.9
0.100	0.004	0.112	0.494	0.891	0.996
0.325	0.004	0.065	0.483	0.910	0.998
0.550	0.004	0.092	0.468	0.913	0.998
0.775	0.003	0.092	0.476	0.915	0.996
1.000	0.001	0.102	0.476	0.898	0.998

Table 7: Positive Belief Fraction for Different $c_{\max}$ and p_{plus} values (without trolls)

A.3.2. With Trolls

c_max	**p_plus**				
	0.1	0.3	0.5	0.7	0.9
0.100	-0.994	-0.880	-0.386	0.357	0.808
0.325	-1.000	-0.913	-0.448	0.318	0.795
0.550	-0.997	-0.916	-0.479	0.304	0.805
0.775	-0.998	-0.904	-0.438	0.307	0.752
1.000	-0.998	-0.932	-0.395	0.281	0.779

Table 8: Average Beliefs for Different $c_{\max}$ and p_{plus} values (with trolls)

c_max	**p_plus**				
	0.1	0.3	0.5	0.7	0.9
0.100	0.010	0.184	0.735	0.747	0.280
0.325	0.000	0.159	0.769	0.864	0.348
0.550	0.005	0.157	0.754	0.891	0.340
0.775	0.004	0.178	0.797	0.892	0.426
1.000	0.003	0.130	0.835	0.910	0.387

Table 9: Belief Variance for Different $c_{\max}$ and p_{plus} values (with trolls)

c_max	**p_plus**				
	0.1	0.3	0.5	0.7	0.9
0.100	0.002	0.057	0.305	0.681	0.909
0.325	0.000	0.043	0.275	0.659	0.898
0.550	0.001	0.043	0.261	0.652	0.903
0.775	0.001	0.048	0.281	0.653	0.876
1.000	0.001	0.034	0.303	0.641	0.889

Table 10: Positive Belief Fraction for Different $c_{\max}$ and p_{plus} values (with trolls)