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

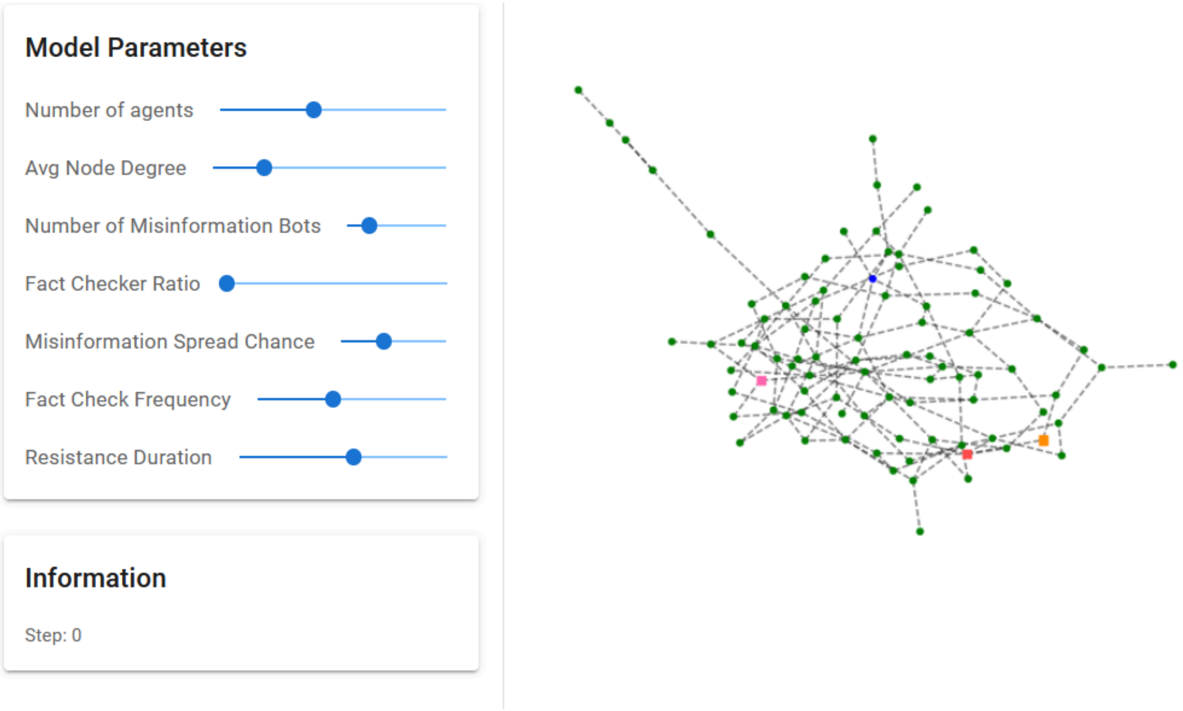
Section 1: Phenomenon Overview

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 through engagement loops 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. The constant state of confusion and hyperreactivity also made them more susceptible to misinformation, which led to some extreme cases such as suicide. (Patel et al., 2020).

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 does 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 the 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 take 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.

Phenomenon Illustration:



Model Parameters

Number of agents

Avg Node Degree

Number of Misinformation Bots

Fact Checker Ratio

Misinformation Spread Chance

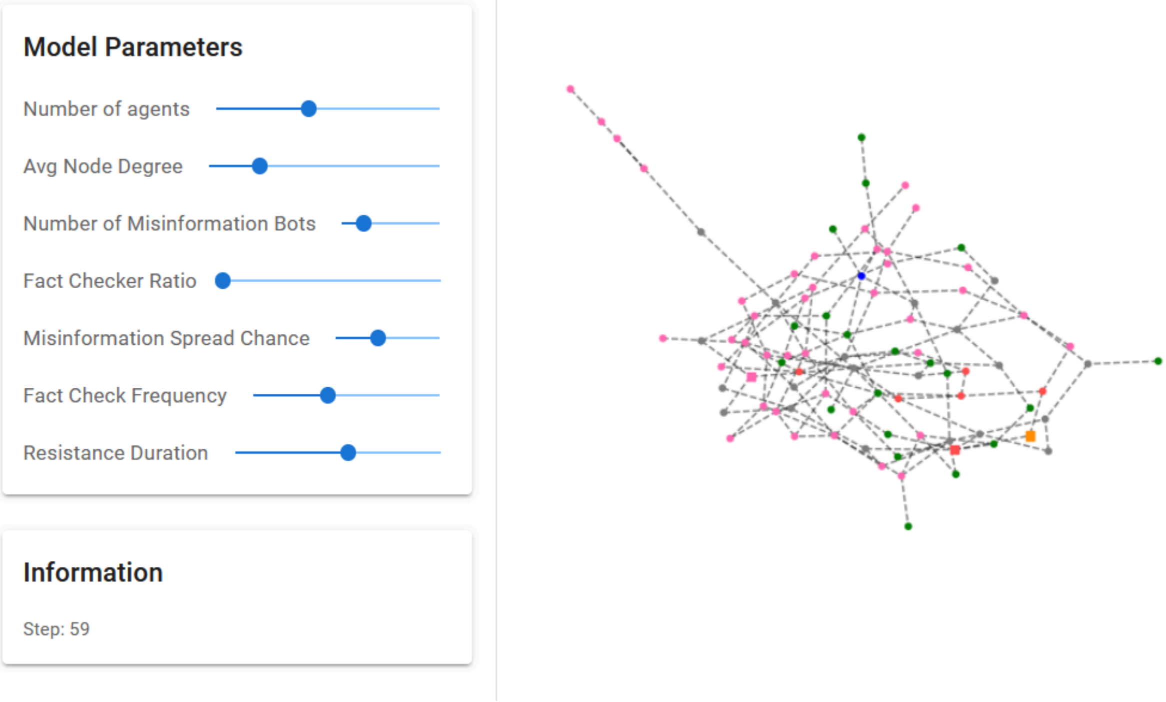
Fact Check Frequency

Resistance Duration

Information

Step: 0

This image represents the initial state of the platform before there is any kind of misinformation spread. At the beginning of the simulation, there are misinformation bots from 3 strains (red squares, pink squares, and orange squares) already present in the network but have not yet influenced users. In the real world, this state represents the initial state where misinformation bots are just deployed. The majority of the network consists of susceptible users (green circles) who have not yet engaged with misinformation bots. There are fact-checkers (blue circles) within the network, but their impact in this state is minimal. Overall, the network structure is loosely connected, with misinformation bots located within different areas of the network, ready to spread misleading content.



Model Parameters

Number of agents

Avg Node Degree

Number of Misinformation Bots

Fact Checker Ratio

Misinformation Spread Chance

Fact Check Frequency

Resistance Duration

Information

Step: 59

By simulation step 59, it is clear that a significant portion of the network has changed. Misinformed users (pink circles, red circles, orange circles) have appeared and are now widely distributed across the network, which indicates that misinformation bots have successfully converted susceptible users to misinformed ones. Some susceptible users remain, which means the whole network has not been convinced by the misinformation yet. Fact-checkers are sparse, illustrating that fact-checking efforts may not have scaled enough to counter misinformation or to slow it down. There are some resistant users (grey circles), which suggests that fact-checking efforts are present, making users resistant to the misinformation, but not enough to cover the whole network.

Section 2: Simulation Design & Implementation

<u>System Overview</u>

Our simulation uses a modified version of the Virus on a Network model from Mesa to simulate how 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 distinct 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.

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).

Agent Design:

Our simulation includes five agents:

Human Users: These are people who used the Twitter social media platform during the COVID-19 Pandemic. Human users may interact with COVID-19 discussions on Twitter. 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.

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. *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.

Early in the development of our modified Virus on a Network model, we identified a critical limitation in its representation of misinformation resistance. The original model allowed *Misinformed User* agents to spontaneously gain resistance against misinformation. This design inaccurately depicted the dynamics of fake news loops in online platforms like Twitter, where resistance typically arises from targeted interventions rather than random occurrences. We adjusted our simulation to enforce intervention agents (*Fact-checker Users* and *Resistant Users)* as the only method for transitioning to a *Resistant* state. This adjustment accurately reflects the real-world scenario observed during the COVID-19 Pandemic, where fake news loops on Twitter were disrupted when *Misinformed Users* interacted with content from educated users or credible sources. By enforcing intervention-driven resistance, our simulation is now more accurate at modeling the dynamics of misinformation control within media ecosystems.

Interaction Dynamics:

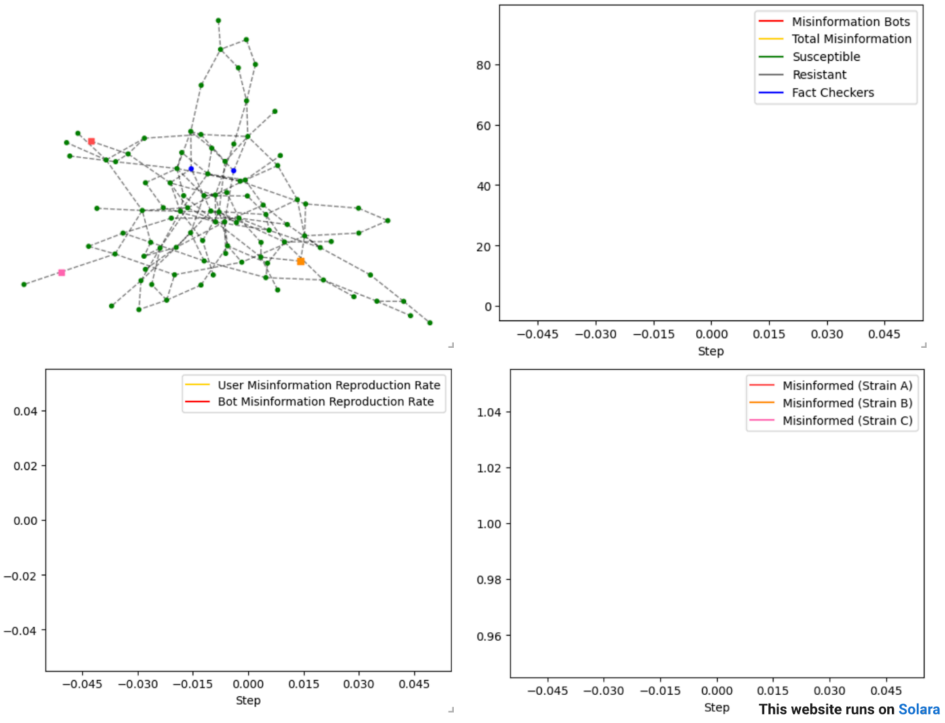
As our simulation is based on the Virus on a Network Mesa Model, we chose to implement the RandomActivation scheduler as it would effectively simulate the unpredictable nature of real-world interactions and virus transmission, or in our case, misinformation propagation. We used the RandomActivation scheduler to ensure that agents are activated in a random order to avoid potential biases that could arise from a fixed activation order. It is important to note that Mesa replaced the RandomActivation scheduler after version 3, which was used in our simulation. To accomplish the same result of RandomActivation scheduling, we consulted the Mesa migration documentation, which outlines the simplified implementation for using RandomActivation in Mesa 3 versions. In our simulation, up to three possible strains of misinformation can be propagated depending on the initial misinformation bot population. These strains simulate different misinformation topics propagated on Twitter by bots during the pandemic. Bot-to-bot interactions emerge in our simulation through the creation and elimination of misinformation clusters, simulating the formation of fake news loops and possible interventions. In our simulation, bots are capable of infecting neighboring nodes with misinformation, and over multiple steps, a pattern emerges where clusters of each strain form around misinformation bots, and fact-checkers and resistant users attempt to contain the spread.

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.

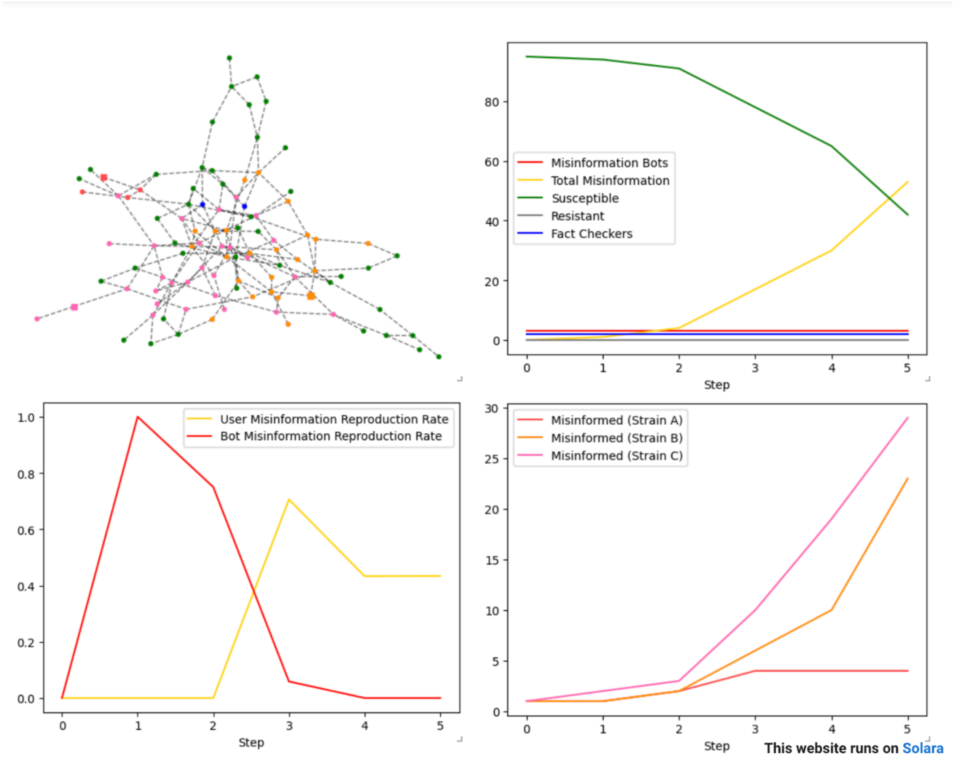
Section 3: Preliminary Observations & Results:

Step 0



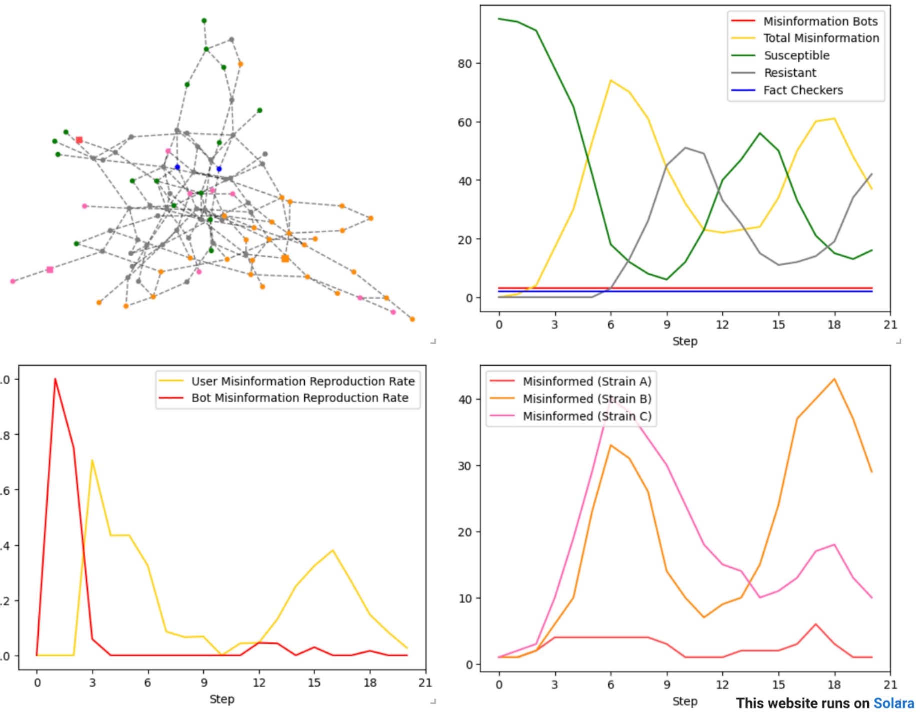
Misinformation Bots
Total Misinformation
Susceptible
Resistant
Fact Checkers
80
60
40
20
0
−0.045
−0.030
−0.015
0.000
0.015
0.030
0.045
Step
User Misinformation Reproduction Rate
Bot Misinformation Reproduction Rate
0.04
0.02
0.00
−0.02
−0.04
Misinformed (Strain A)
Misinformed (Strain B)
Misinformed (Strain C)
1.04
1.02
1.00
0.98
0.96
This website runs on Solara

Step 5



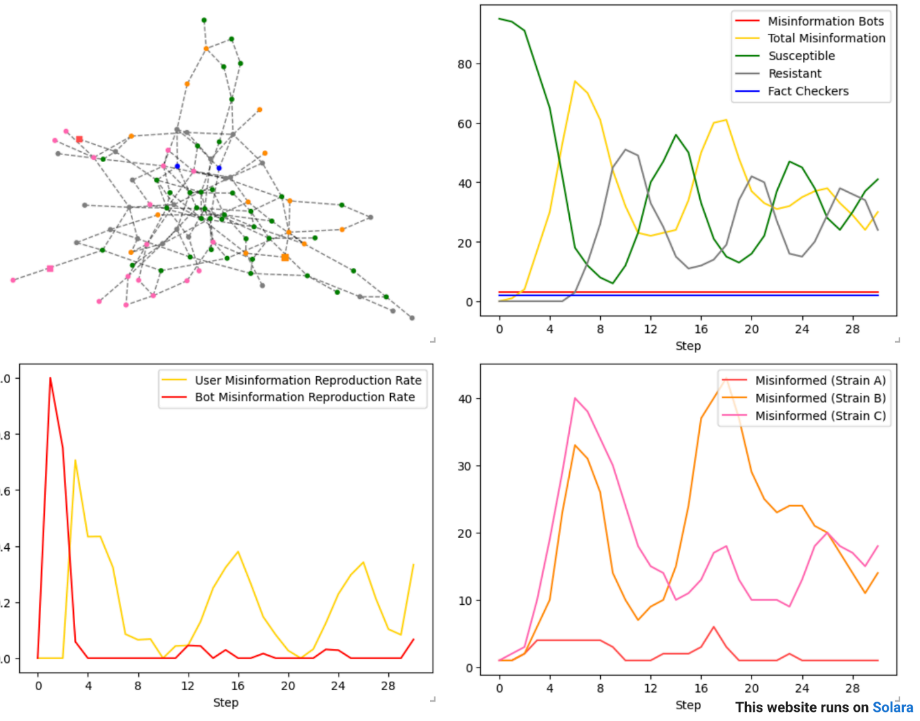
80
60
40
20
0
Misinformation Bots
Total Misinformation
Susceptible
Resistant
Fact Checkers
0
1
2
3
4
5
Step
User Misinformation Reproduction Rate
Bot Misinformation Reproduction Rate
1.0
0.8
0.6
0.4
0.2
0.0
Misinformed (Strain A)
Misinformed (Strain B)
Misinformed (Strain C)
30
25
20
15
10
5
This website runs on Solara

Step 20



This website runs on Solara

Step 30



This website runs on Solara

The simulation demonstrates a dynamic interplay of misinformation spread and intervention, illustrating the cyclical nature of fake news loops. Initially, misinformation spreads rapidly from Step 0 to Step 5, mirroring the explosive growth of false narratives on platforms like Twitter. This initial surge is primarily driven by the activity of *Misinformation Bots*, quickly spreading misinformation across the network. This led to a significant increase in *Misinformed Users* and a rapid decline in *Susceptible Users*, reflecting how users on Twitter were quickly swayed by any new information regarding the COVID-19 pandemic due to uncertainty and panic. As the simulation progresses, the impact of intervention methods, represented by *Fact-checker Users* and *Resistant Users*, becomes more noticeable. The intervention methods in the simulation are able to curb the initial spread of misinformation, however, the misinformation is still present in the network and the fake news loops persist. The reproduction rate plots further emphasize this transition, showing 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. Strain dynamics reveal that various misinformation narratives coexist and fluctuate, reflecting real-world scenarios where multiple false claims circulate simultaneously. The emergence of cyclical patterns in propagation and intervention illustrates that misinformation bots serve as a catalyst for fake news loops as they contribute to the initial formation of echo chambers that isolate users from credible sources.

Emergent behaviors:

The simulation behavior differentiated from our initial predictions of a stabilized network following intervention efforts. Instead, an oscillating equilibrium emerged from Step 20 onward, illustrated in the State plot where cyclical fluctuations occurred in the number of *Misinformed* and *Resistant Users*. Our simulation uses a higher *Misinformation Spread Chance* parameter as we aimed to simulate the desire of Twitter users for any new information regarding COVID-19 during the pandemic. The oscillating nature of the simulation suggests that this parameter effectively created a continuous cycle of infection and recovery. At Step 20, the intervention methods had curbed the initial spread of misinformation, and the network entered a period of oscillation, with the number of *Misinformed Users* and *Resistant Users* fluctuating over time. During this first cycle of misinformation propagation and intervention, the clustering of *Misinformed Users* for strains A and B signified the formation of echo chambers in the network. This illustrates how misinformation propagated by bots can isolate susceptible Twitter users, influencing them to continue engaging with false narratives and sustaining fake news loops. Looking forward to Step 30, the simulation has gone through another cycle of misinformation propagation and intervention, highlighting the persistence of fake news loops and the inability to fully prevent misinformation. This persistence may be attributed to continuous bot activity, user behavior within echo chambers, or network effects that create isolated pockets of misinformation spread. Early runs of our simulation were able to emulate the difficulty in achieving complete control over misinformation in online networks as seen through the emergence of echo chambers which are an inherent part of fake news loops..

Unexpected behaviors/emergent dynamics and their cause:

A key unexpected behavior from the simulation is the tendency of resistant users to become re-infected after some time. While fact-checking efforts can successfully turn misinformed users into resistant users, the resistance will decay over time, turning them back into susceptible users.

This will essentially create a misinformation loop, where users transition between susceptibility and misinformation. This could happen as users can forget corrections, making them vulnerable to misinformation again in the future. Additionally, if resistant users engage with misinformation continuously, they can adopt the false narrative again. Moreover, without drastic measures, verified information can lose its footing over time as misinformation can regain its influence through sheer volume.

Section 4: Challenges & Next Steps

Development Challenges:

One of the most difficult aspects of implementing the simulation was ensuring that the model did not become stagnant, where users were not transitioning between states. Our first version of the simulation contained only one strain of misinformation and after the initial cycle of misinformation propagation, the simulation would become static as there was no new misinformation circulating to change users from a *Resistant* state. To address this issue we implemented a *Resistance Duration* parameter and we implemented multiple strains of misinformation. The *Resistance Duration* parameter allowed us to control how long *Resistant* clusters persisted and it ensured that the simulation would not become static. This parameter was implemented to model that as new information emerges in the network, users can be swayed from previously conceived mindsets.

When implementing multiple strains of misinformation we needed to introduce a mechanism to monitor the current strain of a bot or user and update it for each step. To address this, we implemented a new attribute to handle changes in the current strain of misinformation that was infecting an agent on the network. While the decision-making processes and rule-based interactions were easy to implement for handling multiple strains on the network, it was difficult to represent these changes visually in our network graph. We faced multiple setbacks where nodes would not change color to match their current strain or the graph would be cluttered by up to 9 different user states. To address confusion between bot and user agents of the same misinformation strain, we used a modified version of the mesa visualization library to make all *Misinformation Bots* square-shaped nodes and all human agents circle-shaped nodes. We decided to simplify the various states of our simulation by assigning bots and users of the same strain to the same color, which reduced the amount of visual information users needed to process in the network graph and allowed for a clearer interpretation of the State plot.

Our simulation requires the development and refinement of a few features to produce results that more accurately model the emergence of fake news loops on Twitter during the pandemic. Fake news loops are sustained by expanding the reach of false narratives to new users in a network. Our current simulation is only capable of simulating in a closed, fixed network where the user defines the number of initial agents and the number of users per agent type (*Number of Misinformation Bots* and *Fact Checker Ratio*). This creates an environment where the simulation can become predictable. To address this problem, we aim to implement a method for dynamically adding more nodes to the initial network and if this is not possible we plan to look into how we can change existing user states to add more variability to the simulation. A feature we plan to revise is the allocation of misinformation strains to *Misinformation Bots*. In the current version of our model, each bot is assigned a specific strain of misinformation, however, if the bot is isolated or has few connections, then we see its strain of misinformation does not propagate well. To address bias in misinformation strain propagation we aim to implement a process for strain switching where bots will switch misinformation strain based on lowest propagation. This change aims to model how bots coordinate efforts to bypass prevention methods and circulate multiple false narratives on networks like Twitter.

Our current simulation utilizes various metrics such as reproduction rates, state counts, and strain distribution to monitor the emerging dynamics between bot and user agents in a network environment like Twitter. As our initial simulation uncovered a pattern of echo chamber formation, a metric such as the clustering coefficient will be useful for analyzing how tightly connected *Misinformed Users* are. This metric will complement our existing data by providing an understanding of how network structures contribute to the persistence and amplification of misinformation. For instance, high clustering coefficients within *Misinformed User* clusters would indicate strong echo chamber formation, suggesting that misinformation is reinforced within tightly knit groups, making them more resistant to intervention. Conversely, lower clustering coefficients might suggest a more dispersed spread of misinformation, potentially indicating different intervention strategies are needed. Through analyzing the clustering coefficient, reproduction rates, state counts, and state distributions, we can gain a more comprehensive understanding of how fake news loops evolved and persisted on Twitter during the COVID-19 Pandemic.

Section 5: 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/
- 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
- 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/

Section 6: 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 resistance decay aspect of the simulation along with the addition of multiple graphs and predictive emergent behaviours for the simulation elevated the results and data gathered. In terms of planned future contributions, our team will collectively work on implementing other node behaviors such as dynamically adding nodes throughout the simulation.
- **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 1 and 4, which are about phenomenon overview and challenges respectively. Additionally, he also verified the contents of the video demo by suggesting improvements and making it reach the target time. In terms of planned future contributions, our team will collectively work on implementing other node behaviors such as dynamically adding nodes throughout the simulation along with the report and the presentation.
- **Huy Anh Vu Tran:**
  - Huy Anh Vu Tran served as a writer of the original draft and validator of the contents of this deliverable. By writing a significant amount of the original draft along with validating the code implemented, 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. Additionally, he is also the editor of the video demo, who trims and merges clips into a complete video. For planned future contributions, as

a team, we will be collectively working on the implementation of the improvements of our simulation along with creating a solid and detailed report on the final deliverable.