MOCS Final Project: Diverse Misinformation Models: Deepfakes Spread on Online Social Networks

*Juniper Lovato[1], Jonathan St-Onge[1], Alex Friedrichsen[1], Gabriela Salazar Lopez[1], Ijaz Ul Haq[1]

[1]University of Vermont, Burlington, VT
*juniper.lovato@uvm.edu

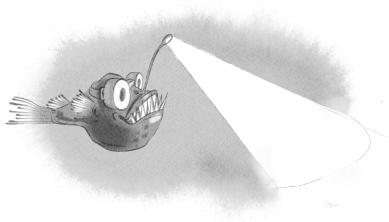


Figure 1: Team Deepfakes Mascot

ABSTRACT

This project focuses on the following research question: how does homogeneity and heterogeneity of users in a network community help or hinder the spread of deepfake videos in an online social network (OSN)? Moreover, how big of an attack would be needed for the deepfakes to dupe most users? We will use three complex modeling approaches to understand the structure and dynamics of this system. We begin with a model of the system's structure by creating a Mixed Membership Stochastic Block Model (MMSBM) voter model. We also model the dynamics of diverse misinformation spread on the network using a contagion model. We then model the interaction between attributes of the users and the system through an agent-based model. Note the first two models we aim to make the model as parameter independent as possible to use the model to compare systems in the future with real-world data. We integrate an agent-based model of the system to look at the effects of introducing various parameters in the system.

1 BACKGROUND

Deepfakes are synthetic images or videos in which the persona in the video is generated synthetically. Creating a deepfake involves training neural network architectures, such as generative adversarial networks (GANs) or existing media, [29]. These neural networks then generate new, synthetic content. Deepfakes are deceptive tools that have gained attention in recent media for their use on celebrity images and their ability to spread misinformation across online social media platforms [26].

Deepfake videos are becoming increasingly more convincing. As they begin to deceive viewers at greater rates, the harms of deepfake videos will further emerge [10]. Deepfakes call into question several ethical considerations: 1) the evidentiary power of video content in legal frameworks [4, 7, 27]; 2) consent of the individual(s) featured in deepfake videos [13]; 3) bias in deepfake detection software and training data [14]; 4) worsening of disinformation and public trust [4]. As deepfakes present harms and require ethical considerations, it is further necessary to understand who gets duped by them and how this impacts the system as a whole.

Deepfakes were easily detectable with the naked eye during the early years of this trend due to their unnatural visual attributes [22]. However, research and technological advancement have allowed adversaries to improve deepfakes, making them more challenging to detect [4].

Biological signals include eye blinking, heartbeat, and head tilting, which are often strong tells that the persona featured in the video is not real. For example, blinking behaviors in humans occur unconsciously and repeatedly. Even though eye blinking is a normal phenomenon, factors such as gender and time of day influence the rate at which it occurs [17]. Humans are pretty good at detecting anomalies in behavior like odd head tilting and off pattern blinking and can, in many instances, detect the authenticity of a video clip using the naked eye. Human aided interventions for deepfake detection can help to assist automated techniques.

We are interested in whether the viewer's implicit or explicit biases help or hinder their ability to detect the authenticity of the deepfake video. Moreover, if bias hinders their ability to detect deepfake videos, does heterogeneity in an online social network helps a community protect themselves from misinformation at a group level?

There are several automated deepfake detection methods [11, 18, 30, 32]. However, as deepfakes become ubiquitous, it will be important for the general audience or viewers to identify deepfakes independently. Also, several issues currently hinder automated methods: 1) they are computationally expensive; 2) there is quite a lot of bias in deepfake detection software and training data - credibility assessments, particularly in video content, have been shown in recent work [14] to be heavily biased; 3) As we have seen

with many cybersecurity issues, there is a cat and mouse evolution that will leave gaps in detection methodology, humans can help fill these gaps. However, we wonder to what extent human biases impact the efficacy of detecting diverse misinformation. If human-aided deepfake detection becomes a reliable strategy, we need to understand the biases it imposes. We also acknowledge that insights into human credibility assessments of deepfakes could help develop more lightweight and less computationally expensive automated techniques in the future.

2 COMMON PARAMETERS ACROSS MODELS

We present three models for this project. As a starting point we developed some common parameters that we could use as foundations for our systems. Here are the parameters that we will be using for most models:

- Initial duped: 1
- N = 1,000
- Network: Mixed Membership Stochastic Block Model or Stochastic Block Model
- Block Sizes = see sizes below for system type
- Block Probabilities = see probabilities below for system type
- Pickiness: the bias against the other group. The pickier a person is about the other group, the more myopic they are to their in-group. The higher their bias, the easier it is for them to get duped by a deepfake of the other group.

Category 1: everyone is equal

Sizes: $\begin{pmatrix}75 & 75 & 75 & 75\end{pmatrix}$

Probabilities: $\begin{pmatrix} 0.2 & 0.001 & 0.001 & 0.001 \\ 0.001 & 0.2 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.2 & 0.001 \\ 0.001 & 0.001 & 0.001 & 0.2 \end{pmatrix}$

Category 2: male has core-periphery

Sizes: $\begin{pmatrix}40 & 20 & 80 & 80 & 80\end{pmatrix}$

Probabilities: $\begin{pmatrix} 0.3 & 0.01 & 0.001 & 0.001 & 0.001 \\ 0.01 & 0.2 & 0.001 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.2 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.001 & 0.2 & 0.001 \\ 0.001 & 0.001 & 0.001 & 0.001 & 0.2 \end{pmatrix}$

Category 3: white mainstream

Sizes: $\begin{pmatrix}120 & 120 & 30 & 30\end{pmatrix}$

Probabilities: $\begin{pmatrix} 0.3 & 0.05 & 0.001 & 0.001 \\ 0.05 & 0.3 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.2 & 0.001 \\ 0.001 & 0.001 & 0.001 & 0.2 \end{pmatrix}$

3 THE PROBLEM

As we mentioned above, this project focuses on the following research question: how does homogeneity and heterogeneity of users in a network community help or hinder the spread of deepfake videos in an online social network (OSN)? Moreover, how big of an attack would be needed for the deepfakes to dupe most users?

To address this problem, we will simulate the introduction of a deepfake adversarial campaign on an online social network and observe the behavior of users as they interact with the videos. In our model, users have a variety of possible attributes that represent a range of demographic diversity and intersectionality. The range and number of attribute types will vary depending on the model type, real-world system, and population of interest. Users will then have probabilities of connecting to other users, being duped by the videos, being informed and not getting duped for the videos, or being susceptible and not seeing the videos.

In our models, we are intentionally trying to understand the broader structure of the system, the dynamics of the system, and the parameters independently so that we can more broadly understand the mechanism and more flexible compare systems from various online social networks using real-world data.

4 NETWORK MODELS OF DIVERSE MISINFORMATION

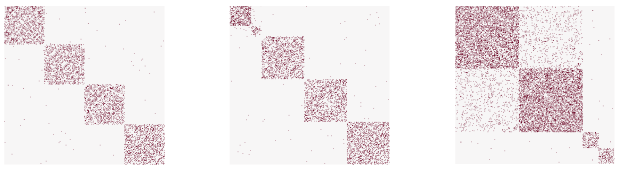
Our first two models will be structured using a Stochastic Block Model (SBM) and a Mixed Membership Stochastic Block Model (MMSBM) and will look at the spreading dynamics of diverse misinformation via a voter model and a contagion model. We initiate three distinct networks in each model by creating an SBM and MMSBM with the above common parameters.

4.1 Background: Stochastic Block Models and Mixed-Membership Stochastic Block Models

Edge probability depends on group memberships instead of connecting node pairs with equal probabilities. **Stochastic Block Model:** A stochastic block model (SBM) is a generative model that generalizes the Erdos-Renyi model to have groups. SBM is a random Poisson graph model in that node degrees within any group are distributed according to a Poisson distribution[23]. The SBM is widely used in complex systems because of its ability to generate different types of network structures, e.g., core-periphery, community structure, hub-and-spoke, etc. Not only is the SBM flexible, but it is also extensible, as we can see below with the Mixed Membership SBM.

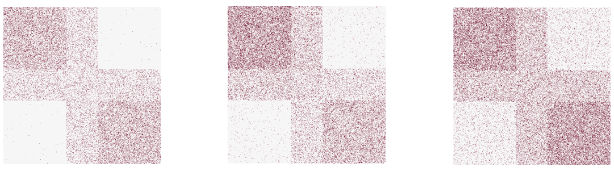
We take advantage of the SBM's flexibility by using three structures as our initial conditions. We refer to these three conditions as (1) 'everyone is equal, (2) 'male has core-periphery,' (3) 'white mainstream' (see Figure 2 for results of the SBM). Note that we have four groups or blocks for each condition to potentially represent different features of a population, here gender and ethnicity. We assume binary gender (male and female) and ethnicity (white and people of color) for simplicity. We hope to include non-binary gender in future work once we have empirical data with gender identity self-reporting. Conditions are created using a matrix of probabilities that specify the degree to which nodes are connected within-group (probabilities on the diagonal) and across groups. In the first condition, we have equal probabilities both within-group ($p_{rr} = 0.2$) and across groups ($p_{rs} = 0.001$). In the second condition, we add a subgroup in the periphery of the white male population. We might hypothesize that this particular subgroup has different biases when consuming deep than the main population. The third condition is analogous to the second one but seeks to represent a minority of people of color against the mainstream of a white population. We created our SBM using NetworkX SBM module.

Mixed Membership Stochastic Block Model: A Mixed Membership Stochastic Block Model (MMSBM)[3] is a Bayesian method of community detection which segments communities into blocks



(a) Everyone equal (b) Male Core-Periphery (c) White Mainstream

Figure 2: from left to right: (a.) Everyone is Equal, (b.) male has core periphery, (c.) white mainstream Stochastic Block Model (N=1000)



(a) Everyone equal (b) Male Core-Periphery (c) White Mainstream

Figure 3: from left to right: less coupling to more coupling in a Mixed Membership Stochastic Block Model (MMSBM) (N=1,000). (a.) Everyone is Equal, (b.) male has core periphery, (c.) white mainstream Mixed Membership Stochastic Block Model

but allows community members to mix with other communities. Assumptions in a MMSBM include a list of probabilities that determine likelihoods of communities interacting. Unlike the Stochastic Block Model explained above a MMSBM allows nodes to belong to multiple communities, can have multiple strengths of membership, adding an extra layer of parameters to the network. MMSBM can either be used to be predictive and identify unknown communities through their patterns of interaction or interpretive where they are used to help understand known communities and how they interact and overlap with each other.

In our project, We use the MMSBM because the system we are interested in (this is something we will consider integrating in future work) real world social networks which are a heterogeneous and are systems where there are overlapping communities. We created our MMSBM in the following manner:

- σi = the label of node i ∈ 0,1 (this vector represents a node's presenting identity and an interest in other identities given their identity)
- i,j = are node IDs ∈ 0,1,2...N-1 (these scalar are the number of sets of blocks in the group, white people, people of color, men, women etc.)
- probabilities (p) = [p1, p2, p3, p4] (this matrix represents the probabilities of blocks interacting)
- Mixed Membership Stochastic Block Model Pseudo Code:
- for i in range(N):
- for j in range(i + 1, N)
- bi = random choice (σ:) (e.g. [[0,0], [0,1],[1,0],[1,1]])
- if Random choice < p[σ1, σ2]
- A[i,j]=1 → add edge (i,j)

In Figure 3 we show preliminary results of the MMSBM for online social media communities that allow for less and more coupling between blocks. Here you see in the most left sub-figure, all coupling must go through the bridge nodes, and a you can see in the furthest sub-figure on the right, coupling can flow from any of the blocks.

4.2 Contagion Model

As in the SIR model, we assume that nodes in our contagion model can be in one of three states; they can be informed, duped, or recovered. We assume that the contagion unfolds in discrete steps where each step represents the possibility that a node i "infects" neighbor j by sharing a deepfake video. As in traditional contagion models, there is a probability that the transmission will succeed, and a duped user will recover. Our model loops over the nodes sequentially for a given number of steps or until the model converges to equilibrium.

Our model differs from the traditional SIR model in two important ways. First, the content propagates on the network concerns either its in-group or the other group. We represent the (deepfake) content in the form of a node attribute as if nodes create a memory of the deepfake they watched and want to share. The transmission probability is then represented by a bias, which can take different values depending on the content of the deepfake being shared. The assumption here is that people will be more skeptical about deepfake from the other groups than their own group. Although we assume an ethnicity bias, our model could be easily extended to other biases such as political ideology or socioeconomic status. We also assume that how uncritical someone is of its own group, or blindness, correspond to blindness = 1 − SKEPTICISM. The second key difference is that we assume duped individuals can recover with a certain probability only if they have a certain fraction of their neighborhood that is informed. The assumption is that having an informed neighborhood might educate duped individuals, thereby helping them recover from being duped by a deepfake video. The form of this recovery mechanism is: $\alpha_0 + \alpha(\frac{\#\text{I neigh}}{k})$, where α_0 is the probability of spontaneous recovery, and α is the social influence weighted by the fraction of informed people around you. In our results, we look at both how social influence influence population recovery with and without spontaneous recovery.

The key parameters of our contagion models are the following:

- Pickiness: the bias against the other group. The pickier a person is about the other group, the more myopic they are to their in-group. The higher their bias, the easier it is for them to get duped by a deepfake of the other group.
- fraction informed: the fraction of neighbors needed to have a chance to recover.
- α_0: the probability of spontaneous recovery.
- α: the probability of recovery by social influence.

Combined with different initial network structures, these parameters allow us to study the relationship between group bias and recovery in the context of a deepfake video outbreak. We now turn to the initial results of the contagion model.

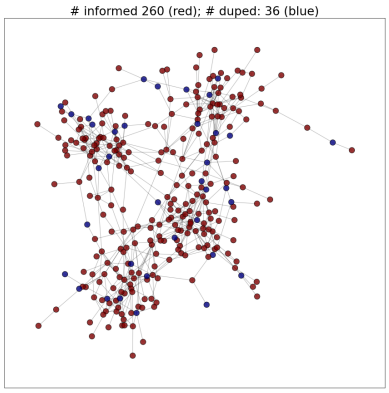


Figure 4: Initial network for the results below. In red we have the initial duped population, in blue the informed population.

In Figure 4, we can see the initial network we use for the subsequent results. It is an "everyone is equal" structure, in which we have rewired within groups so that the degrees are distributed according to a power law. As a first step, we can look at the long-term behaviors of the different populations:

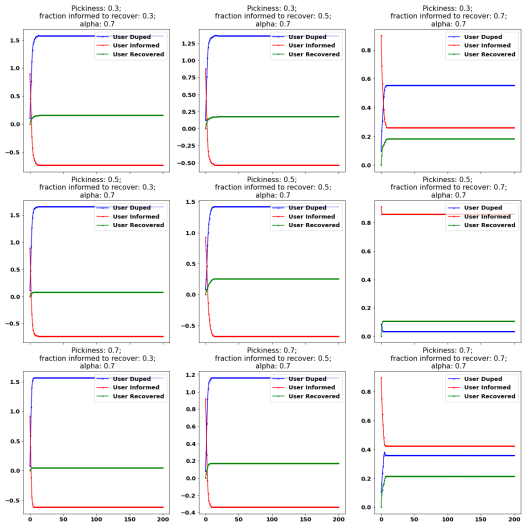


Figure 5: This figure shows the proportions of duped, informed and recovered users throughout the simulation. Parameter sweep of the contagion model. On the x-axis, we have Pickiness at $[0.3, 0.5, 0.7]$. On the y-axis, we have fraction of informed at $[0.3, 0.5, 0.7]$

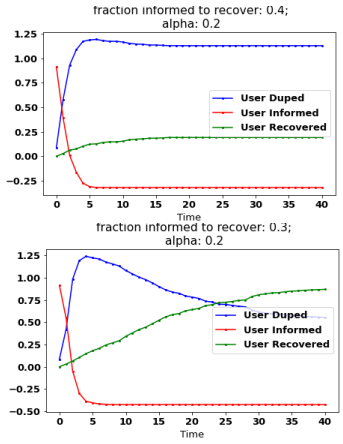


Figure 6: Fixing the bias at .7 and alpha 0.2, we can examine the necessary fraction of informed users that is necessary for the majority to recover. We can see that if users require more than $1/3$ of their surroundings to be informed to recover, not everyone recovers.

Instead of going into the details of all the different panes, we picked one example that we think is promising for further investigation. In Figure 5, we can see that our parameters have strong effects on the ultimate proportion of users we have in each state. Starting from the first row and moving from left to right, we can see that requiring a larger fraction of informed neighbors to recover seems to bring closer together the population, with informed people taking over recovered people (be careful, you have to be wary of the y-axis because it varies freely from one facet to another).

In Figure 6, we see a threshold effect from which the majority seems to be able to recover from the deepfake epidemic. We note that if users need more than $1/3$ of neighbors to recover, the population remains primarily duped by deepfakes. Taking this result at face value, we might hypothesize that one way to counteract a deepfake epidemic is to ensure that informed people are evenly distributed on the network and not concentrated in any particular group.

4.3 Voter Model

The Voter Model is a simple mathematical model of interacting agents in a system formulated by Richard A. Holley, and Thomas M. Liggett [16]. A voter is one node in our network, and edges between networks assume some interaction between connected nodes. In the voter model, we pick uniformly random nodes in the system, which we call listeners, which speakers' opinions can influence their network neighborhood. In our simple model, voters can either be in the state blue or red where these colored states represent duped or informed. Our initial network is based on a Mixed-Membership Stochastic Block Model (MMSBM)[3] with specific probabilities of blocks interacting with each other. At every time-step (t), we pick a random node, and they will adopt the state of neighbor at some probability p. Our model runs until the system reaches the maximum time steps or consensus.

Ideas on how we can extend the model:

- In future work, we would be interested in expanding the model, inspired by Sid Redner's *Reality Inspired Voter Model Overview* [25], we could implement levels of confidence in voter's opinions.
- We are also working on integrating a probability into the model that dictates the relationship between the individual's bias levels and how likely they are to be initially duped.
- We see in the MMSBM that the bridge nodes between communities seem like an ideal population to look at in more detail to answer our question and compare with more peripheral nodes. One question we would like to investigate further is the susceptibility of bridge nodes in the MMSBM because they are connected to opinions of many more blocks, are they more or less susceptible, and perhaps do they suffer a burden of being gatekeepers [8].

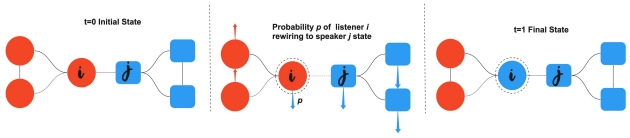


Figure 7: Schematic for voter model

4.4 Voter Model on a Mixed-Membership Stochastic Block Model Initial Results

In this model, we look at the relative spread of duping in the network based on the structure of the network communities' probabilities of interacting with other communities in an MMSBM. The MMSBMs allows for less and more coupling between blocks. Here in 8 you see in the most left MMSBM sub-figure, all coupling must go through the bridge nodes, as seen in the furthest MMSBM sub-figure on the top right, the coupling can flow from any of the blocks. Here we vary heterogeneous and homogeneous mixing and compare the speed at which duping spreads through the system. As seen in Figure 8 the network with the most coupling reaches consensus easier because there is more access to connect to the bridge nodes from other blocks and communicate the vote throughout the system. We run this voter model to see the impact of simple spread dynamics on an MMSBM and to view the role of bridge nodes in this process. from left to right: less coupling to more coupling in a Mixed Membership Stochastic Block Model (MMSBM) (N=1,000). In Figure 8 from second row top to bottom Voter Model results in the probability of being duped=0.3,0.5,0.7 averaged over ten runs and 1e6 iterations show who wins and reaches near consensus in the system. Blue=duped and red=informed.

Degree Distribution for the three voter model MMSBMs:

The average degree is used to measure the connectedness of a network. **Voter Model Network on 1: for the low coupling** network the average degree for this social network is 7.568, which is the total of all the node's degrees divided by the number of nodes in the network. Compared to the other hypothetical social networks, this network has a relatively high level of connectivity.

The average degree is used to measure the connectedness of a network. **Voter Model Network on 2: for the medium allowable coupling** network the average degree for this social network is 4.577, which is the total of all the node's degrees divided by the number of nodes in the network. Compared to the other hypothetical social networks, this network has the lowest level of connectivity.

The average degree is used to measure the connectedness of a network. **Voter Model Network on 3: for the most coupling** network, the average degree for this social network is 5.310, which is the total of all the node's degrees divided by the number of nodes in the network. Compared to the other hypothetical social networks, this network has the median level of connectivity.

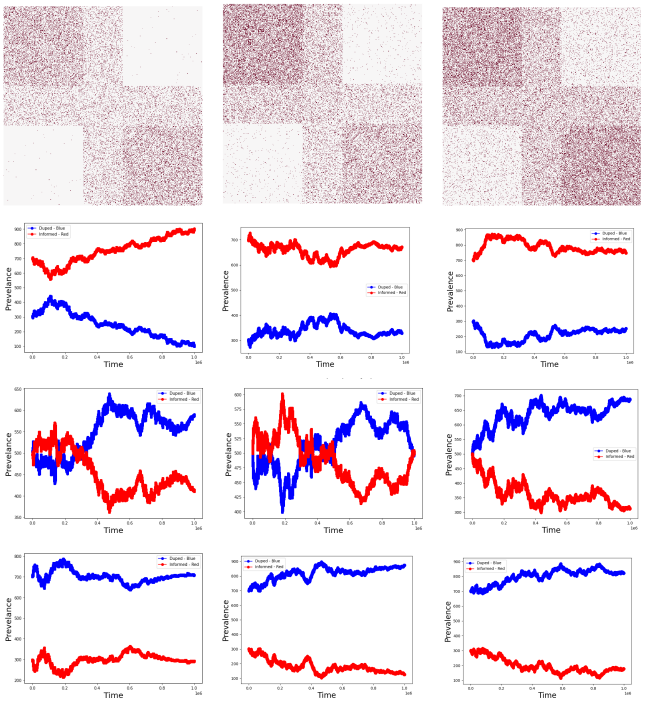


Figure 8: from left to right: less coupling to more coupling in a Mixed Membership Stochastic Block Model (MMSBM) (left column a.) Everyone is Equal, (middle column b.) male has core periphery, (right column c.) white mainstream (N=1,000). From second row top to bottom Voter Model results probability of being duped=0.3,0.5,0.7 averaged over 10 runs and 1e6 iterations shows who wins and reaches near consensus in the system. Blue=duped and red=informed.

5 AGENT-BASED MODEL OF DIVERSE MISINFORMATION

5.1 Methodology

For our agent-based model (ABM), we used Mesa and NetworkX to adapt an existing SIR ABM to track the spread of DeepFake videos as a virus spreading through a population. Each simulation of our

model (Figure reffig:ABMSchematic) follows a single DeepFake video as it spreads through a social network. We chose a network as our environment since this is a more accurate representation of the arrangement of the relationships in an online social framework (e.g., Twitter, Facebook) than structures such as the square grid of a cellular automata (CA) model, and to build on our work from the two network models described previously in this report. The agents are the people in the social network. In this ABM, the structure of the network is static. However, the agents move around the network to simulate the fluctuation of real-life online social platforms, e.g., Facebook users friending new people and unfriending others. Since we are primarily interested in whether the interaction between race and gender of the deepfake viewer and those of the person portrayed in the video affects the spread of the video, the agents have a race and gender attribute. For simplicity's sake, we have reduced the race categories to white and non-white and the genders to male, female, and non-binary. To reduce the number of agent types in the model (since each simulation tracks a single video), we chose to include the race and gender of the persona in the deepfake video as attributes of the viewer (video_race and video_gender). There are four states of the agents, which change as they interact with neighbors: Naïve agents have not watched the video, duped agents believe the DeepFake is true and may spread it to their neighbors, neutral agents are undecided about the video and neither spread nor flag it, and informed agents know the video is fake, do not spread it, and may flag it.

The model is initialized by setting a network structure and assigning all agents a Naïve status, then randomly changing their status to duped based on a tunable probability initialDupedProb. These agents represent the first people in a social group to watch a DeepFake video or perhaps the creators of said video. In each step, the model iterates over all the agents in the model. First, all the agents move to an adjacent node. Next, any neighbor of a duped agent who has watched a video (decided in the previous iteration) decides regarding the video's veracity. Agents may watch the video an unlimited number of times and change their opinion unlimited times (Figure 9 A-B). We have added tunable parameters that can be used to adjust the probability of changing from a duped or informed state, p_StayDuped and p_StayInformed, respectively. Viewers believe a DeepFake video with probability finalPDuped, which is calculated by multiplying a tunable baseline pDuped by tunable modifiers that are applied when the viewer's gender or race matches that of the video's character (Figure 9 C). Duped agents attempt to spread the video by posting it on their social media page, where it is watched by any of their friends with probability pWatch. These agents will then choose their opinion state in the next step. If the video does not dupe the viewer, they will either become neutral with probability pNeutral, or informed with probability 1-pNeutral. Neutral agents do not engage any further with the video. Informed agents flag the video with probability pFlag. When the video meets a tunable threshold of flags, the video is taken offline, and the video stops spreading (Figure 9 A-B).

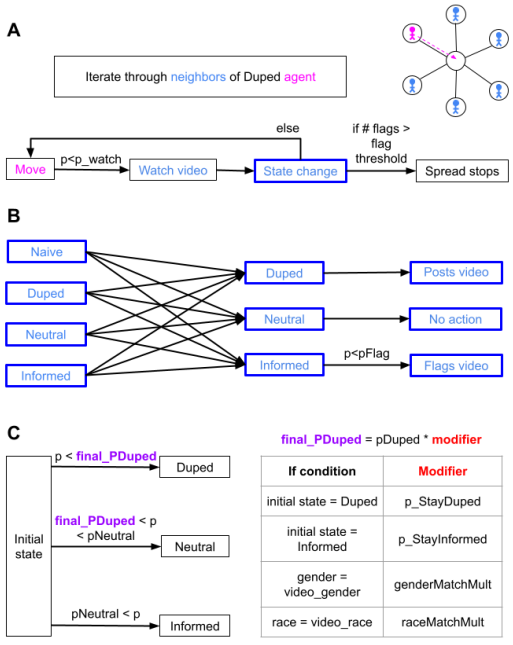


Figure 9: Diverse misinformation Agent Based Model Schematic

5.2 Results

We ran many simulations with our agent-based model using different model interactions, initial conditions/initial network, and different parameters. Shown in this report are 4 of our simulation runs, where only two model parameters were varied in order to see the effect these two parameters had on the model: raceProportions, the proportions of people of each race among the agents, and videoGender, the presenting gender of the individual in the deepfake video that was spreading. Each simulation has seven different output graphs, where the graphs include a total graph that displays the states of all agents and then six different breakdowns of subsections of the agent's population by gender and race. This allows us to pinpoint the effects of changing parameters in the model concerning gender and race.

We observed several general trends in our graphs that will be recorded here before analyzing the specific figures we have chosen to include. For example, varying the probability of becoming duped by a video after watching that video changes the states of nodes causing many more to become duped than informed or vice versa. The level of neutral increases or decreases proportionally with informed. Agents who are naïve to a video decrease at the same rate (this is related to pWatch), except in the case where the flags threshold is reached, and then spread stops because the video is removed. The curves are flattened as the flags threshold is increased. The spread of duped videos is faster and can dupe more people in the final step before reaching the flag threshold when the probability

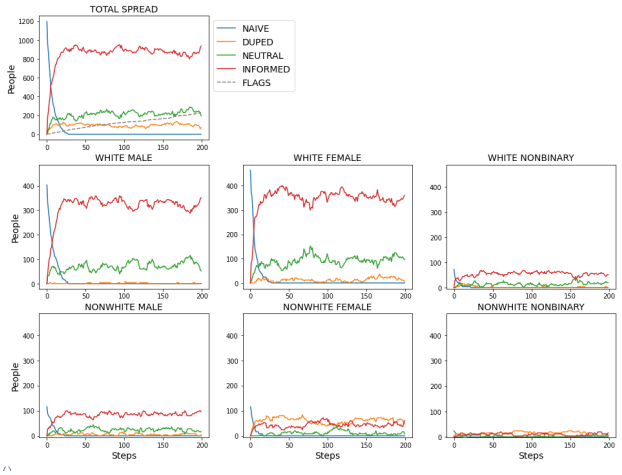


Figure 10: ABM Simulation with raceProportions = (0.8, 0.2), genderProportions = (.45, .4, .1), videoGender = MALE, videoRace = WHITE, genderMatchMult=.1, raceMatchMult=0.1

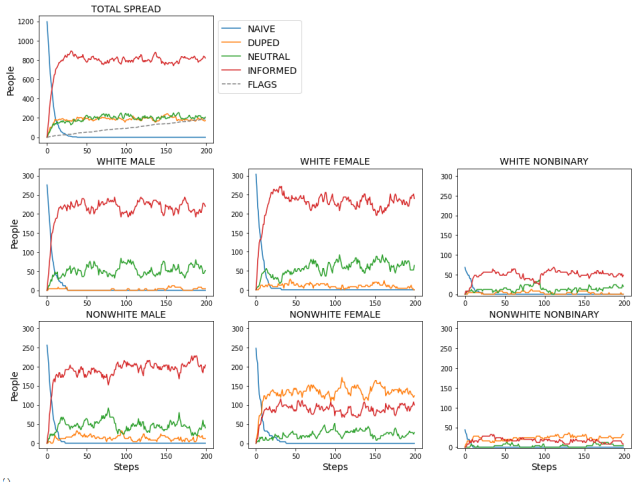


Figure 11: ABM Simulation with raceProportions = (0.5, 0.5), genderProportions = (.45, .4, .1), videoGender = MALE, videoRace = WHITE, genderMatchMult=.1, raceMatchMult=0.1

of watching the video increases. Because varying these parameters caused predictable outcomes in our simulations, we chose not to go into more detail about them here. For the following graphs, our initial parameters that remained unchanged were as follows:

#OfNodes	genProbs	networkGeneration	flagsThreshold	steps	initialDupedProb
1200	allEqual	SBM	100	400	0.01
pWatch	pDuped	pFlag	pStayDuped	pStayInformed	pNeutral
0.01	0.5	0.03	1.5	0.5	0.2

Figure 12: initial parameters that remained unchanged

Figures 10 - 14 show Deepfake spread simulations using different sets of parameters. The top-left graph shows the total spread among the population. The six graphs in the bottom two rows show the opinion states of the subpopulations grouped by gender and race. The total number of people is kept at 1,200. For Figure 10, it can be seen that the number of informed users always stays high as compared to "NAIVE", "DUPED ","NEUTRAL" and the persons who flag the videos. The six graphs below show the distribution of the states among the white or non-white people. Here it can be seen that keeping the race proportion [0.8, 0.2], video gender as "MALE" and video race as "WHITE, the white male, female, and non-binary get more informed. Also, the number of non-white males seems to be more informed about the fake videos than the non-white females, which seem more duped. Similarly, in 11, keeping the race proportion [0.5,0.5], shows us that the ratio of getting informed and duped are almost the same as the previous graph because the gender and race of the video are still the same. In figure 13 , we changed the race proportion from [0.5,0.5] to [0.2,0.8], which means there are 20 percent white people and 80 percent of non-white people. In this case, we can see the rise in the number of informed non-white males, but non-white-females still have high numbers of duped. In the last figure 14, The race proportion is changed again to 0.8 white and 0.2 non-white people, but here we also changed the gender of the video to non-binary. These changes resulted in the rise in the number of duped non-white males compared to the rest of the population. In all these scenarios, the two parameters, i.e., "genderMatchMult" and "raceMatcMult," are kept the same at 0.1. This parameter is used to create biases towards the states. For example, if the gender of the video is the same as the person, we multiply it with the person's probability of getting duped, pDuped. Similarly, if the deepfake video's race is the same as the race, the raceMatchMult is multiplied.

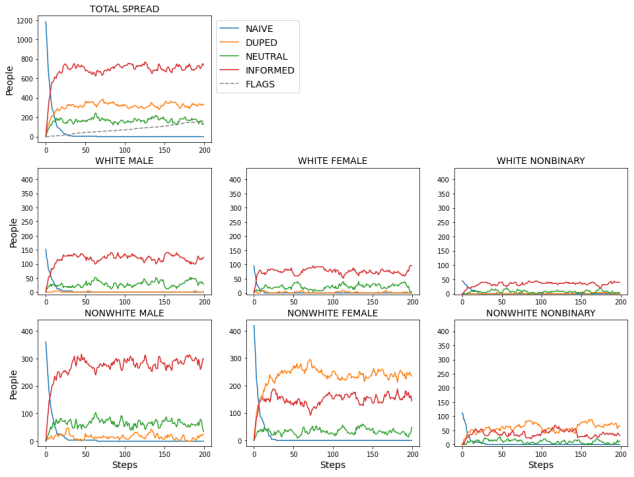


Figure 13: ABM Simulation with raceProportions = (0.2, 0.8), genderProportions = (.45, .4, .1), videoGender = MALE, videoRace = WHITE, genderMatchMult=.1, raceMatchMult=0.1

Notably, varying genderProportion has the same effect on our simulation as varying raceProportion, as these two parameters are

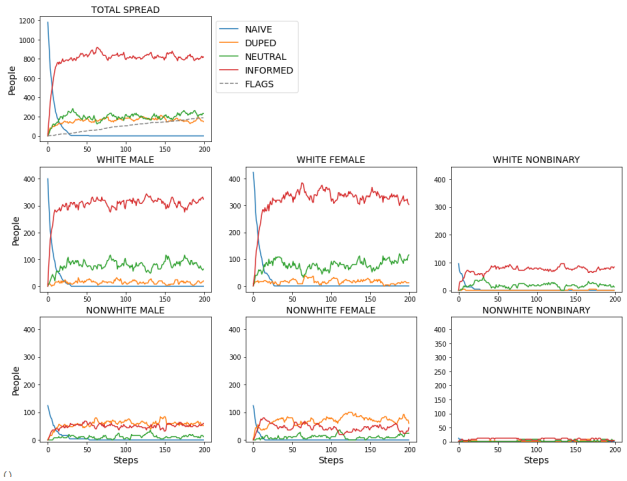


Figure 14: ABM Simulation with raceProportions = (0.8, 0.2), genderProportions = (.45, .4, .1), videoGender = NONBINARY, videoRace = WHITE, genderMatchMult=.1, raceMatchMult=0.1

set up in the same way in our code; the levels increase or decrease proportionally to which videos they match.

6 DISCUSSION

The three models that we explored in this project helped us better understand different parts of our system's structure and dynamics. They also help describe and visualize the system in different ways; for example, the ABM and Voter model is easy to understand for a general audience and may be helpful in science communication about deepfake spreading. Overall, we found the contagion model to be the most illuminating on the system's dynamics. However, in future work, we would like to explore additional mathematical models that can help inform the contagion model. We currently do not have empirical data to inform these models, and we initially created the models to be data-independent. However, we would like to integrate real-world data into our models and possibly compare results across social media platforms with different community structures in future work. Below we outline a summary of each of the model's strengths, weaknesses, and insights:

Pros and cons of the contagion model and what it told us about the system:

The contagion model allows us to study the impact of group and recovery biases on the propagation and convergence of deepfakes in a population. The bias mechanism we implemented revealed a threshold effect in the propagation of deepfakes eg. if you need too many informed people around you to recover, not everyone is recovring. This result holds when a given population has relatively strong biases against the other group. (β = pickiness = .7). The network component of our social contagion model is also interesting because we can examine the impact of network structure on the propagation of deepfakes. In our case, we note that when a population does not recover completely, it is because of a clustering effect among the populations. That being said, one of the downsides of the contagion model is that the analogy breaks down when it comes to the process of transmission, that is, people do not spread deepfakes on contact. For example, someone in Malaysia might assign someone in Quebec if they both share the same interest. Deepfakes are often shared via social media, which have their own peculiarities that we do not discuss here. Also, the spreading process does not reflect the varieties of deepfakes present in reality. Often when we talk about diseases we are talking about a few types of pathogens or viruses. In our case, deepfakes are cultural artifacts, constantly copied and reinvented at a phenomenal rate.

Pros and cons of the MMSBM Voter model and what it told us about the system

In the voter model, we look at the relative spread of duping in the network based on the structure of the network communities' probabilities of interacting with other communities in an MMSBM. The MMSBMs allows for less and more coupling between blocks. We saw from these results that bridge nodes in the MMSBM play a noticeable role in spreading diverse misinformation. In this model, we varied heterogeneous and homogeneous mixing and compared the speed at which duping spreads through the system. As seen in Figure 8 the network with the most coupling reaches consensus faster because there is more access to connect to the bridge nodes from other blocks and communicate the vote throughout the system. We ran this voter model to see the impact of simple spread dynamics on an MMSBM and to view the role of bridge nodes in this process. For the initial process of examining the speed of attacks, the voter model seems insightful. However, this is a simple model that does not show us much about the overall dynamics of our communities.

Pros and cons of the ABM model and what it told us about the system

We noticed several interesting features regarding DeepFake video spreading from the simulations conducted with the ABM. First, the system reaches a dynamic equilibrium quickly, by the 20th step for all the subpopulations the simulations discussed in this report. The system's dynamics can be accelerated or decelerated by changing the probability of neighbors watching the video (p_watch) and stopped very quickly by setting a low flag threshold and a very high probability of becoming informed and flagging the video. Next, the first three simulations represent a small parameter sweep of the race proportions of the agents in the model (Figures 10 - 14). The small effect of the system's race proportions and the gender of the video on the final number of duped, neutral, and informed populations suggests that the outcome is more dependent on the initial pDuped parameter and its modifiers (p/StayDuped, genderMatchMult, and raceMatchMult). Because the model itself has many different parameters to start, changing any given parameter will not, on average, make a large change in the model output (unless it is a key parameter such as pDuped). The key insight of the model is that though the final sizes of the duped, neutral, and informed populations for the overall system may look similar for different simulations, the final results may look vastly different when grouping agents by race and gender. We found such differences both when comparing subgroups in the same simulation and the same subgroup across simulations differing by a single parameter. For example, the nonwhite male and female populations in Figure 13 have very different outcomes due to the effect of the raceMatchMult modifier on the probability of being duped.

Additionally, the nonwhite male population has vastly different opinion states for simulations shown in 10 and 14 where the only difference is the gender of the video. The differences here are due to the genderMatchMult modifier.

Our model's most prominent and polarizing aspect is its inherent flexibility in construction. Flexibility allows for minor adjustments and tuning of the model and incorporates an unlimited number of model parameters or networks that the simulation is run on. This is an excellent boon for attempting to model an issue with many confounding variables that are relatively unexplored in their impact. However, it is also a bane in that we do not currently have real-world data to back up the tuning of our model. For example, we do not know that the effect of a match in gender between a viewer and the perceived gender of the deepfake persona will result in the probability of the viewer becoming duped decreasing by 90%. This uncertainty makes taking away actionable conclusions from our model difficult and leads to sparser conclusions.

ACKNOWLEDGMENTS

The authors would like to thank Professor Laurent Hébert-Dufresne for discussions and feedback on this project.

REFERENCES

[1] Dor Abrahamson and Uri Wilensky. 2005. ProbLab goes to school: Design, teaching, and learning of probability with multi-agent interactive computer models. In *Proceedings of the Fourth Conference of the European Society for Research in Mathematics Education. San Feliu de Gixols, Spain.*

[2] Saifuddin Ahmed. 2021. Who inadvertently shares deepfakes? analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics* 57 (2021), 101508.

[3] Edoardo Maria Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. 2008. Mixed membership stochastic blockmodels. *Journal of machine learning research* (2008).

[4] Bobby Chesney and Danielle Citron. 2019. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *Calif. L. Rev.* 107 (2019), 1753.

[5] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020. How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. arXiv:2008.11363 [cs.CV]

[6] Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2020. Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics* (2020), 1940161220944364.

[7] Don Fallis. 2020. The Epistemic Threat of Deepfakes. *Philosophy & Technology* (2020), 1–21.

[8] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*. 913–922.

[9] graspologic tutorials. 2020. Mixed Membership Stochastic Blockmodel (MMSBM). https://microsoft.github.io/graspologic/tutorials/simulations/mmsbm.html.

[10] Samuel Greengard. 2019. Will deepfakes do deep damage? *Commun. ACM* 63, 1 (2019), 17–19.

[11] David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. IEEE, 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–6.

[12] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances* 5, 1 (2019), eaau4586.

[13] Douglas Harris. 2018. Deepfakes: False pornography is here and the law cannot protect you. *Duke L. & Tech. Rev.* 17 (2018), 99.

[14] Kurtis Haut, Caleb Wohn, Victor Antony, Aidan Goldfarb, Melissa Welsh, Dillanie Sumanthiran, Ji-ze Jang, Md Ali, Ehsan Hoque, et al. 2021. Could you become more credible by being White? Assessing Impact of Race on Credibility with Deepfakes. *arXiv preprint arXiv:2102.08054* (2021).

[15] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social networks* 5, 2 (1983), 109–137.

[16] Richard A Holley and Thomas M Liggett. 1975. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability* (1975), 643–663.

[17] T. Jung, S. Kim, and K. Kim. 2020. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* 8 (2020), 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

[18] T. Jung, S. Kim, and K. Kim. 2020. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* 8 (2020), 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660 Conference Name: IEEE Access.

[19] Elizabeth F. Loftus and Edith Greene. [n.d.]. Warning: Even memory for faces may be contagious. 4, 4 ([n. d.]), 323–334. https://doi.org/10.1007/BF01040624

[20] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[21] Melanie Mitchell, James P Crutchfield, and Peter T Hraber. 1994. Evolving cellular automata to perform computations: Mechanisms and impediments. *Physica D: Nonlinear Phenomena* 75, 1-3 (1994), 361–391.

[22] Masahiro Mori. 2017. The uncanny valley: The original essay by masahiro mori. *IEEE Robots &* (2017).

[23] Mark Newman. 2018. *Networks*. Oxford University Press.

[24] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. 213–222.

[25] Sidney Redner. 2019. Reality-inspired voter models: A mini-review. *Comptes Rendus Physique* 20, 4 (2019), 275–292.

[26] Kevin Roose. 2018. Here come the fake videos, too. *The New York Times* 4 (2018).

[27] Gary T Schwartz. 1990. Explaining and Justifying a Limited Tort of False Light Invasion of Privacy. *Case W. Res. L. Rev.* 41 (1990), 885.

[28] Seth Tisue and Uri Wilensky. 2004. Netlogo: A simple environment for modeling complexity. In *International conference on complex systems*, Vol. 21. Boston, MA, 16–21.

[29] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131–148.

[30] L. Verdoliva. 2020. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101

[31] C Yang and U Wilensky. 2011. Netlogo epidem basic model. *Center for Connected Learning and ComputerBased Modeling Northwestern University Evanston IL, pages Evanston, IL* (2011).

[32] Sergey Zotov, Roman Dremliuga, Alexei Borshevnikov, and Ksenia Krivosheeva. 2020. DeepFake Detection Algorithms: A Meta-Analysis. 2020 2nd Symposium on Signal Processing Systems, 43–48.

7 APPENDICES

APPENDIX I: CODES

- code for SBM basic Contagion Model
- code for SBM Fancy Contagion Model
- code for MMSBM Fancy Contagion Model
- code for MMSBM Voter Model
- code for Agent Based Model

8 APPENDIX II: TEAM ROLES & TIMELINE

8.1 Team Roles

- **Project management:** Juniper
- **Literature review:**
  - Background on deepfakes on online social networks: Juniper, Jonathan, Gabriela, Ijaz, Alex
  - Background on network model methodology and spreading: Juniper and Jonathan
  - Background on agent based model methodology: Alex, Ijaz, Gabriela
- **Models:**
  - Network Model: Juniper and Jonathan
  - Agent Based Model: Alex, Gabriela, Ijaz
- **Visualization** of figures and results: Ijaz, Alex, Gabriela
- **Write final report** and prepare slides: All
- **Record final presentation:** All

8.2 Timeline of due dates

Team meeting times: weekly Sunday from 4-5pm on Teams

- **November 8 -** Complete Project proposal: All
- **November 10 -** Submit Proposal: Juniper will submit on blackboard
- **November 17 -** Complete Literature Review: All
- **November 21 -** Complete models: All
- **November 24 -** complete figures: All
- **November 24 -** Complete methodology and results section: All
- **November 28 -** Complete discussion section and abstract: Juniper
- **December 1 -** Submit first draft of written report: Juniper will submit on blackboard
- **December 4 -** Complete slides - All
- **December 5 -** Complete presentation: All - will record via zoom
- **December 8 -** Submit final presentation: Juniper will submit on blackboard
- **December 16 -** Submit final written report: Juniper will submit on blackboard

9 APPENDIX III: PROJECT PROPOSAL

Prepare a proposal and upload it to the Blackboard website by November 10th at midnight. The proposal should be at least half a page, describing the system or research question your team wishes to tackle. You should discuss what modeling approaches you are considering, and how much you want to achieve in your project. Finally, doing a brief literature search on the topic is highly recommended to find guidance in published work to inspire you

System: our team is working to model spread of deepfake videos on online social networks (e.g. Facebook).

Research Question: how does homogeneity and heterogeneity of users in a network community help or hinder the spread of deepfake videos in an online social network? How big of an attack would you need for the deepfakes to dupe a majority of users?

Modeling Approach 1 Network Model: Our project will model the spread of deepfake videos on a social network, the model will be inspired by an SIR model where users groups will be exposed to deepfake videos and can either take the state of being susceptible (cautious), Infected (duped), or Recovered (informed). Our network structure will be determined by a mixed membership stochastic block model (MMSBM) [9] which will allow probabilities of links between nodes to depend on node attributes. We assume that network neighbors share properties of homophily [20] and will have a tendency to connect to like attributes. The probability of connecting to like attributes will be tunable.

For our network model, we will create 5 distinct user groups on the social network (e.g. by gender (identifies as male, female, or non-binary) and BIPOC status (identifies as a person of color or not a person of color). We will also create distinct types of deepfake persona groups based on perceived demographics of the personas in the deepfake videos (e.g. by gender (user perceived them as male, female, or non-binary) and BIPOC status (user perceived them as a person of color or not a person of color). We will set initial conditions of the populations of users, populations of deepfakes, and deepfake birth (B) and deepfake death (D) rate for deepfake videos entering the and exiting the system.

Once the SBM initial conditions has been established we will release the deepfakes onto the system and observe the spreading dynamics on the network. We will assign tunable probabilities for the likelihood of users seeing deepfake videos (S). Based on which deepfake videos the users see, they will have a likelihood of ending up in 1 of 3 states, namely, cautious, duped, or informed. Once all deepfake videos have been seen we will observe the percolation of the videos in the system and measure which user groups end up in which state.

If we want to take the model to the next iteration, we could then update the system again by setting probabilities of users in states that are duped sharing the deepfake video at some probability and informed users flagging (killing) a deepfake video at some probability. We could then observe how long it takes for the deepfake videos to either die out or take over.

Modeling Approach 2 Agent Based Model: For our agent model we will be model a similar system and dynamic as the network model approach but add additional parameters. We will essentially replicate the network model as an ABM model. We will initially look at how the system looks with parameters zeroed out. Depending on time/how complicated coding is we can introduce/start varying parameters and see how/whether they change things. Such as rules for updating states of neighborhoods based on inter-dependencies between user groups using a majority rules type model approach [21].

What we want to achieve in our project: We hope to finish a basic model of the spreading dynamics of deepfake videos on an online social network. Later we would like to inform the model with various real world network data in order to measure the relationship between the structure (heterogeneity or homogeneity) of the system and the dynamics on the system (misinformation attack size and characteristics). We have decided to leave the model flexible for now so that we can tune parameters for various real world online social network systems later using empirical data.

Literature Search: We have run an initial literature search on deepfake videos (and generally some other misinformation spreading) on online social networks [2, 5, 11, 12, 18, 24, 30, 32], credibility and deception in videos and memory [14, 19, 29], and the ethical harms of deepfake videos [4, 6, 7, 10, 13].

For our model literature review will be conducted on stochastic block models [15], agent based models of spreading [1, 28, 31].