Dynamic Analysis of False Information Spread Over Social Media: 5G–COVID 19 Conspiracy Theory

Orkun İrsoy

Department of Industrial Engineering, SESDYN Lab, Boğaziçi University, Istanbul, Turkey

ABSTRACT

The spread of false information via online social networks is a critical societal issue with various potential harms. Although there are huge efforts both in research and application to mitigate this problem, it persists with increasing magnitude of results ranging from political manipulation to violent attacks. In our research, we built a causal simulation model to combine the existing accumulated knowledge in the literature and provide a formal model to evaluate the governing dynamics for the specific case of the viral spread of the 5G-COVID-19 conspiracy theory. The model makes use of both qualitative and quantitative data and successfully generates the observed dynamics for the 5g narrative. Initial results suggest that the dominance of believers in the active discussion on social media is overrepresented relative to the total population. Moreover, common mitigation strategies proposed in the literature such as limiting the interaction with believers of the misinformation often seem to produce worse outcomes for specific cases. In addition, scenario analysis suggests that corrective information campaigns by the authorities should be timed carefully to maximize the effectiveness of the policy. The current analysis presents several trade-offs while discussing the underlying reasons through posterior analysis. As further research, we plan to expand our analysis on the base model by inclusion of other user profiles, experiment with other mitigation strategies, and discuss the potential similarities and differences of our case with other types of false information dynamics.

Keywords: Misinformation, False information, Social media, 5G, COVID19, Coronavirus, Simulation, System Dynamics, Information diffusion

1. INTRODUCTION

In today's world, communication methods have shifted significantly toward digital communication. The vast majority of people use social media, including people of all ages and socioeconomic backgrounds. As a result, an increasing number of people are using social media to gather and disseminate information on a variety of topics, including critical information. According to Reuters, nearly two-thirds of adult Americans use social media as a source of news (Moon, 2017).

This new mode of communication offers numerous benefits, including the promotion of engagement and the reduction of barriers among people all over the world by providing an alternative to face-to-face socialization. Information spreads faster and to a larger audience on these social networks. However, because the content is created by users without any review

process, unlike traditional media, the content's validity cannot be verified. As a result, whether willingly or unwillingly, people may spread false information. Perhaps the most recent example demonstrating the potential harms of this phenomenon is the “infodemic” during the COVID-19 crisis, with results such as various ineffective and possibly harmful remedies, to outright rejection of the existence of the virus (Pennycook et al, 2020).

A recent example of such viral false information spread is 5G being one of the causes of COVID-19 or increasing its spread was. The debate over the topic quickly erupted in the United Kingdom, particularly on social media platforms. Although fact-checking organizations or experts falsified the concerns related to this link, corrections were insufficient to alleviate the concerns, resulting in 5G tower arsons in Birmingham and Merseyside, United Kingdom (Ahmed et al., 2020).

Given the seriousness of the repercussions of misinformation dissemination, a massive body of research is carried out to tackle various facets of the problem. Researchers from various fields attempt to comprehend the psychological and cognitive drivers underlying the phenomenon, analyze the data at hand to deduce why and how false information spreads, develop novel methods to detect such information, and develop mitigation strategies to combat misinformation.

Unfortunately, due to the complexity of the problem, mitigation measures used today are far from creating a structural solution but instead serve as symptom relief. Use of warning labels, which is one of the dominant tactics used by social media platforms, produce an "Implied Truth Effect" on unlabeled information (Penycook 2020) or may increase online traffic for the labeled content (Ingram, 2017). Fact-checking services attempt to verify the accuracy of the contents, although the rate of information production has increased far faster than the capacity of confirmation services has expanded (Penycook & Rand, 2019). Extensive data science research on false news detecting methods lays the door for the development of smart bots (Ammara, Bukhari, and Qadir, 2020).

Despite great efforts in both research and application, the failure of present mitigation techniques derives from the requirement for a dynamic systems approach. As a result, a systems perspective of the situation that combines current literature findings might identify potential leverage points and policy resistances to obtaining structural solutions to the problem at hand. In this regard, we argue that constructing a causal simulation model will constitute a theoretical basis to discuss the structural properties, derive practical insights, and experiment with different scenarios. As a specific case, we consider the aforementioned rumor spread regarding COVID-19 and 5G technology.

In the following sections, we provide a brief overview of both the general problem of misinformation spread and the COVID19-5G rumor, in particular, develop a System Dynamics model to gain insights into the problem, and discuss preliminary results that reveal the underlying structure and potential implications.

2. BACKGROUND

Although false information dissemination is not a contemporary phenomenon, as evidenced by the 'Great Moon Hoax' in 1835 (Pennycook, Rand; 2021), the availability of highly linked worldwide platforms in the present world, allow anybody to transmit information to millions of individuals in a matter of minutes (Kumar and Shah, 2018) further increasing the reach and severity of the problem. False information causes issues ranging from political manipulation of large groups of people (Varol et al., 2017) and stifling rescue efforts during a crisis to even a terror strike (Kumar and Shah, 2018). One such instance is the Pizza Gate conspiracy, which resulted in a person firing a gun at a neighborhood shop in response to reports of child trafficking (Morstatter, Carley, and Liu, 2019). Another example is Facebook's claim of voter tampering in the 2016 Presidential Election (Lazer et al., 2018). Given the gravity of the effects, the World Economic Forum has identified false information spread on digital platforms as one of the main challenges to society (Lee Howell et al., 2013).

Many definitions and classifications of false information exist in the literature from rumors to Fake News. One prevalent classification dimension is the intention of the agent is where “misinformation” refers to unintentionally spreading information whereas “disinformation” is intentional (Wu et al., 2019; Kumar and Shah, 2018; Caled and Silva, 2021). Another categorization is the knowledge-based differentiation i.e., whether the information is purely factual or opinion-based (Kumar and Shah, 2018).

The problem of false information spread on social media has various drivers, both at the individual and aggregate levels. At the individual level, various psychological and cognitive factors are thought to be effective, and many researchers are trying to find answers to questions: Do political motives drive susceptibility to misinformation? Does repeated exposure lead to higher susceptibility to false beliefs? Which cognitive processes are influential on vulnerability to misinformation, and how can we design better corrective messages? (Penycook and Rand, 2021; Caled and Silva, 2021; Lewandowsky et al., 2021; Chan et al., 2017) From a more holistic perspective, another line of research focuses on the properties of these social networks, such as whether preferential attachment in these networks forms echo chambers (repetitive exposure of specific information due to homogeneous social clusters), whether there are any structural reasons that make specific networks more susceptible to misinformation spread, and whether there are any distinguishing characteristics that differentiate the propagation dynamics of misinformation compared to other networks (Vosoughi, Roy, and Aral, 2018; Vicario et al., 2016; Zhao et al., 2020). A huge effort is put into misinformation detection with machine learning using either content or context-based cues to design early interventions (Wu et al., 2019). Finally, simulation studies act as a testing platform to test the effectiveness of various intervention strategies or develop novel hypotheses about the underlying mechanisms of the problem (Kauk, Kreysa, and Schweinberger, 2021; Lotito, Zanella, Casari, 2021; Ammara, Bukhari, and Qadir, 2020).

Perhaps the most recent and critical forms of misinformation are experienced during the COVID-19 outbreak. Because of the ambiguity surrounding the situation, misleading information swiftly disseminates across borders, including conspiracy theories, fictitious miracle cures, and

material that trivializes the infection (Bridgman, 2021). One such case that emerged in early January 2020 was the conspiracy theory suggesting a link between the installation of new 5G towers with the spread of the virus (Ahmed et al., 2020; Bruns, Harrington, and Hurcombe, 2020). Unfortunately, the spread of rumors did not solely become an instance of misinformation but the escalated panic yielded multiple attacks on 5G towers in the UK (Brewis, 2020; BBC, 2020).

Many researchers investigate different aspects of the "5G-COVID-19" conspiracy theory and its spread as it epitomizes the potential harms of viral misinformation. Ahmed and colleagues (2020) used Social Network Analysis to analyze the Twitter chatter during the peak time of the chatter. Their analysis reveals that the number of people genuinely believing the conspiracy is rather low compared to the volume of the tweets. They conclude that apart from the believers, anti-conspiracy tweets, click baits or satiric tweets also contribute as much as believers, which further increase the dissemination of the false information. Agley and Xiao (2021) conducted an online survey on the believability of various conspiracy theories including the 5G narrative. In their work, they analyze the relationship between believability scores for different profiles and their relationship with trust in science. Bruns, Harrington, and Hurcombe (2020) use both quantitative and qualitative methods to understand how such misinformation escalated quickly up to violent attacks. They analyze the Facebook conversations from the start of the first rumor until the arson attempts by defining different phases and providing in-depth analysis for each phase. They discuss how pre-existing conspiracy networks or super-spreaders such as celebrities affected the propagation and the potential pitfalls that resulted in such virality. From a more quantitative perspective, Kauk, Kreysa, and Schweinberger (2021) use epidemiology modeling to build SIR (Susceptible-Infected-Recovered) model to simulate different mitigation strategies such as fact-checking and tweet deletion and evaluate their effectiveness. The authors also point out few shortcomings of the SIR model such as reappearing incidence bursts and extreme peak observed and call for more complex models that account for such behavior.

Given the severity and complexity of the problem at hand, the huge magnitude of the literature on misinformation is reasonable. However, the current attempts to combat misinformation are far from being effective. Since the research on this domain usually focused on one specific dimension of the problem such as propagation, detection, psychological factors, or network properties; the holistic view of the problem is yet to be achieved. In this regard, we argue that developing a formal dynamic simulation model will help to i) identify the causal feedback structure to gain insights into governing dynamics, ii) evaluate the effectiveness of potential structural mitigation strategies, and iii) discuss the similarities and disparities of the general structure for different cases of misinformation. System Dynamics methodology is an ideal fit for such a task as it allows the integration of main dynamic factors and provides a causal interpretation of the emerging dynamics. To narrow it down, we consider 5G-COVID-19 conspiracy theory as the exemplary. We believe that from a methodological perspective the constructed model will serve as a roadmap of potential improvements of SIR models of information diffusion to account for more complex cases, and from the perspective of the problem domain it will contribute a deeper understanding of the problem and provide a testing environment for different policies.

3. MODEL DESCRIPTION

3.1. Causal Structure & Loops

A simplified version of the stock-flow diagram is presented in Figure 1. Fundamentally model is an enriched version of the traditional SIR (Susceptible- Infected- Recovered) model of information diffusion. Misinformation spread is often followed with corrective information either initiated by the informed people or by different authorities such as fact-checking organizations, scientific institutions, experts, etc. Thus, to differentiate between such a difference and reveal the competing dynamics between these groups, 'Infected' stocks are separated as "Believer" and "Informed" stocks. *Believer* stocks represent the people who think that the misinformation is true whereas *Informed* stocks denote the people who are educated enough to know the rumor is false.

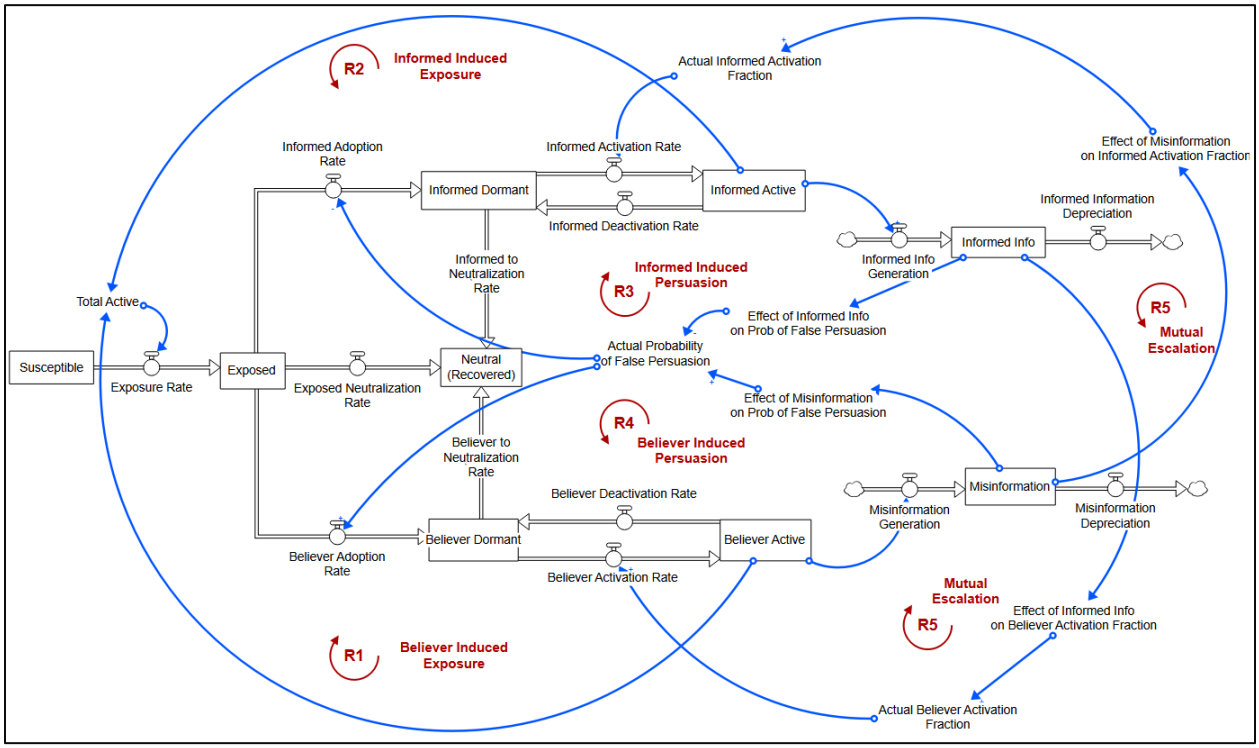


Figure 1: The simplified structure of the model and causal loops

Another distinction is made to differentiate between people who are actively spreading their views (either believer or informed) on the issue or remain silent. Therefore, *Believer Dormant* stock represents the people who believe the false information but remain silent on the issue whereas *Believer Active* is the people who believe and actively contribute to the spread of misinformation. Apart from Believer and Informed stocks, people who are exposed can also remain neutral which is denoted as *Neutral ('Recovered')*. The amount of information generated by Believers and Informed people is represented in *Misinformation* and *Informed Info* stocks respectively.

The causal structure of the proposed model presents five main reinforcing loops:

R1- Believer Induced Exposure & R2 - Informed Induced Exposure: Exposure rate is affected by amount of active people in the population. Thus, an increase in either Informed Active or Believer Active stocks will result in an increased number of exposed people. Eventually, exposed people would proceed in this stock chain and increase the number of Informed & Believer Active, closing the reinforcing loops.

R3- Believer Induced Persuasion & R4 - Informed Induced Persuasion: A constant fraction of *Exposed* becomes *Neutral (Recovered)*. The remaining fraction is split into Believer Dormant with Actual Probability of False Persuasion (p) and Informed Dormant with the complementary probability (1-p). This probability is not constant as it is assumed that the available information would alter such fraction depending on the type of information. Thus, as the amount of *Misinformation* increases, the *Actual Probability of False Persuasion* also increases, resulting in more people adopting the false information. In turn, more believers would produce more misinformation closing the vicious cycle. A symmetric causal loop is present for the informed people as the increment in the *Informed Info* would result in a smaller *Actual Probability of False Persuasion* thus increasing *Informed Adoption Rate*.

R5- Mutual Escalation: A more complex loop emerges from the competing dynamics between the opposing groups. As Misinformation increases, more dormant informed people are inclined to share their opinion followed by an increase in the Informed Info. Similarly, an increase in the Informed Info would result in more people becoming active for Believers. Such behavior is also reported in experiments conducted on digital social networks (Ma and Zhang, 2021).

3.2. Stock-Flow Diagrams & Effect Formulations

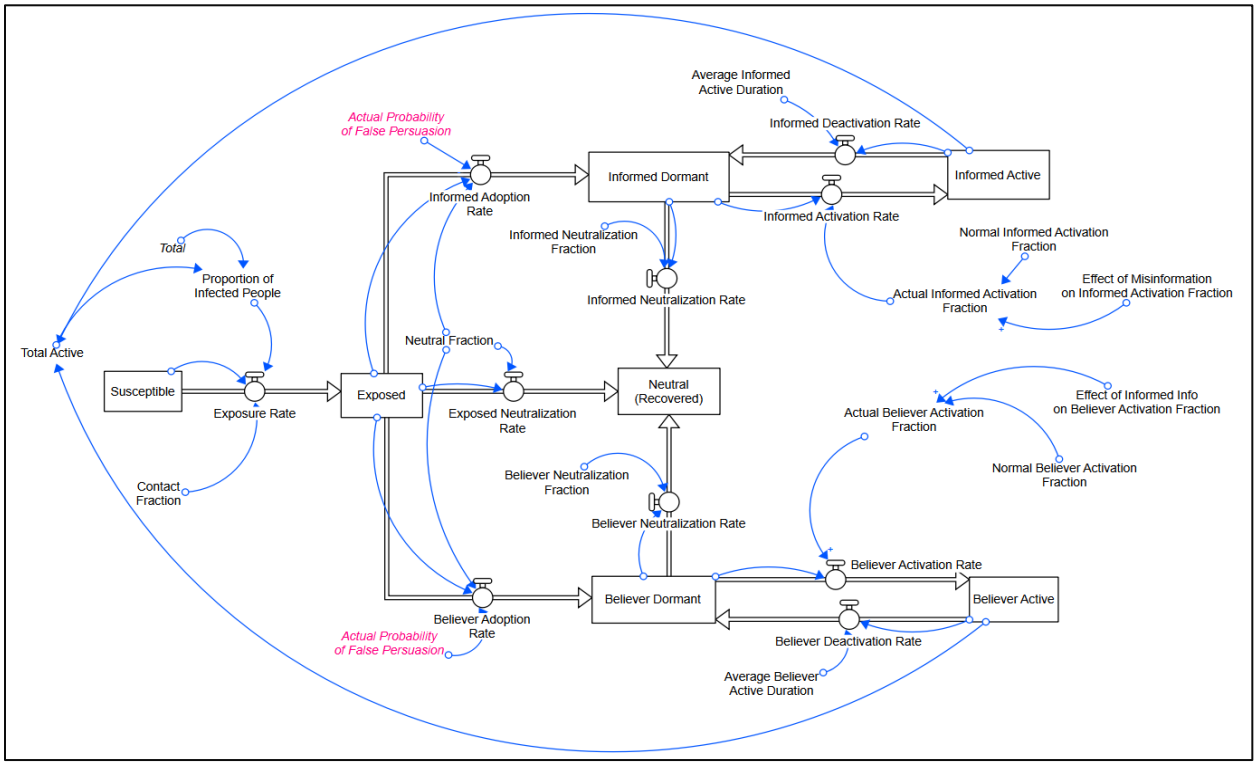


Figure 2: Population stocks

Figure 2 depicts the stock-flow structure for population stocks. *Susceptible* people contact other people with the *Contact Fraction*. The probability of such contact being with an active person is calculated by the ratio of *Total Active* to the *Total* number of people. A constant fraction of *Exposed* remains neutral after the first exposure whereas the remaining is split between the opposing groups. The distribution is calculated using *Actual Probability of False Persuasion*. *Believer Dormant* and *Informed Dormant* stocks flow to the *Neutral (Recovered)* with constant fractions *Believer Neutralization Fraction* and *Informed Neutralization Fraction* respectively. Activation in each group is determined by the actual activation fractions which are multiplications of graphical effect functions with normal values. Deactivation flows are modeled as typical delay formulations using average active durations as delays.

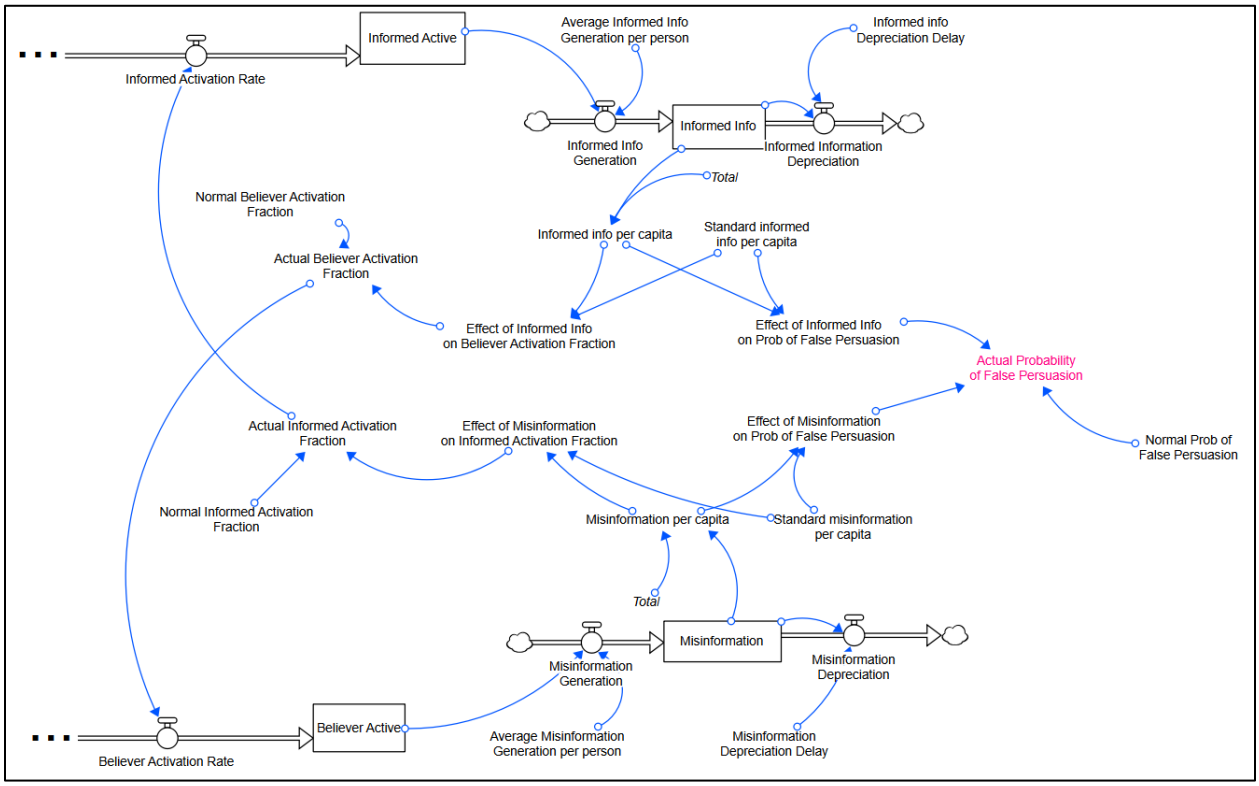


Figure 3: Information stocks and effect functions

Figure 3 represents the information stocks and related effect formulations. *Misinformation Generation* and *Informed Info Generation* are calculated by the multiplication of active stocks with the *Average Information Generation per person*. Generated information depreciates with the depreciation delays. All effect functions are standardized using *Standard Misinformation per capita* & *Standard Informed Info per capita*. Actual values for parameters are derived by multiplying effects with the normal values of the parameters employing standard multiplicative effect formulation (Equation 1).

$$\begin{aligned} &\textit{Effect of Informed Info on Believer Activation Fraction} = f\left(\frac{\textit{Informed Info per capita}}{\textit{Standard Informed Info per capita}}\right) \\ &\textit{Actual Believer Activation Fraction} = \textit{Normal Believer Activation Fraction} * \textit{Effect of Informed Info on Believer Activation Fraction} \end{aligned}$$

Equation 1: Multiplicative effect formulation for the *Effect of Informed Info on Believer Activation Fraction*

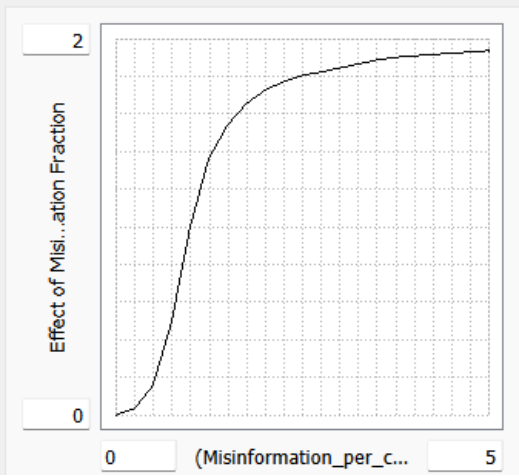


Figure 4: Graphical effect function of *Effect of Misinformation on Informed Activation Fraction*

There are four graphical functions utilized in the model: two effects regarding the *Actual Probability of False Persuasion* and the other two regarding the activation fractions. The *Effect of Misinformation on Informed Activation* is presented in Figure 4. The graphical function takes the ratio of *Misinformation per capita* and *Standardized Misinformation per capita* as input. A recent study reports that the decrease in the perceived peer support increases opinion expression (Ma and Zhang, 2021) thus it is fair to assume that the effect function should be an increasing function of misinformation per capita. Therefore, it is assumed that initially if misinformation is not present informed people should stay in the dormant state whereas as *Misinformation* reaches a theoretical standard point then the *Actual Informed Activation Fraction* assumes its normal value (at point (1,1)). As *Misinformation per capita* passes beyond that standard point; informed people are getting more active as they encounter *Misinformation* more frequently. Such an effect should reach saturation level as the remaining people in the dormant informed stock will be the least motivated ones to speak up. Thus, the increasing function is modeled as having S-shape. *Effect of Informed Information on Believer Activation Fraction* is modeled with the same logic, i.e. the effect is a logistic increasing function of *Informed Information per capita*, only differing in minimum-maximum values as it is assumed that the effect should be less dominant compared to the *Effect of Misinformation on Informed Activation Fraction*.

The effects regarding the information effects on the *Probability of False Persuasion* are formulated using additive graphical effect functions. Thus, the Actual Probability of False Persuasion is calculated as the sum of *Normal Probability of False Persuasion*, *Effect of*

Misinformation on Prob of False Persuasion, and *Effect of Informed Info on Prob of False Persuasion*. Graphical function regarding the misinformation effect on false persuasion is an increasing function as there is more misinformation available we would expect more people to be convinced by the misinformation on average (Figure 5). Similarly if competing *Informed Info* is present, the believability of the false information should decrease, hence corresponding effect is a decreasing function of Informed Info per capita. The initial limits for these graphical functions are taken as 0.1 and -0.1 as a simplification.

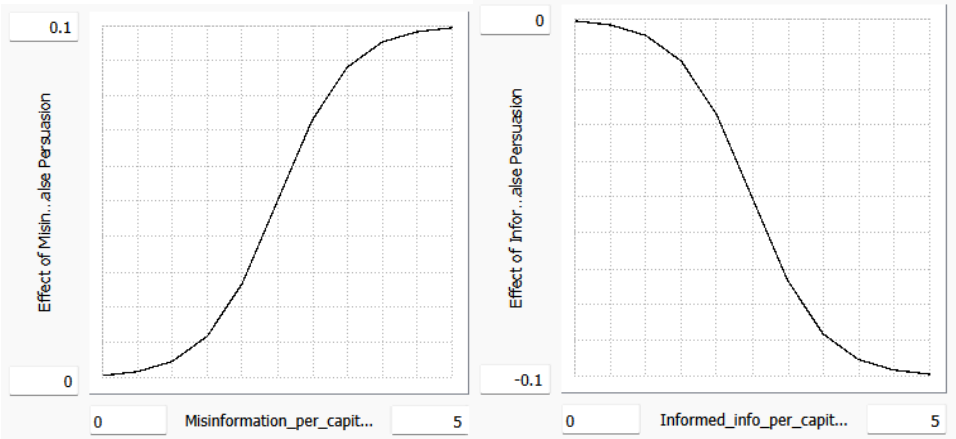


Figure 5: Graphical effect functions for *Effect of Misinformation on Prob of False Persuasion* (left) and *Effect of Informed Info on Prob of False Persuasion* (right)

Some additional clarifications on model assumptions & simplifications:

- Only active stocks (*Believer Active* and *Informed Active*) contribute to the exposure rate. Thus if there are no active people in both groups there wouldn't be any propagation even though there are dormant people.
- False persuasion probability is assumed to be independent of the exposure pathway (whether exposed by an informed or believer). Thus instead of using separate stocks for "Exposed by Believer" and "Exposed by Informed" these two stocks are aggregated in the *Exposed* stock.
- Since the exposure rate formulation is based on the contacts between *Susceptible* and *Active Stocks* (and not affected by *Misinformation* or *Informed Information* directly) in case of no misinformation, there can still be a propagation in this specific social media platform. Therefore the corresponding assumption is that even though we silence all misinformation on a specific platform, there will still be some people that encounter the information from other sources or social media platforms thus still causing a propagation but not contributing to chatter on this platform.
- *Normal Probability of False Persuasion* is assumed to have some constant value. The normal value dynamically changes depending on existing information ecosystems which implicitly assumes that there should be some fraction of the total population who can either believe or disbelieve based on the availability of the competing information. Thus the range

of *Actual Probability of False Persuasion* presents an estimate of the fraction of people that can change their minds based on whether there is competing information or not.

- *Believer Active Duration* and *Believer Activation Fraction* are assumed to be larger than the *Informed Active Duration and Informed Activation Fraction* respectively, as people having a conspiracy mindset have a larger tendency to insist on their viewpoint since they are personally involved in the issue.
- *Misinformation* and *Informed Information* have an artificial unit of "info" instead of tweets or posts. The reason is that since they are used in effect formulations, they should represent the effect of that type of information which should depreciate as time progress. Thus, rather than the physical unit of expression of that specific social media platform, these stocks are defined as soft variables and they correspond to sustained effects of their corresponding information type.
- The possibility of transition from believer to informed (or vice versa) is not allowed, as well as the transition from *Neutral Stock* to any other stocks.
- All other variables except the ones with effect function formulations (*Activation Fractions* and *Probability of False Persuasion*) are assumed to be constant during the simulation horizon in the base model, although the sensitivity results for each one are presented as supplementary material.

3.3. Parameter Selection & Structural Validity

Parameter Name	Unit	Value
Normal Prob of False Persuasion	-	0.22 [1]
Neutral Fract	-	0.1 [3]
Contact Fraction	day $^{-1}$	0.78 [3]
Believer Neutralization Delay	day	9.09 [2]
Informed Neutralization Delay	day	9.09 [2]
Average Believer Active Duration	day	3 [3]
Average Informed Active Duration	day	1 [3]
Normal Believer Activation Fraction	day $^{-1}$	0.68 [3]
Normal Believer Activation Fraction	day $^{-1}$	0.2 [3]
Average Informed Info Generation Per people	information/(day*person)	1 [3]
Average Misinformation Generation per people	information/(day*person)	1 [3]
Informed info Depreciation Delay	day	2 [3]
Misinformation Depreciation Delay	day	2 [3]
Standard informed info per capita	information/person	0.02 [3]
Standard misinformation per capita	information/person	0.02 [3]

Stock Name	Unit	Initial Value
Believer Active	person	5
Believer Dormant	person	0
Exposed	person	0
Informed Active	person	5
Informed Dormant	person	0
Informed Info	information	0
Misinformation	information	0
Neutral (Recovered)	person	0
Susceptible	person	10000

Table 1: Parameter values and initial levels of stocks. [1]: Agley and Xiao, 2021; [2]: Kauk, Kreysa, and Schweinberger, 2021; [3]: Calibrated using data from: Ahmed et al., 2020; Kauk, Kreysa, and Schweinberger, 2021.

Initial parametrization is provided in Table 1. To deduce the model parameters, various research from the literature is utilized. Based on the believability scores obtained in the study of Agley and Xiao (2021), the *Normal Probability of False Persuasion* is kept at around 0.2 during calibration. Initial stock values of *Misinformation* and *Informed Info* are assumed zero as the

beginning of the rumor is assumed at the t=0. Since we are mainly concerned with the dynamics of the system rather than acquiring a perfect fit to data, we select the total number of people in the system as 10000- a close number (approx. 9999) of the total estimated in Kauk, Kreysa, and Schweinberger, (2021)-, assume the initial level of 10 for *Believer Active* to start the propagation, and assume the initial value of 0 for the other stocks.

The analysis conducted by Ahmed and colleagues (2020) revealed that the prevalence of pro-conspiracy (34.8%) and anti-conspiracy (32.2%) tweets about the issue is quite close for the 1 week during the peak time of the debate. Thus, the calibration is made so that the *Believer Active* should be slightly larger than Opposer Active during the peak of the chatter. To calibrate the remaining parameters a novel dataset that is used in a recent study (Kauk, Kreysa, and Schweinberger, 2021) is utilized. In their work, the authors use Twitter hashtag data to approximate the level of Infected people. The hashtag data consist of the number of tweets containing specific hashtags. mostly pro-conspiracy although some hashtags correspond to more broad topics which involve both pro/anti-conspiracy tweets. Due to the volatility of daily hashtag data cumulative hashtag data is used similar to Kauk and colleagues (2021). To obtain a proxy for the cumulative incidence of tweets, it is assumed that an average active believer posts one tweet per week (Average Time to Tweet = 7 days) under these hashtags. Thus the *Cumulative Total Believer Tweets* are defined as a stock with a constant inflow of *Believer Active /Average Time to Tweet*. Although a tweet per day might seem like a small number, since the data at hand is the number of tweets under specific hashtags rather than all tweets about the issue, it is better to assume a smaller number to avoid underestimating the number of believers.

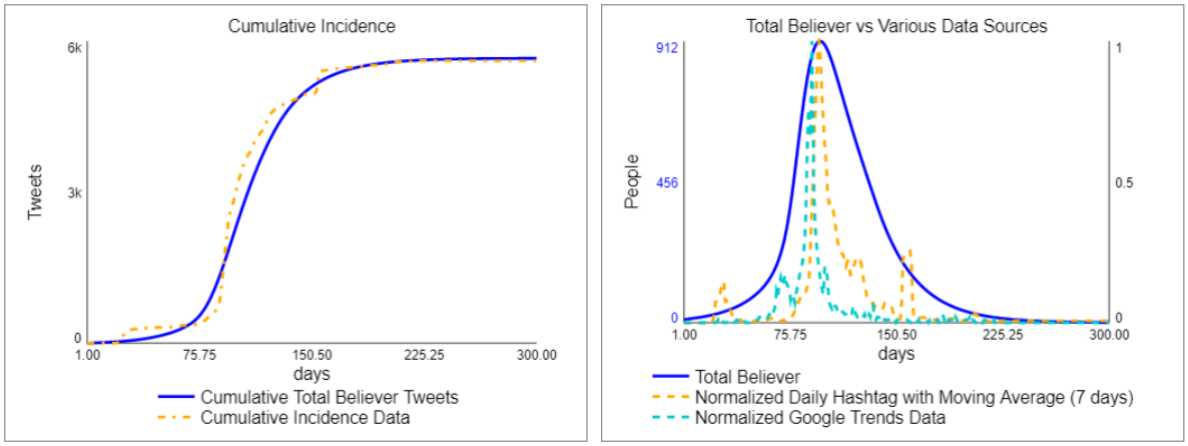


Figure 6: Cumulative Total Believer Tweets simulated (blue) and Cumulative Incidence Data (yellow) on the left; Total Believer (blue), Daily Hashtag Data (cyan), and Daily Hashtag Data with Moving Average (7 days) (yellow) on the right (data from Kauk, Kreysa, Schweinberger, 2021).

As provided in Figure 6, the resulting *Cumulative Total Believer Tweets* provide a good fit with the data and the peak time of Total Believers coincides with the maximum frequency of hashtag data. Moreover, the Believer Active to Opposer Active ratio is close to 1 during the peak which is in line with the findings. Although the current set of parameters does not explain the

reoccurring peaks in hashtag data, we build the base model with this parameter setting and later evaluate whether such differences in the behavior can be obtained by further expanding the dynamic hypothesis by including other causal links.

Regarding the structural validity of the model, we evaluated the model behavior for several extreme conditions such as having no initial active people, very small probability of false persuasion, sudden decrease in the *Misinformation*, or sudden decrease in the *Believer Active*. As for the small probability of false persuasion and no initial active cases, the model shows minuscule and no changes respectively. The results of the other two tests are presented in Figure 7. To test the behavior of the model, external outflows that become active after day 70 are implemented to *Misinformation* and *Believer Active* stocks. Elimination of *Misinformation* after day 70 (Figure 7, top panel) results in no dispute on the subject for this platform whereas the propagation still exists with a lesser impact as the active people continue to spread the misinformation outside this social media platform. Elimination of *Believer Active* on day 70 on the other hand, stops the propagation as observed in the stabilized levels in Susceptibles, thus also ending the dispute in this specific social media platform.

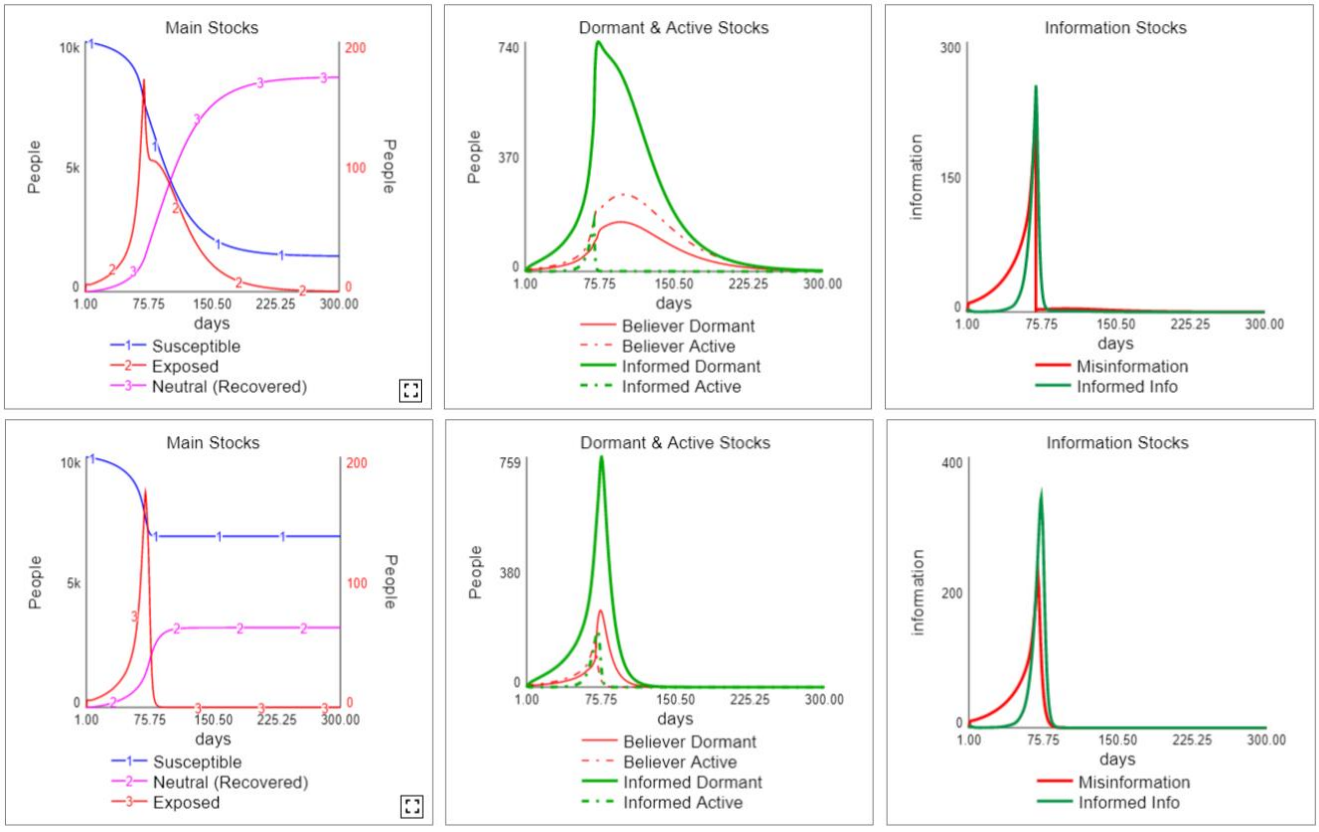


Figure 7: Stock levels for the extreme case of a sudden decrease in Misinformation (top three plots) and the case of a sudden decrease in Believer Active (bottom three plots).

4. SENSITIVITY ANALYSIS

Sensitivity analysis is conducted to analyze the model behavior for changes in model parameters. Full results of the sensitivity analysis are presented as a supplementary material whereas important runs are discussed in this section. To avoid combinatorial complexity, the analysis is conducted one parameter at a time. Overall, the model behavior seems consistent with the model assumptions and real-life implications. To compare the effectiveness of different runs three outcomes of interest are defined as *Total Believer AUC*, *Total Believer Peak*, and *Total Neutralized from Believer*. *Total Believer AUC* represents the area under the *Total Believer* curve during the simulation horizon, *Total Believer Peak* is the maximum level of *Total Believer* and the last one is the cumulative number of people in the *Neutral (Recovered)* stock coming from the *Believer Dormant*.

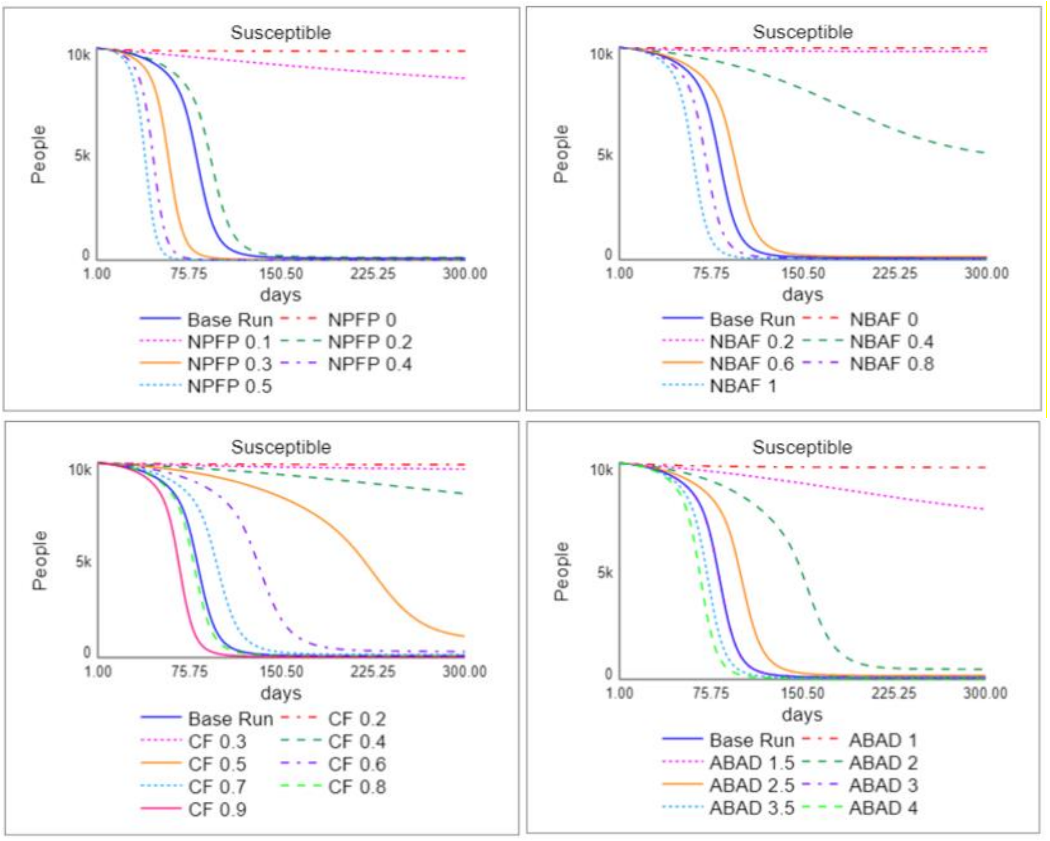


Figure 8: Susceptible stock levels with (a) Changes in Normal Probability of False Persuasion (NPFP), (b) Changes in Normal Believer Activation Fraction (NBAF), (c) Changes in Contact Fraction (CF), (d) Changes in Average Believer Active Duration (ABAD)

For the parameters *Contact Fraction*, *Normal Probability of False Persuasion*, *Average Believer Active Duration*, and *Normal Believer Activation Fraction* there seems to be a tipping point beyond which the epidemic occurs similar to traditional SIR models (Figure 8). Interestingly, none of the changes in informed people parameters can prevent the full spread in the current parameter settings although they are effective in restricting the outcomes of interest which imply

that informed people parameters might be promising leverage points to develop efficient mitigation strategies.

Another observation is that changes in *Believer* parameters typically create a unidirectional change in the outcomes whereas responses to changes in *Informed* parameters often have nonlinear outcomes. For example, increasing *Believer Active Duration* or *Believer Activation Fraction* always results in worse outcomes as observed by increased *Total Believer AUC* and *Total Believer Peak* (see Supp 2.1 and Supp 3.1). However, increasing *Informed Active Duration* or *Informed Activation Fraction* produces changed outcomes based on the trade-offs created by Informed Induced Exposure (R2), Informed Induced Persuasion (R3), and Mutual Escalation (R5) which will be further analyzed in the upcoming section.

5. RESULTS & ANALYSIS

5.1. The Base Run

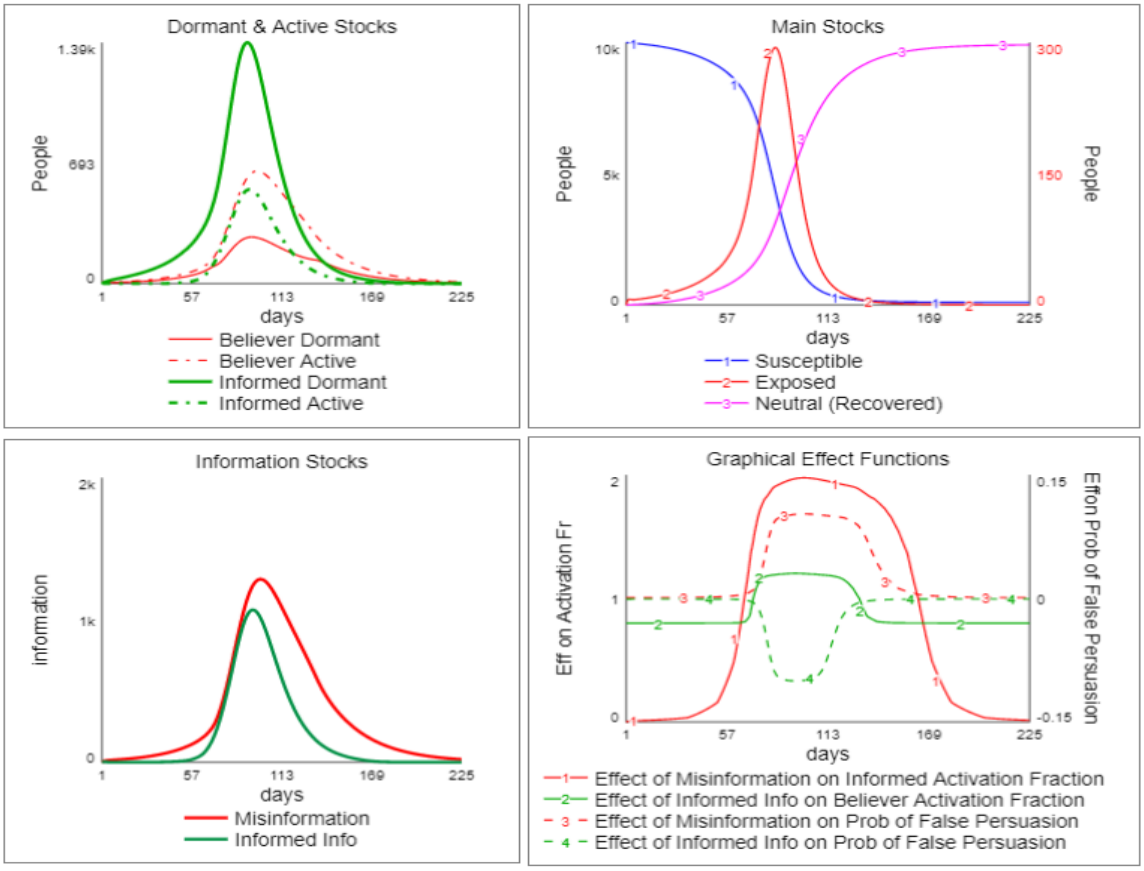


Figure 9: Base run, (a) Dormant & Active Stocks, (b) Main Stocks, (c) Susceptible Exposed and Neutral Stocks, (d) Graphical Effect Functions

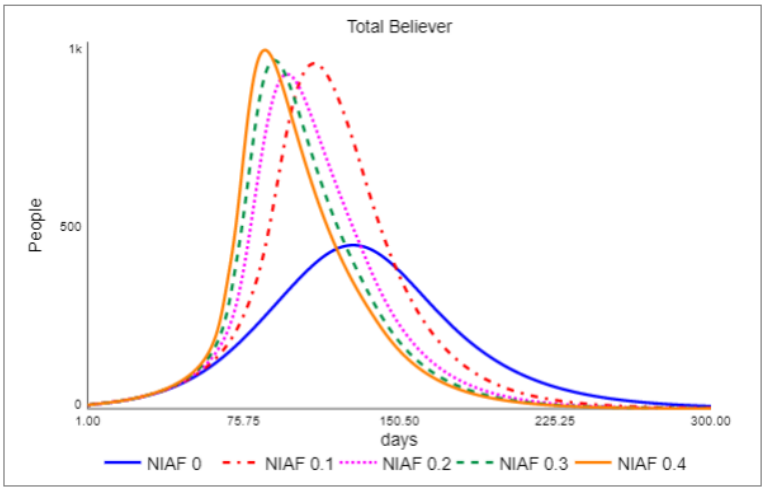
The base run represents 225 days from the 4th of Jan to the 15th of Aug where the debate is prominent on Twitter as hashtags. The dynamics of the main stocks and variables are presented in Figure 6. The behavior of the four people stocks (Fig. 9.a) is similar, as the peak times and the

shapes are nearly the same for all 4 stocks. *Misinformation* seems to exceed the *Informed Info* for the whole period and nearly all of the Susceptible is depleted with the end value of approximately 90 people at the end of the simulation. Considering our assumption that *Susceptible* represents the people on the social media platform that has the potential to participate in the discussion, we can say that the maximum potential is reached for this case. It seems intuitively consistent, as the 5G narrative is one of the most viral instances of misinformation involving distribution channels such as national TV, celebrity super-spreaders, and conspiracy theorists with preexisting social connections (Bruns, Harrington, and Hurcombe, 2020) thus resulting in a wider reach to various audiences.

The initial observation in Figure 9.a is that although the magnitude difference between the *Believer Dormant* and *Informed Dormant* is huge, the *Active Stock* levels are quite close for the two groups. Therefore, one simple insight is even when the pro-conspiracy people in the population are in minority, their presence in the digital sphere (i.e. *Active Stocks*) can dominate the educated people, as believers are more inclined to engage in social media. It should be noted that such an insight is provided by the enriched model whereas the traditional SIR models lack the necessary resolution for such an analysis.

Looking at the graphical effect function values for activation fractions, we see that the *Effect of Misinformation on Informed Activation* is much more effective in comparison to its counterpart effect (Figure 9.d). This should be an intuitive observation as we would expect that informed people will become active and start to speak up only if they are subjected to false claims whereas the motivation of believers would be less sensitive to the existence of an opposition.

5.2. Comparative Runs for Changing Informed Activation Fraction



Total Believer AUC	Final
NIAF 0:	48043
NIAF 0.1:	64720
NIAF 0.2:	58057
NIAF 0.3:	56185
NIAF 0.4:	55106

Total Believer Peak	Final
NIAF 0:	446
NIAF 0.1:	941
NIAF 0.2:	912
NIAF 0.3:	950
NIAF 0.4:	978

Total Neutralized from Believer	Final
NIAF 0:	2007
NIAF 0.1:	2383
NIAF 0.2:	2029
NIAF 0.3:	1918
NIAF 0.4:	1848

Figure 10: Comparative runs for different *Normal Informed Activation Fraction*

A simple analysis can be done for different values for *Normal Informed Activation Fraction*, as in their work, Ahmed and colleagues (2020) suggest the lesser interaction of informed group on the subject would be a better option in terms of isolation of the believer group. The comparative graph of *Total Believer* values and comparative tables of outcomes of interest for

different values of *Normal Informed Activation Fraction* is presented in Figure 10. The minimum peak and AUC are observed when informed people are not involved in the discussion at all which is consistent with the strategy suggested in the literature. However, looking at the changes in *Total Believer AUC*, the relationship is far away from being linear. Starting from zero, increasing *Informed Activation Fraction* worsens the outcome initially, as observed in increased AUC, whereas after 0.1 we observe better outcomes as we continue to increase. Moreover, worst-case outcomes for the *Total Believer Peak* and *Total Believer AUC* happen in different settings.

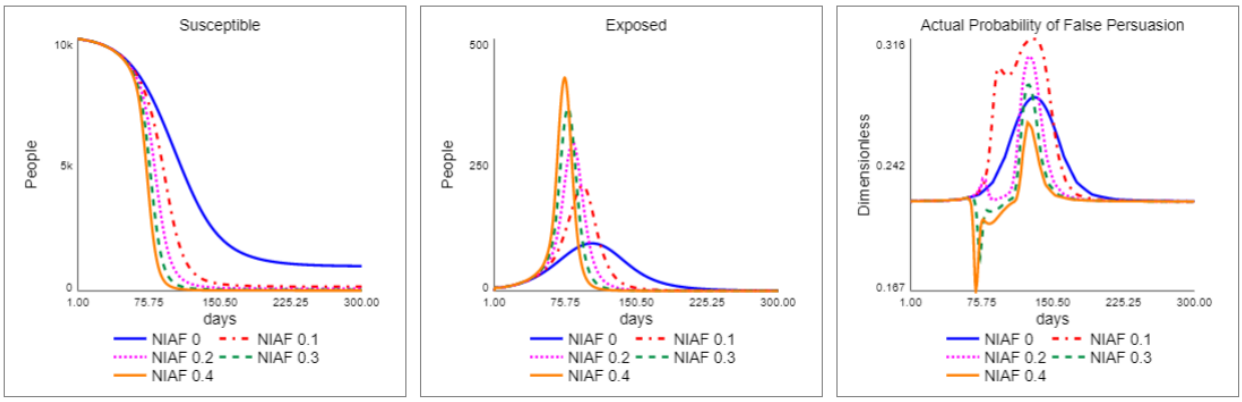


Figure 11: Comparative runs for different *Normal Informed Activation Fraction*

To analyze the tipping points, comparative graphs of *Susceptible*, *Exposed*, and *Actual Probability of False Persuasion* are presented in Figure 11. Initially, when the *Normal Informed Activation Fraction* is low enough, it allows some portion of the population to remain in the *Susceptible*, thus causing less number of *Believers* in turn (Figure 11.a). In other words, if educated people remain silent they might restrict the magnitude of the spread and result in more people who haven't encountered the misinformation at all which also limits the number of believers. Therefore, as informed people become more involved in the discussion, an increased number of *Informed Active* triggers more exposure and increases *Exposed* and *Active Believers* in turn (Figure 11.b). However, a competitive effect is observed through the *Effect of Informed Info Actual Probability of False Persuasion*. By joining the discussion, informed people contribute to convincing evidence against the misinformation. Therefore as informed people become more active, *Informed Info* starts to increase and compete with *Misinformation.* As the *Informed Info* gains dominance over *Misinformation* the probability to be convinced by the false information decreases (Figure 11.c). Decreased probability of false persuasion causes fewer people from the *Exposed* to proceed to the *Believer Dormant*. In sum, increased involvement of the informed people has two conflicting impacts on the eventual number of believers: It causes more people to be exposed to the information but simultaneously decreases the probability of false persuasion. These two competing impacts offer different trade-offs for different outcomes of interest:

For the *Total Believer AUC* and *Total Neutralized from Believer*; we observe that if the *Susceptible* stock is close to zero at the end of the simulation, i.e. if the misinformation becomes viral and reaches its full potential spread, it is better to speak up rather than remaining silent as the silence will result in more people who haven't decided to become believers. On the other hand, if

the misinformation does not have the potential to spread to the whole population, speaking up will result in a worse outcome which is triggered by the exposure.

For the *Total Believer Peak*; the increased activation of informed people will inevitably result in higher and sharper peaks for the *Exposed* stock. However, unlike the observation made for the other outcome of interests, even after the point where the *Susceptible* eventually zeroed out, it is not always better to speak up. As seen in Figure 8, increasing the *Normal Informed Activation Fraction* beyond 0.3 worsens the *Total Believer Peak* even though it causes better outcomes for *Total Believer AUC* and *Total Neutralized from Believer*. Thus it is evident that the increased peak impact in the exposed stock eventually shadows the decrease caused by the smaller probability of false persuasion.

5.3. Comparative Runs for Information Campaign

One of the main criticisms of the management of the 5G-COVID 19 conspiracy spread was the lack of explanations and denouncements by the authority figures (Ahmed et al., 2020). Using this idea, an "Information Campaign" on the social media platform is tested as a mitigation strategy. The information campaign is implemented as an inflow to the *Informed Information* starting at *Campaign Start* for the duration of *Campaign Duration*. The amount of inflow is denoted as *Campaign Intensity* and assumed to be constant during the campaign. The analysis is conducted for different values of *Campaign Start* and *Campaign Duration*, whereas *Campaign Intensity* is kept constant (equal to 300) for the analysis to simplify the comparisons.

Neutralized From Believer		Duration				
		10	20	30	40	50
Start (date)	40	2008	1960	1817	1631	1611
	50	1985	1813	1661	1644	1635
	60	1875	1721	1705	1695	1687
	70	1843	1826	1816	1809	1799
	80	1995	1986	1980	1971	1966
	90	2020	2013	2005	2000	1998
	100	2022	2014	2009	2007	2006
	110	2019	2013	2011	2010	2010
	120	2023	2020	2019	2019	2019

Neutralized From Believer		Duration				
		10	20	30	40	50
Start (date)	40	2008	1960	1817	1631	1611
	50	1985	1813	1661	1644	1635
	60	1875	1721	1705	1695	1687
	70	1843	1826	1816	1809	1799
	80	1995	1986	1980	1971	1966
	90	2020	2013	2005	2000	1998
	100	2022	2014	2009	2007	2006
	110	2019	2013	2011	2010	2010
	120	2023	2020	2019	2019	2019

Table 2: Comparative tables of *Total Neuralized from Believer* for different values of *Campaign Start* and *Campaign Duration*. Color scale corresponds to higher and lower values for the red and white respectively. The table on the left is colored by column, whereas the table on the right is colored by row.

The complete analysis of all three of the outcomes of interest for different scenarios is presented in Appendix A. Since the insights derived are mostly consistent for all outcomes of interest, only results of *Neutralized From Believer* are discussed in this section. The table below (Table 2) shows the final *Neutralized From Believer* for different values of *Campaign Start* and *Campaign Duration*. The color scale is applied based on columns for the table on the left and based on rows on the right where red indicates the highest value in a row/column and white represents the lowest.

Given a fixed start time, increasing the duration of the campaign seems to be effective for all starting dates which presents a consistent scenario with intuitive thinking. However, although the change of direction remains the same, the observed improvement is very little for the interventions after 70, given that the *Neutralize From Believer* count was 2029 for the base case scenario without the information campaign.

Regarding the start time of the campaign, unidirectional thinking would suggest intervening as early as possible would produce better outcomes. However, results indicate that early interventions might be ineffective if the duration of the intervention is not lasting. For example, if the duration of the information campaign is determined to be 10, starting it on day 40 instead of day 70 would result in 165 more believers (around 10% increase).

6. CONCLUSIONS & FUTURE RESEARCH

We have reviewed the literature on misinformation spread on social media specifically for the 5G-COVID-19 narrative and built a system dynamics model for the problem. The model is constructed using both quantitative and qualitative literature and validation with respect to real data and extreme condition tests are presented. Using the proposed interventions from the literature, analysis is conducted for two different cases, namely decreasing *Informed Activation Fraction* and intervention of corrective information campaign.

The results from the scenario analysis indicate that the nonlinear relationships in the system result in several counterintuitive outcomes. Firstly, decreasing the *Informed Activation* which is commonly cited as a prevention method in the literature does not follow a unidirectional pattern in terms of effectiveness. Our analysis indicates that limiting the involvement of the informed people in the online discussion would produce better outcomes only if it results in more people who have not encountered the information at all as compared to the no intervention case. In other words, if the virality of the misinformation is high enough to spread to all of the potential users, i.e. *Susceptible*, even after the intervention, then the resulting outcome will be worse than no intervention case. Thus a practical insight that can be derived from such an analysis is that such an intervention should not be planned to be implemented "as much as possible". On the contrary, it should have a specific target below which the policy-makers are sure that, or at least somewhat confident that, the tipping point will be passed and the outcome will be better if the target is reached.

Another result is that the outcome of interests does not always follow the same direction of change for the mitigation scenarios. As an example, regarding the *Total Number of Neutralized from Believers*, increasing *Informed Activation Fraction* unidirectionally produces better outcomes (Figure 10). On the other hand, the same policy would produce worse outcomes in terms of *Total Active Peak* as evident in the increased level of peak values in Figure 10. Such a difference is important as one outcome might be more important for specific types of misinformation. For instance, in 5G-COVID 19 case, looking at the reported violence in the comments on social media platforms (Bruns, Harrington, and Hurcombe, 2020) one can focus on minimizing the maximum

number of believers to avoid violent protests but, for a case about information that can change long-term behaviors such as public health measures, decision-makers may prioritize minimizing the total number of believers during the spread.

Results from the scenario analysis of the information campaign indicate that such an intervention should be planned carefully to maximize the improvements in the outcomes. If the duration of the campaign is limited then it should be timed carefully, while if such a limitation is not present the impact will be maximum when the intervention starts as early as possible and is sustained until the end of the spread.

To sum, the initial analysis of the model represents several trade-offs that can result in unintended consequences for the proposed mitigation strategies in the literature. The next step for this research is to test the model assumptions, outputs, and robustness of the insights further by making use of richer cross-sectional and dynamic data. Moreover, other possible mitigation strategies can be incorporated into the model to assess the effectiveness for different scenarios. Another agenda is to expand the model by including more user profiles such as "like-seekers" instead of just two opposing sides as there are different motivations for other groups to engage. Different susceptibility of these groups may also be incorporated as such a classification and differentiation is presented in the current literature (Agley and Xiao, 2021). Finally, a discussion on similarities and disparities between our specific case and other types of misinformation spread can be useful to infer the harmony of interventions for different cases of false information spread.

7. REFERENCES

Agley, J., & Xiao, Y. (2021). Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*, *21*(1), 1–12. https://doi.org/10.1186/s12889-020-10103-x

Ahmed, W., Vidal-Alaball, J., Downing, J., & Seguí, F. L. (2020). COVID-19 and the 5G conspiracy theory: Social network analysis of twitter data. Journal of Medical Internet Research, 22(5), 1–9. https://doi.org/10.2196/19458

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. 31(2), 211–236.

Ammara, U., Bukhari, H., & Qadir, J. (2020). Analyzing Misinformation Through The Lens of Systems Thinking. Tto, 55–63.

BBC. (2020, April 4). *Mast fire probe amid 5G coronavirus claims*. BBC News. Retrieved March 17, 2022, from https://www.bbc.com/news/uk-england-52164358#

Brewis, H. (2020, April 14). *Nightingale Hospital Phone Mast attacked as 5G conspiracy theory rages*. London Evening Standard | Evening Standard. Retrieved March 17, 2022, from https://www.standard.co.uk/news/uk/nhs-nightingale-phone-mast-arson-attack-5g-conspiracy-a4414351.html

Bridgman, A., Merkley, E., Zhilin, O., Loewen, P. J., Owen, T., & Ruths, D. (2021). Infodemic Pathways: Evaluating the Role That Traditional and Social Media Play in Cross-National

Information Transfer. *Frontiers in Political Science*, *3*(March). https://doi.org/10.3389/fpos.2021.648646

Bruns, A., Harrington, S., & Hurcombe, E. (2020). 'Corona? 5G? or both?': the dynamics of COVID-19/5G conspiracy theories on Facebook. *Media International Australia*, *177*(1), 12–29. https://doi.org/10.1177/1329878X20946113

Caled, D., & Silva, M. J. (2021). Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. In *Journal of Computational Social Science* (Issue 0123456789). Springer Singapore. https://doi.org/10.1007/s42001-021-00118-8

Chan, M. pui S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Kauk, J., Kreysa, H., & Schweinberger, S. R. (2021). Understanding and countering the spread of conspiracy theories in social networks: Evidence from epidemiological models of Twitter data. *PLOS ONE*, *16*(8), e0256179. https://doi.org/10.1371/journal.pone.0256179

Kumar, S., & Shah, N. (2018). False Information onWeb and Social Media: A Survey. ArXiv, April.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee Howell et al. Digital wildfires in a hyperconnected world. WEFReport, 3:15–94, 2013.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest, Supplement*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lotito, Q. F., Zanella, D., & Casari, P. (2021). Realistic aspects of simulation models for fake news epidemics over social networks. *Future Internet*, *13*(3). https://doi.org/10.3390/fi13030076

Ma, S., & Zhang, H. (2021). Opinion Expression Dynamics in Social Media Chat Groups: An Integrated Quasi-Experimental and Agent-Based Model Approach. *Complexity*, *2021*. https://doi.org/10.1155/2021/2304754

Moon, Angela. "Two-Thirds of American Adults Get News from Social Media: Survey." Reuters, Thomson Reuters, 8 Sept. 2017, www.reuters.com/article/us-usa-internet-socialmedia/two-thirds-of-american-adults-get-news-from- social-media-survey-idUSKCN1BJ2A8.

O. Varol et al., in Proceedings of the 11th AAAI Conference on Web and Social Media (Association for the Advancement of Artificial Intelligence, Montreal, 2017), pp. 280–289.

Sterman, J.D. Business Dynamics: Systems Thinking and Modeling in a Complex World. McGraw-Hill, Boston, 2000.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. ACM SIGKDD Explorations Newsletter, 21(2), 80–90. https://www.public.asu.edu/~huanliu/papers/Misinformation_LiangWu2019.pdf

Vicario, M. Del, Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences of the United States of America, 113(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., Li, D., Wu, J., & Havlin, S. (2020). Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, *9*(1). https://doi.org/10.1140/epjds/s13688-020-00224-z

Zhou, X., & Zafarani, R. (2018). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. https://doi.org/10.1145/3395046

Appendix A. Complete Tables for Scenario Analysis of Information Campaign

AUC		Duration				
		10	20	30	40	50
Start (date)	40	57742	56743	52887	47387	46866
	50	57365	52586	48066	47659	47463
	60	54117	49549	49152	48963	48827
	70	52684	52265	52080	51969	51843
	80	57160	57000	56919	56848	57127
	90	57899	57822	57757	58049	58624
	100	57972	57907	58201	58779	59321
	110	57932	58228	58807	59350	59785
	120	58401	58988	59537	59976	60313

AUC		Duration				
		10	20	30	40	50
Start (date)	40	57742	56743	52887	47387	46866
	50	57365	52586	48066	47659	47463
	60	54117	49549	49152	48963	48827
	70	52684	52265	52080	51969	51843
	80	57160	57000	56919	56848	57127
	90	57899	57822	57757	58049	58624
	100	57972	57907	58201	58779	59321
	110	57932	58228	58807	59350	59785
	120	58401	58988	59537	59976	60313

Total Believer		Duration				
		10	20	30	40	50
Start (date)	40	908	889	805	682	671
	50	907	802	701	692	687
	60	834	731	722	718	718
	70	792	783	779	779	779
	80	891	887	887	887	887
	90	909	909	909	909	909
	100	912	912	912	912	912
	110	912	912	912	912	912
	120	912	912	912	912	912

Total Believer		Duration				
		10	20	30	40	50
Start (date)	40	908	889	805	682	671
	50	907	802	701	692	687
	60	834	731	722	718	718
	70	792	783	779	779	779
	80	891	887	887	887	887
	90	909	909	909	909	909
	100	912	912	912	912	912
	110	912	912	912	912	912
	120	912	912	912	912	912

Neutralized From Believer		Duration				
		10	20	30	40	50
Start (date)	40	2008	1960	1817	1631	1611
	50	1985	1813	1661	1644	1635
	60	1875	1721	1705	1695	1687
	70	1843	1826	1816	1809	1799
	80	1995	1986	1980	1971	1966
	90	2020	2013	2005	2000	1998
	100	2022	2014	2009	2007	2006
	110	2019	2013	2011	2010	2010
	120	2023	2020	2019	2019	2019

Neutralized From Believer		Duration				
		10	20	30	40	50
Start (date)	40	2008	1960	1817	1631	1611
	50	1985	1813	1661	1644	1635
	60	1875	1721	1705	1695	1687
	70	1843	1826	1816	1809	1799
	80	1995	1986	1980	1971	1966
	90	2020	2013	2005	2000	1998
	100	2022	2014	2009	2007	2006
	110	2019	2013	2011	2010	2010
	120	2023	2020	2019	2019	2019