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: Phenomena of interest

The COVID-19 Pandemic was one of the largest outbreaks recorded recently. It prompted widespread misinformation regarding the origin of the virus, potential treatments or protections, and the severity and prevalence of the disease. This misinformation was largely propagated through content amplification bots, also known as super-spreaders. These bots engage in automated activities such as liking, sharing, retweeting, commenting, and reposting content to artificially boost the visibility of misinformation. A key feature of this process is the interaction between bots themselves, misinformation-spreading bots engage with each other to reinforce and sustain false narratives. In one sample of tweets related to COVID-19, 24.8% of tweets included misinformation and 17.4% included unverifiable information. The authors found no difference in engagement patterns with misinformation and verified information, suggesting that myths about the virus reach as many people on Twitter as truths (Himelein-Wachowiak et al., 2021). This AI-to-AI interaction creates self-reinforcing misinformation cycles, known as fake news loops, where falsehoods repeatedly resurface and gain credibility through repetition. During the pandemic, these loops ensured that misinformation remained persistent on Twitter, reaching many users as verified information. Understanding the role of bot-driven amplification and false news loops is critical for analyzing the spread of misinformation in digital media ecosystems.

Section 2: Relevant works (Reference)

1. ***Bots and Misinformation Spread on Social Media: Implications for COVID-19***
    - During the COVID-19 pandemic, social media bots play a critical role in spreading misinformation and swaying public opinion. These automated accounts amplified false information and narratives, including unverified medical advice

and conspiracy theories, which undermined public health aid and advice. The research focuses on bots strategically engaging with human users which increases the credibility and visibility of the false content. The dissemination of misinformation on mainstream platforms like Twitter and Facebook contributed to public confusion and hesitation towards the vaccine. Mitigating and identifying bot activities poses an obstacle to social media companies and policies. Addressing this issue would need improved bot detection algorithms and public awareness to combat misinformation.

- 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*. Journal of medical Internet research. https://pmc.ncbi.nlm.nih.gov/articles/PMC8139392/#:~:text=Bots%20also%20employ%20the%20strategy,articles%2C%20and%20are%20more%20likely

2. ***Lies Kill, Facts Save: Detecting COVID-19 Misinformation on Twitter***
    - This paper explains how misinformation on COVID-19 spreads through Twitter and proposes machine learning as a foundation to detect it. The authors utilized an ensemble-learning model by combining multiple machine-learning algorithms to identify tweets as non-credible or credible. Their dataset has 400,000 tweets that are labeled based on credibility. The model leverages user and tweet features to achieve high accuracy in misinformation detection. The study highlights the dangers of misinformation which creates the need for better detection tools and emphasizes the role of fact-checking organizations in combating false information online.
    - Al-Rakhami, M. S., & Al-Amri, A. M. (2020). Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter. *IEEE Access*, *8*, 1–1. https://doi.org/10.1109/access.2020.3019600

Section 3: The Core Components of the Simulation

The closest Mesa model for simulating our phenomenon is the Virus on a Network model. This model effectively captures how misinformation spreads across social media, similar to how a virus propagates through a network. In our adaptation, misinformation bots act as infectious agents, susceptible users function as potential hosts, and fact-checkers serve as intervention mechanisms that reduce the spread. Our simulation will incorporate 5 different entities that each represent various types of users on Twitter. By modifying the model, we can incorporate key entities that both amplify and counteract misinformation on Twitter, simulating the dynamics of misinformation outbreaks and the role of AI-to-AI interactions in sustaining fake news loops

Entities

- **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
    - **Fact-Checkers:** These are Twitter users who participated in the Birdwatch Program (now Community Notes). They are given the authority to verify claims and provide context to misleading posts by submitting explanatory notes. If a note reaches a high approval rating, it appears with the misinformation content. This encourages public accountability by challenging misinformation and it increases transparency in fact-checking. Their role is to counter misinformation and educate users regarding false narratives.
    - **Susceptible Users:** These are Twitter users who can be influenced, manipulated, or deceived by information from external sources like misinformation, propaganda, or scams. Lacking the awareness or tools to critically assess credibility, they are the most vulnerable to misinformation. They can either become misinformed users (if influenced by bot-amplified content) or resistant users (if exposed to fact-checking). Their role is to act as neutral users on Twitter who have not yet formed strong opinions about COVID-19-related information.

    - **Misinformed Users:** These are Twitter users who believe, share, or engage with misleading content posted by misinformation bots or other misinformed users. Their actions (liking, commenting, sharing, etc.) contribute to the amplification of falsehoods. However, exposure to fact-checking may reverse their misinformation stance, making them resistant over time. Their role is to consume misinformation and possibly influence susceptible users.
    - **Resistant users:** These are Twitter users who are reluctant to believe information posted on the platform without sufficient evidence. These users are not influenced by misinformation content shared by misinformation bots or misinformed users. They remain immune to bot-driven misinformation. Their role is to prevent misinformation spread by being reluctant to engage in false content
- **Misinformation amplification bots:**

  For the purpose of this study, we will refer to this type of content amplification bot as Misinformation bots. These bots aim to spread misleading content about COVID-19, including false treatments. They amplify the visibility of misinformation by liking, sharing, retweeting, commenting, and reposting content on Twitter. These artificial visibility-boosting activities can manipulate the platform's trending algorithms. Some of them are even programmed to interact with each other to simulate an organic human discourse to avoid detection by Twitter. These bots can act like viruses on a network, sending infectious misinformation that exploits user attention to spread content and mimic human behavior to avoid detection algorithms

Affordances

The content in our simulation consists of Twitter posts, specifically COVID-19 misinformation and general content. These posts serve as the basis for interactions between agents in our simulation. The spread of misinformation is driven by affordances such as liking, sharing, commenting, and tagging which act as engagement mechanisms that sustain fake news loops. These affordances allow misinformation to gain more visibility through increased engagement, increasing the likelihood of further amplification by misinformation bots or misinformed users. The affordances that emerge from Twitter posts contributed to sustaining

fake news loops regarding COVID-19 during the Pandemic. The affordances can be grouped into 3 categories based on how they contribute to the Twitter platform:

1. Metric Boosting – Includes liking, upvoting, and hashtags, which increase visibility by influencing ranking algorithms
2. Comments – Allows users and bots to reinforce misinformation narratives and create the illusion of credibility
3. Sharing (retweeting, tagging, etc.) – Includes retweeting and tagging, which ensure rapid content exposure across the platform

Algorithms

- **Communities Note Ranking Algorithm:** this algorithm ranks fact-checking contributions from real users based on their credibility. It evaluates whether contributors from diverse perspectives agree on a note's accuracy. If it is not deemed helpful by an audience, it is displayed as a misleading tweet. The algorithm prioritizes well-supported explanations and prevents bias by showing that notes are approved by different viewpoints before being shown on the platform. Community notes detect misinformation and neutralize them. If responses from the note gain enough credibility it stops the spread of misinformation to other users.
- **Recommendation Algorithm**: this algorithm suggests content based on individual user engagements and interests. It prioritizes tweets with high interactions, such as likes and replies, and also factors in personalized content preferences, trending topics, and account activity. It amplifies tweets with high engagement while misinformation bots create artificial engagement with each other which creates the spread of misinformation.
- **Trending Algorithm:** this algorithm promotes and identifies topics receiving rapid engagement. By analyzing tweet volume, geographic relevance, and other factors it enables other information to surface and create viral discussions. It prioritizes emerging trends by factoring in hashtags, keywords, and user interactions to maintain trend authenticity. It creates an outbreak where misinformation goes trending faster when bots artificially amplify content through bot-to-bot interaction.

Section 4: Simulation Anticipated Outcomes

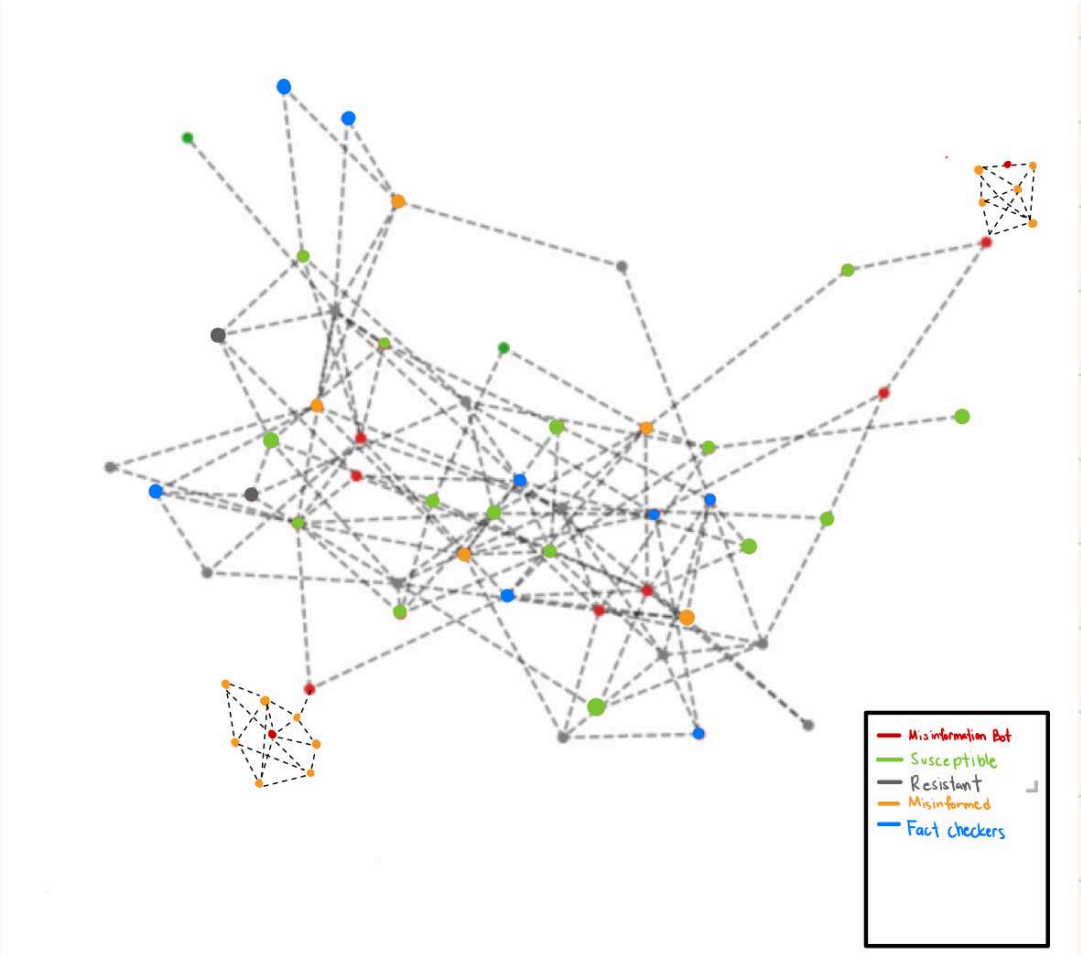

Figure 1. Diagram of virus-on-a-network simulation

The goal of this simulation is to visualize how misinformation spread by content amplification bots can create and sustain fake news loops on Twitter. The simulation will demonstrate how misinformation bots interact with susceptible users to reinforce false narratives and how fact-checkers attempt to counteract this spread. A successful simulation will show self-reinforcing misinformation cycles, where engagement affordances (likes, retweets, etc.) increase exposure, leading to further amplification. If fact-checkers are effective, the model should reveal a disruption in these cycles, reducing the spread and visibility of misinformation over time. Based on information from our research articles, in one sample of tweets related to COVID-19, 24.8% of tweets included misinformation (Himelein-Wachowiak et al., 2021). In our simulation, we hope to find a similar distribution of misinformed users. One key hypothesis is

that misinformed users may have a higher chance of spreading misinformation to susceptible users than misinformation bots because user-generated content appears more believable. Unlike bots, human users add credibility to misinformation by attaching personal opinions, emotions, and perceived authenticity, making their posts more persuasive. If this hypothesis holds, we suspect that some misinformation bots may contribute to the formation of echo chambers containing primarily misinformed users. To evaluate our simulation results we will track key metrics that indicate how misinformation propagates:

- Misinformation Reach – Measures the percentage of users exposed to misinformation over time, showing whether misinformation is spreading or being contained through fact-checking mechanisms
- User State Transitions – A time series visualization of how users shift between different states (susceptible, misinformed, resistant, etc.), revealing trends in misinformation reach and resistance
- Engagement Metrics – Tracks the impact of the activities on Twitter such as liking, sharing, and retweeting which amplifies misinformation visibility
- User-Based Misinformation Reproduction Rate (R_0, m) – Measures how efficiently misinformed users spread misinformation to susceptible users
    - If this rate is higher than the Bot-Based Misinformation Reproduction Rate, it would support our hypothesis that user-generated misinformation is more persuasive than bot-generated misinformation
    - (R_0, m) = $\frac{new\ misinformed\ users\ created\ by\ misinformed\ users}{total\ misinformed\ users}$
- Bot-Based Misinformation Reproduction Rate (R_0, b) – Measures how efficiently misinformed bots spread misinformation to susceptible users
    - (R_0, b) = $\frac{new\ misinformed\ users\ created\ by\ misinformation\ bots}{total\ misinformation\ bots}$
- Echo Chamber Density – This refers to the percentage of users in a network who engage with misinformation sources rarely encountering fact-checkers or opposing viewpoints. This leads to a cycle where misinformation is left unchecked and users are repetitively exposed to similar content without any form of verification from credible sources.