Retweet, Repeat, Deceit: How Content Amplifiers Created Fake News Loops on Twitter During the COVID-19 Pandemic

Github URL: *https://github.com/HuyAnhVuTran/Team-9_group-project/tree/main*

Kyle CHANDRASENA

Yuri MATIENZO

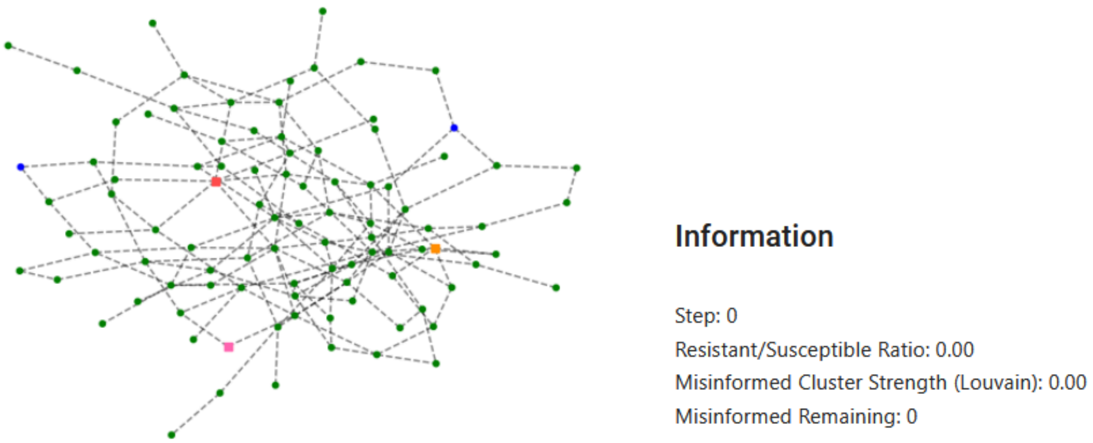
Huy Anh Vu TRAN

Section 1: Final Report Overview

The COVID-19 Pandemic was one of the largest outbreaks recorded in recent years. It prompted widespread misinformation regarding the origin of the virus, potential treatments or protections, and the severity and prevalence of the disease. During this pandemic, AI bot interactions on social media platforms like Twitter were responsible for creating self-reinforcing misinformation cycles, or fake news loops, where falsehoods repeatedly resurfaced and gained credibility through repetition. This misinformation was largely propagated through content amplification bots, which will be referred to as Misinformation Bots throughout this report. These entities engage in activities such as liking, sharing, retweeting, and commenting to amplify misinformation, reinforcing its spread through engagement loops (Himelein-Wachowiak et al., 2021). Misinformation bots amplify content and interact with one another by forming coordinated networks that systematically boost false narratives. These AI-driven interactions create high-engagement misinformation clusters that algorithms interpret as trending, prioritizing their visibility in users' feeds and reinforcing the fake news loop. The rapid dissemination of COVID-19 information and the exponential spread of the virus led to a surge of contradictory content on social media, creating what has been termed an 'infodemic' (Himelein-Wachowiak et al., 2021). The COVID-19 infodemic fostered confusion and fear on social media as people sought out any news related to the pandemic, regardless of whether the information was credible or not. This report aims to provide a comprehensive analysis of how media ecosystems can be exploited by social bots during times of crisis and urgency to foster persistent fake news loops.

Social media platforms like Twitter are useful tools for sharing information in an open forum, however, they also present a great risk as content posted by accounts on the platform do not sustain the same credibility as other news sources. Misinformation bots played a crucial role in sustaining fake news loops related to COVID-19 information on Twitter during the pandemic. In a sample of tweets related to COVID-19, 24.8% of tweets included misinformation, and 17.4% included unverifiable information (Himelein-Wachowiak et. al, 2021). Additionally, during the early months of the outbreak, a study shows that approximately 14% of accounts spreading pandemic content were automated, in other words, bots. (Suarez-Lledo et al., 2022) The inability to distinguish bot-generated and human-generated content overwhelmed Twitter's information ecosystem, creating an infodemic where misinformation gained credibility through sheer volume. The COVID-19 Pandemic is an example of how social bots can amplify the reach of misinformation and contribute to sustaining fake news loops, threatening the stability and credibility of information sharing in media ecosystems. Due to fact-checking organizations taking at least a day to take action and verify content, finding the public's initial reaction is a challenge (Al-Rakhami and Al-Amri, 2020). This project aims to develop an agent-based model to evaluate how specific bot characteristics and network structures influence the persistence and reach of misinformation on Twitter, aiming to simulate how bots contribute to the formation of fake news loops during the COVID-19 Pandemic.

Agent-based modeling (ABM) is an effective method for studying fake news loops because it allows us to model the iterative and interactive nature of these phenomena. ABM can simulate how individual bots and users react to misinformation based on their unique characteristics and decision-making processes. Using an agent-based model, such as Virus on a Network, we can simulate network-based media ecosystems like Twitter and evaluate how the collective behaviors of individual agents contribute to the emergence and persistence of fake news loops.

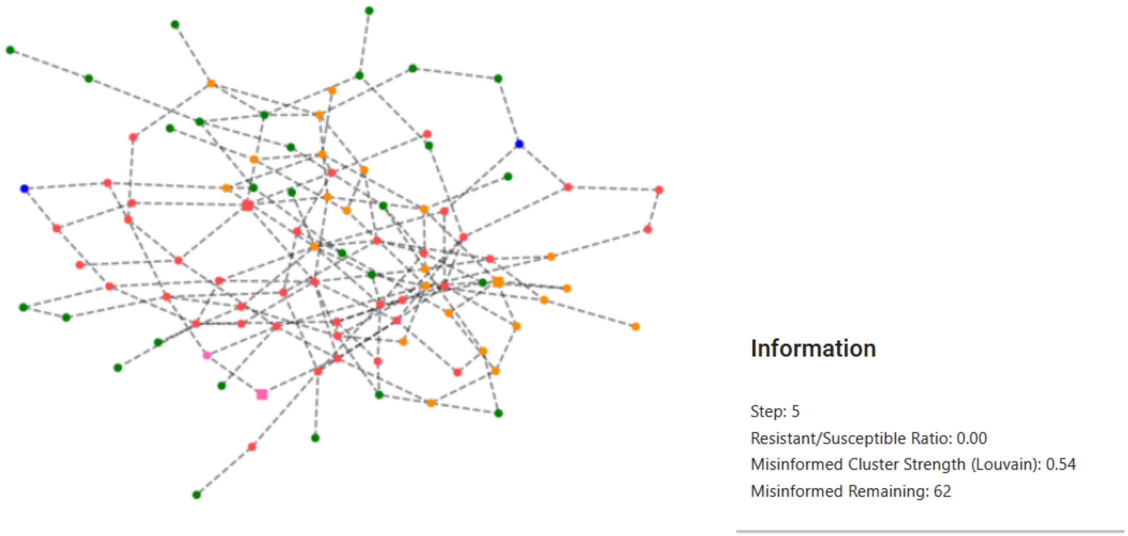


Information

Step: 0
Resistant/Susceptible Ratio: 0.00
Misinformed Cluster Strength (Louvain): 0.00
Misinformed Remaining: 0

Figure 1. Step 0 of the simulation - Initial stage of the simulation

The initial state of the simulation represents the Twitter network before any misinformation is circulated. The simulation has a misinformation bot for each of the three distinct strains (Strain A = red, Strain B =orange, and Strain C = pink). The majority of the network consists of susceptible users (green circles) and fact-checkers (blue circles). Overall, the network structure is loosely connected, with misinformation bots located within different areas of the network, ready to spread misleading content.

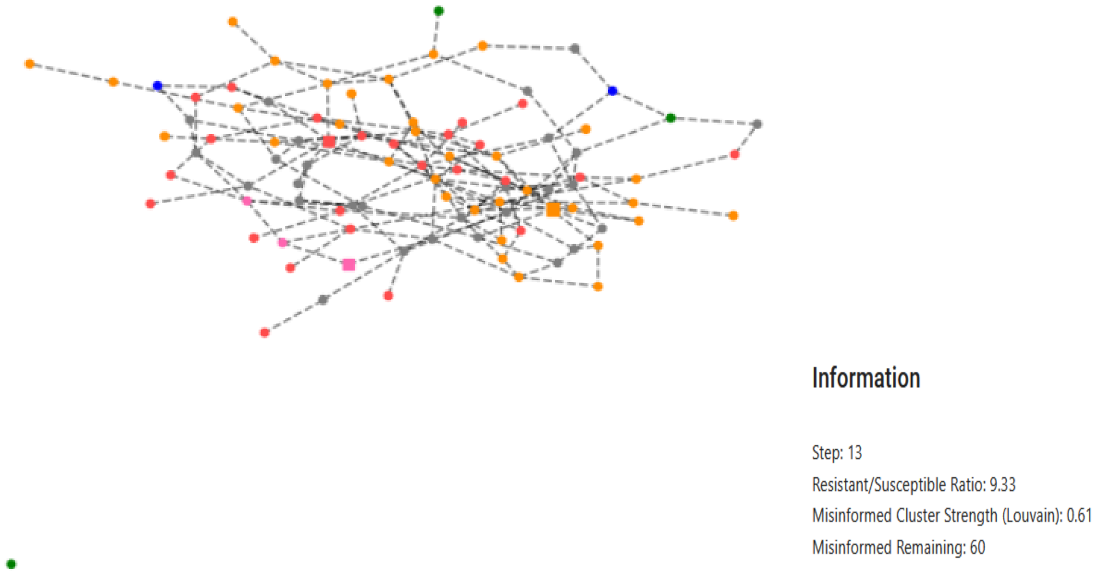


Information

Step: 5
Resistant/Susceptible Ratio: 0.00
Misinformed Cluster Strength (Louvain): 0.54
Misinformed Remaining: 62

Figure 2. Step 5 of the simulation - Initial propagation state

This is the initial propagation state of the simulation where misinformation bots influence neighboring nodes and echo chambers begin to form throughout the network. A recent study using a sample of collected data from Twitter uncovered that more than 85% of bots' tweets are liked, and they have a large number of followers and friends (Zhang et al., 2023). This stage of the simulation reflects how bots leverage Twitter to influence user perceptions about disease transmission and public health. In this example, the bots in central locations on the network were able to influence numerous users and dominate the network.

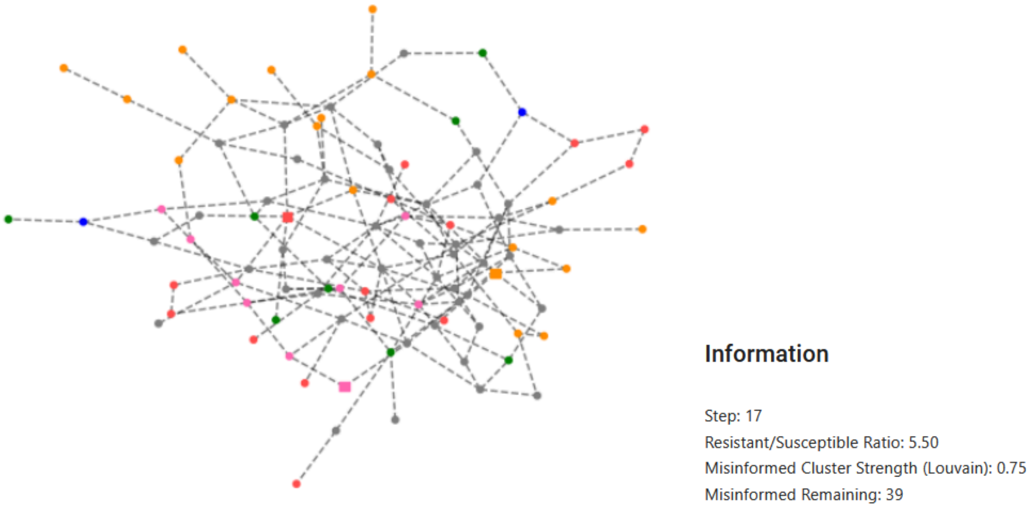


Information

Step: 13
Resistant/Susceptible Ratio: 9.33
Misinformed Cluster Strength (Louvain): 0.61
Misinformed Remaining: 60

Figure 3. Step 13 of the simulation - New user added to the network

In step 13, there is a new susceptible user node that has yet to be added to the network. This reflects the real-world phenomenon that social media like Twitter is constantly growing, meaning there are always new users waiting to be a part of the platform

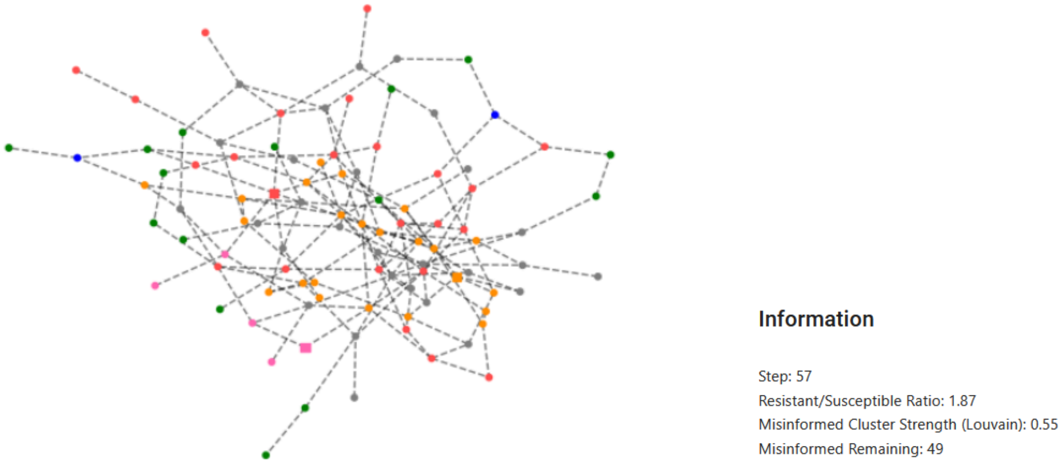


Information

Step: 17
Resistant/Susceptible Ratio: 5.50
Misinformed Cluster Strength (Louvain): 0.75
Misinformed Remaining: 39

Figure 4. Step 17 of the simulation - Initial intervention state

In step 17, there is a rise in the number of resistant users in the network, meaning that fact-checkers have stepped in to slow down the spread of misinformation across the network.



Information

Step: 57
Resistant/Susceptible Ratio: 1.87
Misinformed Cluster Strength (Louvain): 0.55
Misinformed Remaining: 49

Figure 6. Step 57 of the simulation - Reemergence of misinformation

By step 57 of the simulation, the network has been overtaken by misinformation once again. Misinformed users with different misinformation strains are once again widely distributed across the network. There are still some susceptible users remaining, meaning the whole network has yet to be convinced of this reemergence.

Section 2: Simulation Design & Implementation

System Overview

Our simulation is based on a modified Mesa Virus on a Network model to simulate misinformation spreads on Twitter and the creation of fake news loops. The core components of our model consist of a network-based environment used to simulate a simplified version of a Twitter ecosystem populated by five agent types. The model uses three distinct strains of misinformation to simulate the diversity of stories and information propagated by bots during the COVID-19 Pandemic (i.e., political conspiracies, side effects of vaccines, potential cures for COVID-19, etc). To differentiate between bots and human agents of the same strain, we implemented bot agents as square-shaped nodes and human agents as circle-shaped nodes. These agents interact within the network environment, with their behaviors and state transitions determined through probabilistic rules and predefined parameters provided by the user when the simulation is initialized. In our simulation, fake news loops emerge as a dynamic consequence of the network's behavior, specifically when it exhibits a cyclical pattern between phases of misinformation propagation and intervention attempts. This pattern arises from the interaction of Misinformation Bots, which initially spread misinformation, and Fact-checker and Resistant Users, which attempt to counter it. The resulting cycles of misinformation spread and intervention simulate the self-reinforcing nature of fake news loops.

Simulation Environment

Our simulation operates in a network-based environment, where agents represent users and bots on Twitter. This environment is supposed to reflect a real-world environment on social media, specifically Twitter, where misinformation can spread through interconnected nodes (users). Each agent of the simulation is represented as a node in a social graph, interacting with each other through edges that represent interactions on Twitter (e.g., follow, retweet, like).

Our simulation uses several user-defined parameters to control the initialization of the network graph and influence the potential outcome. Key parameters to focus on are the *Misinformation Spread Chance*, *Fact Check Chance*, *Resistance Duration*.

Misinformation Spread Chance is a value used in probabilistic decisions when the model decides if a user will become misinformed. During the COVID-19 Pandemic, governments and public health agencies were slow to disseminate information online and build public trust, leaving room for the spread of COVID-19 misinformation and conspiracy theories (Weng & Lin, 2022). The sense of urgency and the desire to understand what was going on made users on Twitter more likely to share misinformation before it had been fact-checked (Bryant et al., 2025).

In our simulations, we used a slightly elevated value (0.4) for *Misinformation Spread Chance* to simulate how urgency and the desire for new information made Twitter users more susceptible to misinformation. *Fact Check Chance* is a value used in probabilistic decisions when the model decides if a user will become resistant. To attain an ethical simulation without biasing a specific outcome, we used a value of 0.4 to provide an equal chance of being influenced by misinformation or verified information. *Resistance Duration* is a value to determine how long an agent is resistant to misinformation after interacting with verified information from a *Fact Checker User* or a *Resistant User*. This parameter is used to avoid stagnant simulations and it effectively models how preconceived notions about trending information may change over time.

Agent Design:

Our simulation includes five agents:

Human Users: These are people who used the Twitter social media platform to interact with COVID-19 discussions during the Pandemic. These users are split into four specific entities that have different behaviors, roles, and goals on the platform. In our simulation, all human user agents are given circle-shaped nodes, and we will focus on the importance of the node-shape convention when discussing Misinformed Users and Misinformation Bots.

- **Fact-checker Users:** These agents represent Twitter users who participated in the Birdwatch Program (now Community Notes). These users were given the authority to verify claims and provide context to misleading posts by submitting explanatory notes. In our simulation, these human agents represent the intervention method employed by Twitter to counteract the formation of fake news loops. At each step of the simulation, *Fact-checker User* agents check neighboring nodes for strains of misinformation. These agents attempt to convert *Misinformed User* agents to a *Resistant* state using the *Fact Check Chance* parameter defined by the user. The decision to intervene is probabilistic: a random value is generated and compared to the *Fact Check Chance* parameter. If the random value is lower, the intervention succeeds, and the *Misinformed User* agent changes to a *Resistant* state.
- **Susceptible Users:** These agents represent Twitter users who can be influenced, manipulated, or deceived by information from external sources like misinformation, propaganda, or scams. These human agents are the base state for users in the simulation. Currently, these agents can only be influenced by Misinformation Bots and Misinformed User agents. In our final version of the simulation, we aim to implement preemptive intervention to allow Fact-checker User agents or Resistant User agents to Influence Susceptible User agents to transition to a Resistant state. These agents do not have any decision-making processes because they are usually acted upon by other agents in the simulation.

- **Misinformed Users:** These agents represent Twitter users who believe, share, or engage with misleading content posted by misinformation bots or other misinformed users. In our simulation, these human agents represent users propagating misinformation to create fake news loops. *Misinformed User* agents can be created when *Susceptible User* agents interact with either *Misinformation Bots* or existing *Misinformation User* agents. The decision to propagate misinformation is probabilistic: a random value is generated and compared to the *Misinformation Spread Chance* parameter defined by the user. If the random value is lower, the propagation succeeds, and the *Susceptible User* agent changes to a *Misinformed* state for the respective misinformation strain. *Misinformation Bot* agents are given circle-shaped nodes, with their color indicating the specific strain of misinformation they are propagating. When the propagation method succeeds, the misinformation strain and node color of the agent are passed to any *Misinformed User* agents it created.
- **Resistant users:** These agents represent Twitter users who are reluctant to believe information posted on the platform without sufficient evidence. In our simulation, these human agents represent users who are not influenced by misinformation content shared by *Misinformation Bots* or *Misinformed User* agents. In real-world media ecosystems like Twitter, users who are not verified fact-checkers may engage with misinformation content to discuss and educate others based on credible information or personal knowledge. These users work as an intervention method to slow the spread of misinformation on Twitter. However, without the authority and platform affordances given to fact-checking users, they may not achieve the same success rate. To incorporate this element in our simulation, we developed a new *Influence Chance* parameter, which is an adjusted, lowered value that uses the *Fact Check Chance* parameter as its base. The decision to intervene is probabilistic: a random value is generated and compared to the *Influence Chance* parameter. If the random value is lower, the intervention succeeds, and the *Misinformed User* agent changes to a *Resistant* state. We implemented a *Resistance Duration* parameter, which is used to control when *Resistant User* agents return to a *Susceptible* state. This was implemented to avoid situations where the simulation becomes stagnant and to model that as new information emerges online, previously conceived assumptions and mindsets change. Our final simulation adds a new parameter, *Relapse Chance*. This parameter is used to reflect that informed users can be swayed by misinformation on topics of which they do not have prior knowledge. In our simulation, a *Resistant User* will store their previous strain of misinformation, and if they interact with a different strain, the *Relapse Chance* parameter is used in a probabilistic decision-making process to determine if the user will be infected.

Misinformation amplification bots: These agents represent bot accounts on Twitter that are designed to spread misinformation across the platform. These automated accounts engage in activities that increase the visibility and perceived credibility of false or misleading information. They play an important role in AI-to-AI interactions in media ecosystems by reinforcing fake news loops through content engagement. In our simulation, these bot agents act as the catalyst for the emergence of fake news loops, which are visualized through cluster formation. At each step of the simulation, *Misinformation Bot* agents check for neighboring *Susceptible User* agents to potentially propagate misinformation. These bot agents are assigned a specific strain of misinformation (A, B, C), which is passed to each *Misinformed User* agent they create. The decision to propagate misinformation is probabilistic: a random value is generated and compared to the *Misinformation Spread Chance* parameter defined by the user. If the random value is lower, the propagation succeeds, and the *Susceptible User* agent changes to a *Misinformed* state for the respective misinformation strain. During the simulation, if one strain of misinformation has dominated the network for a long period, a *Misinformation Bot* will be switched from the strain with the maximum number of users to the strain with the minimum number of users. *Misinformation Bot* agents are given square-shaped nodes, with their color indicating the specific strain of misinformation they are propagating. When the propagation method succeeds, the misinformation strain and node color of the bot is passed to any *Misinformed User* agents it created.

To make the simulation feasible and as realistic as possible, we decided to simplify and abstract several factors that contribute to interactions on social media platforms like Twitter.. Firstly, human behaviors in the real world are influenced by various psychological and social factors, which are challenging to implement in our simulation. Therefore, we abstract these factors by transforming agent decision-making into probabilistic rules based on specific parameters (e.g., misinformation spread chance, fact-check chance). This helps reduce the psychological complexity of users' content engagement into manageable actions. However, this simplification may limit the model's ability to fully capture the nuances of individual responses to misinformation and the potential for unpredictable behavior. Moreover, users on Twitter can interact with each other through retweeting, liking, and following. In our model, we simplify these interactions by representing them as edges between nodes. Additionally, in real-world scenarios, some social media influencers can amplify content more than a misinformation bot or an average user while algorithmic recommendations suggest content to users. These two factors add complexity to the simulation, implying that each agent in the network has a different level of interaction with others. In our simulation, we decided to have all agents operate with the same level of interaction.

Interaction Dynamics:

Our simulation implements the RandomActivation scheduler within the Mesa framework to govern agent activation. This scheduler was selected to represent accurately the unpredictable nature of real-world interactions and the random spread of misinformation on social media networks. The RandomActivation scheduler ensures that agents are activated in a randomized order, preventing potential biases that could arise from a predetermined activation sequence. Our implementation leverages the built-in scheduler functionality of newer Mesa versions (3+), which replaces traditional defined schedulers with direct method calls as specified in the Mesa migration documentation.

Bot-to-bot interactions in our simulation are not direct and emerge through competition between misinformation strains. Initially, bots propagate their assigned misinformation strains, leading to the formation of distinct misinformation clusters. Over time, these clusters compete for dominance in the network and may coordinate efforts to alter the trending topics of the network. For example, if one strain becomes dominant, bots may switch to propagating a less prevalent strain to increase its visibility. This dynamic simulates the evolving nature of misinformation campaigns and the way fake news loops emerge and persist in media ecosystems.

Our simulation incorporates several key assumptions to simplify the complex dynamics of misinformation spread. Human agents' susceptibility to misinformation is modeled probabilistically, influenced by factors like the *Misinformation Spread Chance* parameter and interactions with neighboring agents. This simplification abstracts the complex psychological and social factors that influence individual beliefs and sharing behavior. We also assume that the verified platform fact checkers and resistant users are the only prevention methods in the network. While platform policies and regulations may control misinformation in the real world, they are difficult to simulate and were not included in the model.

Data Collection & Visualization:

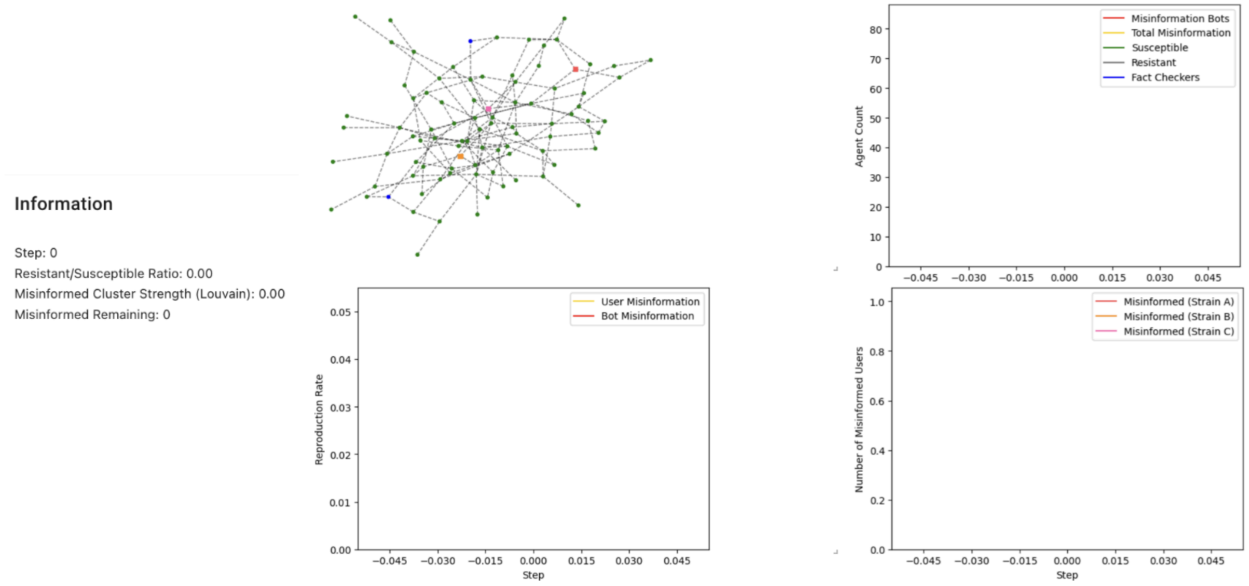
The simulation data collection focuses on tracking the spread of misinformation and its propagation methods, as well as the efficacy of countermeasures with our simulated Twitter network. The model records the counts of *Misinformation Bots*, *Misinformed Users*, *Susceptible Users*, *Resistant Users,* and *Fact-checker Users*, providing a method to plot agent state transitions over time. The model calculates the reproduction rates of misinformation from bots and misinformed users to compare how these agents contribute to fake news loops. The model tracks the number of bots and users influenced by each strain to monitor patterns in propagation over time. The collected data is used to create three line graphs: the State plot, the Reproduction plot, and the Strain plot. The State plot allows us to monitor agent transitions and fluctuations in total misinformation spread on the network. This graph gives a method for tracking the impact of misinformation on other agent types at each step of the simulation. The Reproduction plot allows us to monitor our hypothesis proposed in Deliverable 2, where we stated an assumption that misinformed users may have a higher chance of spreading misinformation to susceptible users than misinformation bots. The reproduction rate metrics used in this graph calculate the new misinformed users generated by bots and users for each step. The Strain plot is used to monitor fluctuations in strain distribution when clusters form and how these strains respond to intervention methods.

We wanted to implement a metric for analyzing the strength of clustering in the network as it would help us connect how echo chamber formation contributes to sustaining fake news loops. Our initial implementation used the average clustering coefficient function from networkx. This implementation provided inconsistent results because of the distributions of edges to various nodes, which would often make the simulation crash. We decided to use the Louvain Community Detection Algorithm to extract the community structure misinformation in the network. By applying this algorithm to a subgraph of the network containing only misinformation nodes (bots and users), we could calculate a value to represent the strength and connectivity of clusters of misinformation in the network. The final version of the model defines this value as the Louvain Modularity, used to measure the connectivity between misinformed nodes in the network. This metric is used to measure how well misinformation clusters in the network

Section 3: Observations & Results

The simulation demonstrates a dynamic interplay of misinformation spread and intervention, illustrating the cyclical nature of fake news loops

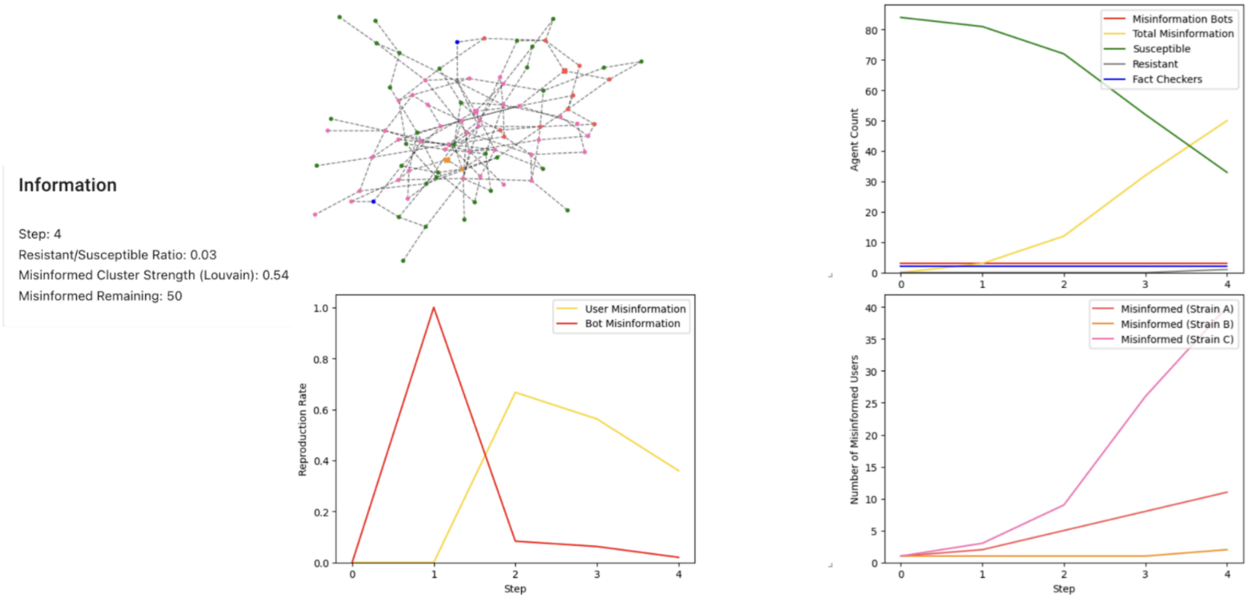
Initial Network State



Information

Step: 0
Resistant/Susceptible Ratio: 0.00
Misinformed Cluster Strength (Louvain): 0.00
Misinformed Remaining: 0

Initial Propagation Stage



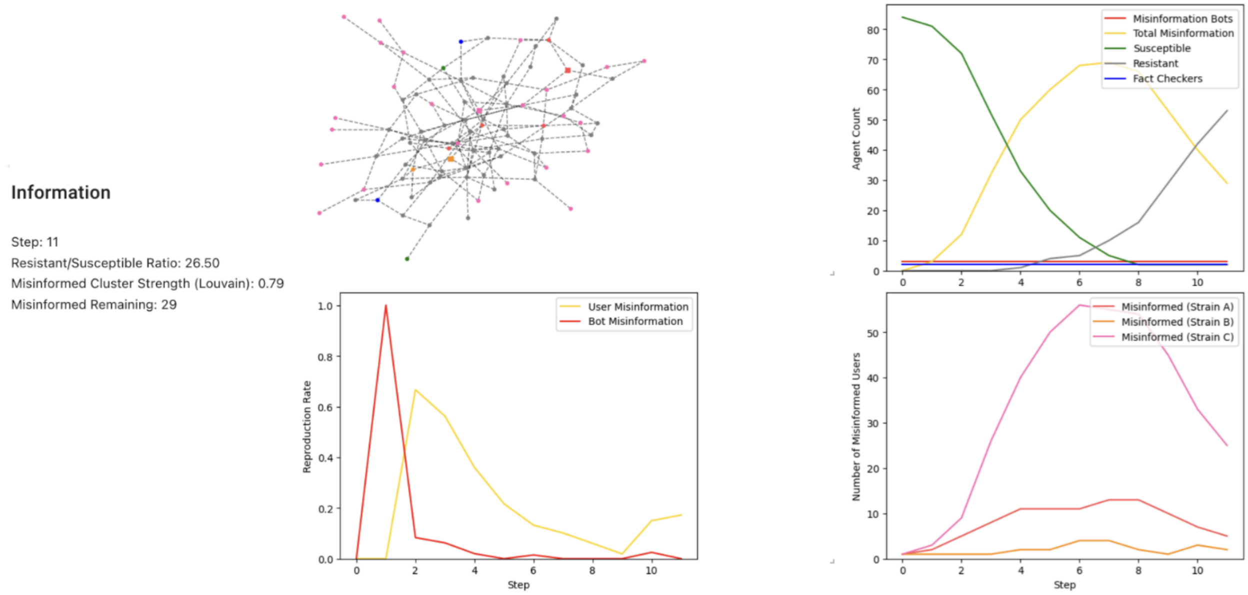
Information

Step: 4
Resistant/Susceptible Ratio: 0.03
Misinformed Cluster Strength (Louvain): 0.54
Misinformed Remaining: 50

This stage of the simulation models the network transitioning from a neutral state to a state of propagation. The model simulates the early formation of echo chambers in the network as misinformation bots begin to influence neighboring nodes.

- State Plot: Rapid decline in the number of susceptible users in the network as the total number of misinformed agents in the network (bot and humans) rises
- Misinformed Cluster Strength (Louvain): this metric rises from 0.00 at Step 0 to 0.54 by Step 4, suggesting the early formation of misinformation clusters and the beginning of echo chamber formation
- Reproduction Plot: shows bots as initial catalysts and users as sustained propagators, mirroring how fake news loops during the COVID-19 pandemic became self-sustaining through user engagement

Initial Intervention Stage



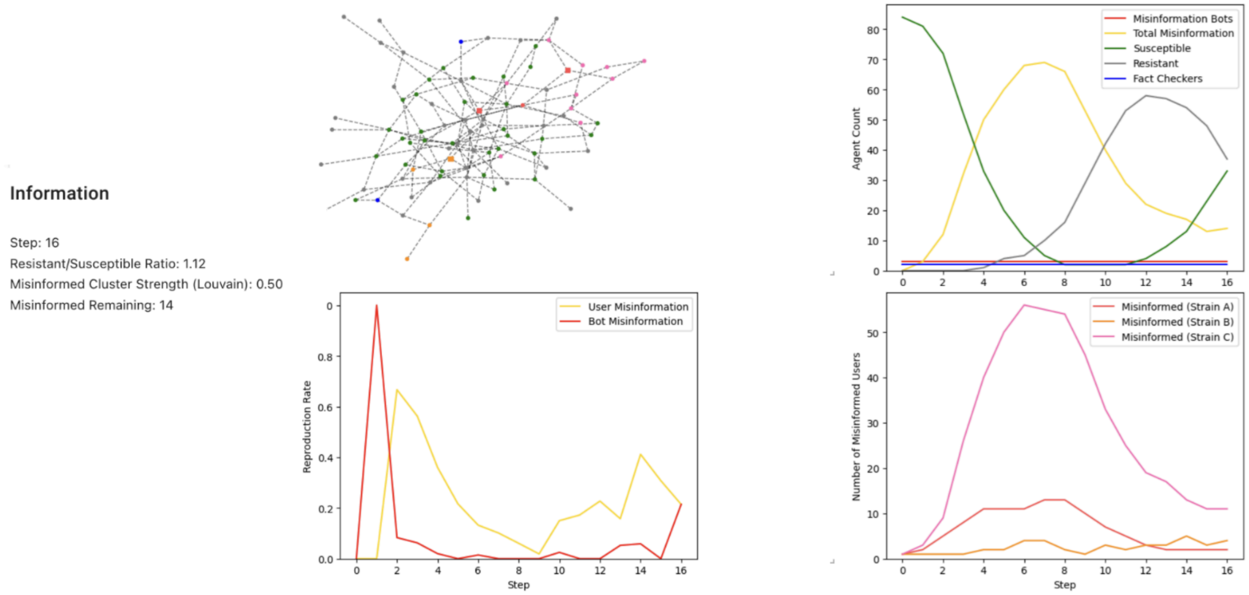
Information

Step: 11
Resistant/Susceptible Ratio: 26.50
Misinformed Cluster Strength (Louvain): 0.79
Misinformed Remaining: 29

This stage of the simulation models the network transitioning from a propagation state to a state of intervention. The model simulates how fact-checkers on Twitter interact with misinformation to break down echo chambers in the network by educating and informing users about inaccurate content. Once users transition to a *Resistant* state, they seek to influence *Misinformed Users* on the network. This simulates how users on Twitter who are not verified fact-checkers might contribute to intervention methods through discussion threads on posts or comments.

- State plot: Rapid decline in the number of total misinformation users in the network as resistant users in the network rise

Misinformation Bot switches from Strain C to Strain A

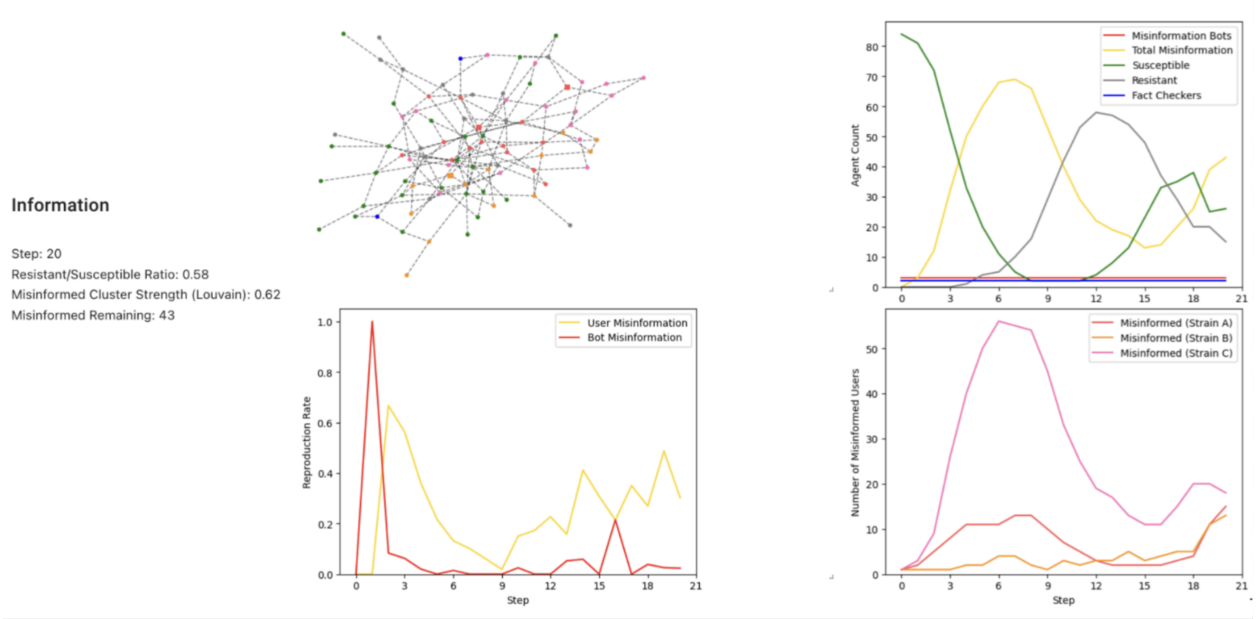


Information

Step: 16
Resistant/Susceptible Ratio: 1.12
Misinformed Cluster Strength (Louvain): 0.50
Misinformed Remaining: 14

This stage of the simulation illustrates how bots switch between infecting strains of misinformation based on trending content in the network. If one strain of misinformation is dominating the network for a long period of time, the model will allocate one of the max strain bots to the misinformation strain with the least interactions. In this simulation, Strain C was dominating the network for the initial fake news loop cycle, and at Step 16, the bot for Strain C was allocated to Strain A to boost interactions with this misinformation campaign. This functionality was implemented to model how groups of bots can function in coordination with each other, forming what are called botnets (Himelein-Wachowiak et. al, 2021)

Second Propagation Stage



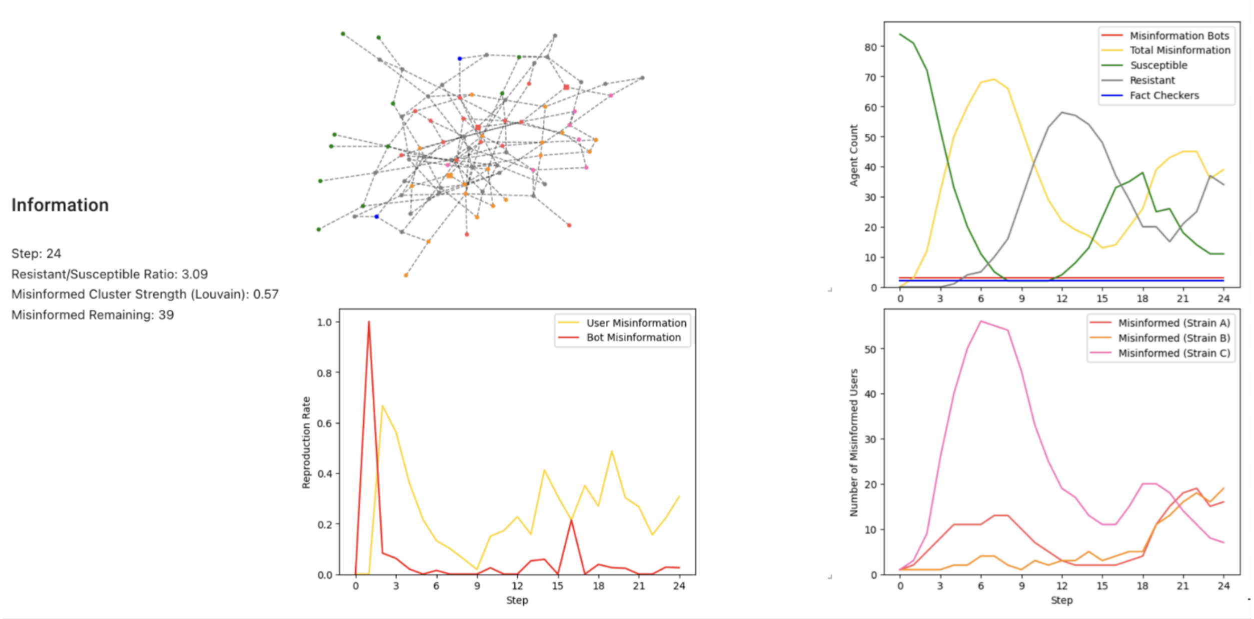
Information

Step: 20
Resistant/Susceptible Ratio: 0.58
Misinformed Cluster Strength (Louvain): 0.62
Misinformed Remaining: 43

This stage of the simulation models the reemergence of misinformation in the network as bots have formed new echo chambers and a new strain of misinformation begins to dominate the network.

- Strain Plot: Strain C begins to rise as an echo chamber forms around the newly allocated bot located in the center of the network

Second Intervention Stage



Information

Step: 24
Resistant/Susceptible Ratio: 3.09
Misinformed Cluster Strength (Louvain): 0.57
Misinformed Remaining: 39

This stage represents the end of the second cycle of propagation and intervention as *Fact Checker Users* and *Resistant Users* curb the spread of misinformation and break down the central echo chamber.

Emergent behaviors:

The simulation created recurring cycles of misinformation as the network went through stages where bot-driven activities influenced users and intervention methods attempted to respond in return. During stages of propagation, *Misinformation Bots* acted as a catalyst, influencing users in the network to create echo chambers to isolate *Fact Checker Users* from intervening. Afterward, the simulation would enter an intervention stage as *Fact Checker Users* and *Resistant Users* attempted to reduce the spread of misinformation. This cyclical pattern continued throughout the simulation, emulating the persistence of fake news loops during the COVID-19 Pandemic. Social media platforms can be exploited when they are used as an information source because content posts do not have inherent credibility. Our simulation suggests that fake news loops persist in media ecosystems as a result of continued user engagement and the urgency for interactions. In a real-world context, the initial spread of misinformation during the COVID-19 pandemic was facilitated by the sparsity of verified information, which both hindered the ability to counter false narratives and increased people's susceptibility to believing and sharing unverified claims due to fear (Geronikolou & Chrousos, 2021). Similarly, in our simulation, once misinformation circulates on the network, it gains credibility and becomes difficult to fully eliminate as long as *Susceptible Users* remain on the network. The emergence of misinformation cycles in our simulation represents how self-reinforcing fake news loops were established on Twitter as a result of bot-driven amplification during the COVID-19 Pandemic.

Unexpected behaviors/emergent dynamics and their cause:

A key unexpected behavior from the simulation is how the location of a *Misinformation Bot* in the network impacts its ability to amplify narratives. Our simulation randomly places bots in the network when initialized and forms connections based on the user-defined value for node degrees. The random placement can create networks where a *Misinformation Bot* is isolated from the central cluster of users. Alternatively, a *Misinformation Bot* can be placed in the center of the network with several connections to human users. We discovered that a bot's initial location impacts the number of users it influences, with isolated bots leading to fewer *Misinformed Users* for a given strain. This emergent behavior is primarily attributed to the *Average Node Degree* parameter, which determines the average number of connections each node has. Our simulations used a value of 3 to produce a network that balanced connectivity and neutrality for observing misinformation trends. Increasing this value would likely increase interactions for isolated bots but further concentrate interactions in the network's center. To mitigate the location bias, we implemented the strain switch functionality, simulating how bots might coordinate efforts to boost less successful misinformation campaigns.

Section 4: Ethical & Societal Reflections

Ethical Considerations

Our simulation models the spread of misinformation on social media platforms like Twitter during the COVID-19 pandemic. It is important to emphasize that this project did not involve the direct collection or use of real-world user data from Twitter or any other platform. The data within the simulation is synthetically generated to represent agent behaviors and network interactions. We avoided the use of real data to adhere to ethical considerations regarding user privacy, data breaches, and unauthorized use of personal information. The design of our simulation is informed by real-world observations of the impact bots made in sustaining fake news loops during the pandemic. As previously mentioned in Section 2, we used a slightly higher *Misinformation Spread Chance* value to simulate the increased susceptibility to misinformation by users on Twitter due to the scarcity of credible information during the early stages of the pandemic. While our simulation does not directly use real-world data, the insights gained have implications for understanding and addressing the ethical challenges of misinformation in media ecosystems.

Societal Implications

The findings of our simulation have broader implications for understanding the impact of misinformation in media ecosystems. At the micro-level, the simulation illustrates how individual users can be influenced by misinformation and how echo chambers can isolate them from alternative perspectives, leading to polarization and distrust. At the meso-level, the simulation demonstrates the role of bot-to-bot interactions in amplifying misinformation narratives and sustaining fake news loops that distort the information landscape on social media platforms and erode user trust. At the macro-level, the simulation uncovers how states of urgency, like the COVID-19 pandemic, facilitate the mass spread of misinformation. Twitter users during the early months of the pandemic were more likely to believe, share, or engage with COVID-related misinformation. The fake news articles spread faster than true news articles because real human users, not bots, were more likely to retweet fake articles (Himelein-Wachowiak et al., 2021). Furthermore, the spread of the infodemic not only fueled the rise of racist attitudes and behaviors but also posed a significant global risk by putting both the health of the population and the ability of governments to implement effective preventive measures at risk (Weng & Lin, 2022).

Our simulation demonstrates a strong alignment with real-world patterns of misinformation propagation and intervention. As mentioned in Section 2, our simulation effectively modelled the cyclical nature of fake news loops on Twitter, where bots continued to propagate misinformation trends after intervention efforts as long as susceptible users remained on the network. Our simulation emphasized that urgency and uncertainty are key factors that make fake news loops persistent. Penn Medicine's Anish Agarwal had the following to state regarding how social media ecosystems set the stage for misinformation to arise and persist in cycles,

> We saw this all the time in early COVID-19, with false information spreading about the vaccines and people using buzzwords like 'hydroxychloroquine' and 'nano-robots' and the like. I didn't think people actually believed what they were reading, but when I asked those who weren't vaccinated against COVID why not, they would say it's either because they saw this thing or heard that thing. It gives you a deeper understanding of how people are processing what they see on social media, be it true or not. With COVID-19, we're watching it happen every single day, in real time (Penn Today, 2025)

Agarwal stated, "A decade ago, what a doctor or nurse said was highly trusted, but in today's landscape, I don't know whether that's still true, partially because on social media anyone can say almost anything" (Penn Today, 2025). Agarwal's perspective highlights how social media platforms lack inherent regulations and credibility to act as information networks. As our simulation illustrates, the vulnerability of the network lies in susceptible user engagement, which bots capitalize on to perpetuate misinformation and enhance its perceived credibility via fake news loops.

This simulation has the potential to be repurposed for malicious intent through using patterns and emergent behaviors of the model to develop strategies for targeted misinformation campaigns. By modeling the impact of bot placement and the effectiveness of strategies like coordinated botnets, the simulation provides insights that could be misused to design more effective misinformation campaigns, exploit vulnerabilities in social media platforms, or manipulate public opinion for malicious gain. This highlights the importance for a broader AI governance perspective, including ethical guidelines for AI research and development, transparency and accountability in the use of algorithms and social simulations.

Section 5: Lessons Learned & Future Directions

Development Reflections

Developing our agent-based simulation provided our team with valuable insights into the technical and ethical challenges of modeling the complex interactions of bots and human agents. One of the earliest challenges we encountered was preventing simulation stagnation. In the early stages of implementation, misinformation spread rapidly in the initial steps and then plateaued as most users transitioned into a resistant state. This behaviour failed to reflect the persistence of misinformation loops as seen in real-world social media ecosystems.

To address this problem, we implemented a resistance duration parameter, enabling resistant users to revert to susceptible states after a duration of time. This behavior was implemented to reflect that user mindsets and preconceived opinions may change due to repeated exposure to misinformation on the network. This parameter allowed us to simulate the cyclical pattern and longevity of misinformation clusters more accurately. Our early versions of the model only used a single strain of misinformation, which produced static, predictable results as the network would become fully influenced by misinformation and then transition to resistant states, and then the cycle would repeat. Our solution was to include multiple strains of misinformation, which would be propagated simultaneously throughout the network. This created a competition dynamic where each strain of misinformation raced to establish echo chambers that would facilitate its spread throughout the network. This behavior accurately represents how bots coordinated efforts during the pandemic to push various false narratives about vaccine side effects, conspiracy theories, and more. Implementing multiple strains introduced a problem for tracking agent state transitions as bots and users shared the same color depending on their misinformation strain. We implemented unique node shapes where bots were represented by square-shaped nodes and human users were represented by circle-shaped nodes. This allowed us to use various states in our network simulation while maintaining a clear distinction between agent types.

Model Limitations & Areas of Improvement

Although our simulation captures multiple aspects behind misinformation spread and fake news loops, several oversimplifications remain. Our agent behaviours rely on probabilistic decision-making rather than cognitive thinking. Susceptible and resistant states are treated as binary states, while real users are influenced by factors such as beliefs, peer pressure, individual environments, and trust. Bots are also pre-assigned a specific strain and do not dynamically create content tailored to user vulnerability and trending narratives. Real-world misinformation is highly adaptive and often AI-generated. Fact checkers and resistant users are our only intervention methods, which makes our misinformation-spread and fake news loop mitigation methods limited.

Several refinements still need to be made despite our implementations. Attribute modeling can be introduced where agents would possess traits such as media literacy or influence level. These can affect how prone users are to believing, resisting, or sharing misinformation, enabling more accurate simulations. The integration of real-world datasets, such as topic trend data, along with ethical protocols for user data safety would be critical for empirical observations. Additionally, our model currently treats all affordances as uniform interactions. A future improvement for this would be adding unique weights for interaction edges that would impact the likelihood for misinformation to spread based on different affordances (e.g., liking, sharing, retweeting, commenting). This improvement could help us track which affordances contribute the most to misinformation campaigns on Twitter.

For future applications, our findings can inform the development of automated mitigation strategies by helping platforms like Twitter to identify early indications of misinformation clusters or where fact-checking efforts would prove to be the most effective. Using bot-to-bot and bot-to-human interactions in our simulation can also serve as test areas for fact-checking prioritization or misinformation de-amplification. When it comes to AI governance and safety, our model highlights how autonomous agents can coordinate to undermine public discourse. This information can be important for researchers to develop preventative measures for AI systems being used to spread misinformation. Our simulation can also support digital literacy initiatives by demonstrating the mechanics of echo chambers and cyclical behaviours of misinformation to promote data-driven opinions and narratives along with critical media consumption.

Our model offers a field for potential prototyping by simulating the impact of content moderation approaches and bot detection algorithms in a controlled environment before deploying in real-world environments. Through this, we can illustrate how our model doesn’t only serve as a project but also as a platform for exploring the dynamics of misinformation automation along with other social and ethical questions revolving around misinformation in the real-world social media ecosystems.

Section 6: References

Al-Rakhami, M. S., & Al-Amri, A. M. (2020, August 26). *Lies kill, facts save: Detecting covid-19 misinformation in Twitter*. U.S. National Library of Medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC8043503/

Bryant, L., Bem, G., & Forney, M. (2025, February 4). *Breaking the misinformation cycle*. AFT Washington. https://aftwa.org/newsletters/breaking-misinformation-cycle

Geronikolou, S., & Chrousos, G. (2021, February 3). *Covid-19-induced fear in Infoveillance Studies: Pilot Meta-Analysis Study of preliminary results*. JMIR formative research. https://pmc.ncbi.nlm.nih.gov/articles/PMC7860927/

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., & Curtis, B. (2021, May 20). *Bots and misinformation spread on social media: Implications for covid-19*. U.S. National Library of Medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC8139392/#:~:text=Bots%20also%20employ%20the%20strategy,articles%2C%20and%20are%20more%20likely

Patel, M., Kute, V., & Agarwal, S. (2020). "Infodemic" of COVID-19: More pandemic than the virus. *Indian Journal of Nephrology*, *30*(3), 188. https://doi.org/10.4103/ijn.ijn_216_20

Penn Today. (2025, January 31). *Why covid misinformation continues to spread.* https://penntoday.upenn.edu/news/Penn-research-why-covid-misinformation-continues-spread

Suarez-Lledo, V., & Alvarez-Galvez, J. (2022, August 25). *Assessing the role of Social Bots during the COVID-19 pandemic: Infodemic, disagreement, and criticism*. Journal of medical Internet research. https://pmc.ncbi.nlm.nih.gov/articles/PMC9407159/

Weng, Z., & Lin, A. (2022). Public Opinion Manipulation on Social Media: Social Network Analysis of Twitter Bots during the COVID-19 Pandemic. International Journal of Environmental Research and Public Health, 19(24), 16376. https://doi.org/10.3390/ijerph192416376

Zhang, Y., Song, W., Shao, J., Abbas, M., Zhang, J., Koura, Y. H., & Su, Y. (2023, February 13). *Social Bots' role in the COVID-19 pandemic discussion on Twitter*. International journal of environmental research and public health. https://pmc.ncbi.nlm.nih.gov/articles/PMC9967279/

Section 7: Attestation

- **Kyle Chandrasena:**
    - Kyle Chandrasena served as the lead software developer and contributed significantly to the implementation of the different aspects and concepts of the simulation. By taking charge of the conceptualization, software development, and visualization, Kyle was able to improve the depiction of real-world dynamics of misinformation spread in our simulation. Implementing the dynamic nodes and echo-chamber formation using the Louvain method helped make the simulation more accurate.

- **Yuri Matienzo:**
    - Yuri Matienzo served as a software developer alongside Kyle as well as a validator and writing editor. He contributed a great deal to the implementation of multiple misinformation strains along with the graphs to display them in our simulation. For the report, Yuri was in charge of sections 4 and 5, which are about ethical and societal reflections, along with lessons learned and future directions.

- **Huy Anh Vu Tran:**
    - Huy Anh Vu Tran served as a writer of the original draft, validator of the contents of this deliverable, and software developer. By writing a significant amount of the original draft, validating the code implemented, and implementing dynamic nodes of the simulation, Huy was able to create a solid base of how and what we needed to include in our report while also double-checking the quality of the work.