Testing a Decentralized Fact-checking Approach with Spatial Games of Fake News

Matthew I Jones, Scott D Pauls, Feng Fu

November 21, 2022

Abstract

Online social media is plagued by fake news, but centralized fact-checking by governments is fraught with complications about free speech and censorship and content publishers like Facebook and Twitter are unmotivated to clamp down on the misinformation that is driving traffic and therefore advertiser funding. Recent studies have proposed an online crowdsourced fact-checking approach as one possible intervention method to curb the spread of misinformation online. However, it remains unclear under what conditions such efforts are successful in containing misinformation. To study this issue, we develop a mathematical model in which fact-checking is done by specified individuals who reward those around them that share real news and punish those spreading false information. By simulating our model on synthetic square lattices and small-world networks, we show that the presence of social network structure enables fake news spreaders to self-organize into echo chambers, thereby boosting the efficacy of fake news and more than doubling its resistance to fact-checking efforts. Additionally, we use a Twitter network dataset and show that fake news can be contained with relatively few fact-checkers if they are chosen deliberately. Finally, by studying our model analytically, we determine the conditions under which real or fake news is favored in the limit of weak selection. Our work has practical implications for developing mitigation strategies to control the spread of misinformation that interferes with the political discourse.

Significance Statement

Modern social media has been inundated by false and misleading headlines and articles. We advance the study of such fake news by developing a game-theoretic model of the spread of fake news in a social network. We are able to test the effectiveness of peer fact-checkers and we estimate that the structure of a social network can increase misinformation's resistance to fact-checking in a population by a factor of two to three, but by carefully choosing which individuals

are equipped to be fact-checkers can significantly boost fact-checking efforts in certain networks.

1 Background

By creating an environment that encourages reposting inflammatory headlines instead of developing nuanced understandings of complex topics, social media platforms seem designed to spread fake news [1, 2, 3]. Online misinformation has negatively impacted every aspect of society, including terrorist attacks [4, 5], the COVID-19 pandemic [6, 7, 8], and of course, politics and elections [9, 10, 11].

Unfortunately, there are no easy solutions to the problem of fake news. Regulation of false content by government authorities opens a "Pandora's Box" of concerns around censorship and free speech rights, with many experts rightly pointing out that such efforts could impact everyone's online experiences, including those who do not currently create or consume fake news [12, 13]. On the other hand, because almost all news media is advertiser-driven, social media sites are incentivized to boost false information to increase engagement from consumers [14]. Because of these obstacles, we will focus on fact-checking at the level of the individual consumer. One recent study suggested crowdsourced fact-checking as a possible intervention to reduce misinformation [15], which avoids the twisted motivations of government agencies and social media companies. This work is designed to test the feasibility of such an approach.

Recent research shows that "inoculation" with exposure to a weakened version of misleading arguments (similar to vaccination ideas) is effective at reducing susceptibility to misleading persuasion and thus confers psychological resistance to fake news [16, 17, 18, 19]. By training some subset of the population to identify and respond to fake news, we can create a decentralized fact-checking system where innoculated individuals will be positioned to apply pressure to their social neighbors that share fake news while also supporting their real news sharing neighbors. Inspired by "zealot models" from the field of opinion dynamics [20], we assume that these fact-checkers will never change belief because of the behavior of those around them, having been successfully inoculated against fake news,

The study of misinformation has grown rapidly in an effort to limit its impact and has shed some light on how these stories reach such a wide audience despite containing blatant falsehoods. In [2], Shin, Jian, Driscoll, and Bar traced the lifecycle of 17 popular political rumors that circulated on Twitter over 13 months during the 2012 U.S. presidential election; they found that misinformation tends to come back multiple times after the initial publication, while facts do not. Using massive Twitter datasets, it was recently reported that the spread of true and false news follow distinctive patterns: falsehood diffused significantly faster, deeper, and more broadly than the truth in all categories of information [3]. These studies and others suggest that fake news has some innate advantage over real news when shared on online platforms.

Making matters worse, social influence, following, and unfollowing can create

polarized and segregated structure in online social networks like Twitter [21]. These *echo chambers* create conditions for confirmation bias and selection bias and thus can facilitate the spread of misinformation [22]. In [23], Evans and Fu investigated opinion formation on dynamic social networks and, using the voting records of the United States House of Representatives, presented and validated the conditions for the emergence of partisan echo chambers [24, 25, 26]. Integrating publicly available Twitter data with an agent-based model of opinion formation driven by socio-cognitive biases, Ref. [21] recently found that open-mindedness of individuals is a key determinant of forming echo chambers under dueling campaign influence.

Whether these echo chambers form because individuals copy those around them or tend to make connections with like-minded individuals, it is clear that something about the structure of a social media network (e.g. friend/follower networks) is allowing misinformation to fester. To attempt to quantify the effect network structure has on the proliferation of fake news, we develop a mathematical model of fake news sharing and test it on a variety of social networks. There is an established tradition of using spatial game theory to study problems of coordination and collective action. The structure of a network has been found to reinforce good behavior [27, 28], and the evolution of the system can exhibit interesting spatial phenomenon that is not present in the well-mixed case [29]. We use a similar strategy to study the spread of fake news through a social network.

We also draw inspiration from the field of opinion dynamics on networks. It has been a major research concern to effectively understand circumstances that will lead to consensus of opinion in a population and others that will lead to divergence of opinion and a weakening of information transfer [30, 31, 32, 33, 34, 20, 35]. In our work, there are two competing choices of opinion: real news and fake news. A successful outcome for our model is population-wide consensus on real news, but we frequently see a middle ground where isolated communities form echo chambers and continue to hold minority beliefs.

Here we study spatial games of fake news by modeling distributed fact-checking efforts which will discourage sharing false information while rewarding the spread of reliable information. Predetermined fact-checkers will be placed into the population to represent fake news inoculation efforts. Our agent-based model, where individuals can share real or fake news depending on the behavior and success of their neighbors, is studied with simulations as well as a rigorous mathematical analysis. We find that the presence of subtle network structures resembling echo chambers impede crowdsourced fact-checking, thereby requiring a much higher threshold of fact-checkers across the population to contain online misinformation.

2 Methods & Model

This paper centers on a model of fake news that is inspired by the virtual interactions that occur all the time on online social media sites. To begin, an

individual makes a post containing a news story (either true or false) for all of her friends or followers to see. Those who believe the story is true can react positively to the post by liking or sharing, while those who disagree may simply ignore the post or even attempt to debunk a false story by pointing out flaws or sharing a link to a fact-checking website, inflicting a social penalty for sharing fake news.

We also want to take into account that these interactions happen on social networks with limited connectivity, not in an open space where everyone knows everyone and sees everything. Consider a network where vertices represent individuals that can exhibit three kinds of behavior: sharing real news (labelled A), sharing fake news (B), and fact-checking (C). Individuals will receive a payoff depending on their behavior and the behavior of their neighbors. These payoffs are encoded in the payoff matrix like the one below.

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & -4 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \tag{1}$$

To read a payoff matrix, look at the row corresponding to an individual's strategy and the column corresponding to her neighbor's strategy. For example, when an A player interacts with a C player, the A player gets a payoff of 1 and the C player gets a payoff of 0.

In our model, real and fake news sharers both get positive payoffs when interacting with like-minded individuals, which represents the social capital gained by likes, shares, retweets, comments, etc. Because fake news tends to generate more interest than real news on social media [3], our model gives a higher payoff to individuals sharing fake news. To contain the spread of fake news, this natural advantage given to B players will have to be counterbalanced by a penalty inflicted when meeting C players, who behave like real news sharing A players but are also willing to publicly call out fake news when they see it on social media.

Having been successfully immunized against fake news, our fact-checkers will never change strategy, playing C during every time step. Therefore, the proportion of fact-checkers p_C is prescribed and static, representing the level of inoculation effort. The payoff to fact-checkers is irrelevant as the fact-checker population is unchanging, so for simplicity we set it to zero. A selection strength parameter controls how much impact an individual's payoff has on her reproductive success in the update step. The payoffs and selection strength can take arbitrary numerical values, but for the rest of this paper, unless otherwise noted, we will use a selection strength of $\beta = 0.5$ and the payoff matrix for this symmetric, two-player game will be the matrix in Equation (1).

Over time, if individuals see that only certain types of stories receiving positive feedback, they may be convinced of the accuracy of those (potentially false) narratives [36] and begin sharing those same stories themselves. We will use a death-birth process for the evolutionary strategy update [27] to capture this

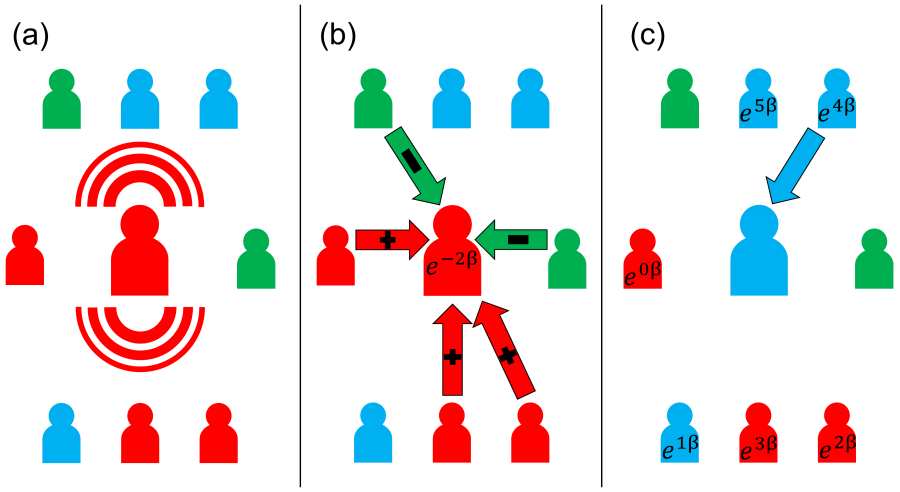

Figure 1: Model schematic. We model information sharing and fact-checking through the lens of spatial games. First, individuals share news that is either true (blue) or false (red), shown in (a). In (b), we see a focal individual receiving positive or negative feedback from her neighbors depending on their relative beliefs. The presence of crowdsourced fact-checkers (green) can reduce the fitness of fake news sharers substantially. Finally, in (c), individuals update their strategy by copying a neighbor proportional to fitness. However, fact-checkers do not update strategy and are never chosen to be replicated.

social imitation phenomenon. After computing the expected payoff π_i for every individual i, a focal individual imitates the strategy of one of her neighbors, chosen with probability proportional to their fitnesses $f_i = \exp(\beta\pi_i)$. Thus, individuals with high payoff are likely to be selected, but even individuals with a low payoff due to repeated fact-checks or social isolation could be chosen to reproduce occasionally.

In our investigation, we use two flavors of this update rule: synchronous and asynchronous. In the synchronous update, used in our simulations, every individual simultaneously updates their strategy, while in the asynchronous update, which lends itself to easier mathematical analysis, a single individual is chosen uniformly at random to update. These two update rules will lead to very similar outcomes, and the minor differences between them are manifested only in edge cases that occur rarely for reasonable fact-checker densities. Keeping this in mind, we will treat them as the qualitatively same process operating on different time scales.

The basic outline of our model is shown in Figure 1. First, individuals play the fake news game with neighbors by broadcasting a real or fake post. The expected payoff from these games is then converted into a fitness. Figure 1c demonstrates the asynchronous update, where only a single focal individual

updates strategy by considering the fitness of all neighbors. In the synchronous update, all individuals would select a neighbor simultaneously.

Our study of the spread of fake news focuses on three types of networks with different properties: a 30×30 square lattice [29], Watts-Strogatz small-world networks [37] (also with $N = 900$), and a portion of the Twitter follower network [38] ($N = 404719$). Our small-world networks are calibrated to have the desired high clustering coefficients and short path lengths using the following parameters: base degree 8 and rewiring probability 0.03, giving us approximately 200 shortcuts. The Twitter network is interesting for its size but also its natural clustering and the gatekeeping individuals that control the flow of information through the network. Although edges in the network were originally directed, we symmetrized the network before using it to match the bidirectional flow of information in our model.

To initialize our simulations, we assign some fraction p_C of the individuals as fact-checkers, and the rest we set to be A or B players with probability $\frac{1}{2}$. After initializing the system, we allow it to evolve using one of the update processes described above until all possible players are sharing the same type of news or a predetermined number of time steps is reached. At the end of the simulation, the type of news with more sharers is said to be dominant, and if there are no individuals sharing one type of news, we say that that strategy has gone extinct and the other strategy has fixated.

3 Results

We used both computer simulations and analytic techniques to study this spatial game of fake news. Using simulations, we demonstrate that the spontaneous formation of echo chambers can be driven by local variation in fact-checker density, and that these echo chambers are extremely resistant to invasion. We also test the hypothesis that network structure seems to protect fake news from fact-checking efforts, and we examine the viability of targeted inoculation efforts, when fact-checkers are carefully selected to maximize fact-checking impact with minimal resources. Finally, we use analytic techniques to determine which collections of payoff values favor real news, favor fake news, or favor neither.

3.1 Echo Chamber Formation

When there are very few fact-checkers, the natural advantage that fake news sharers have allows them to drive the real news sharing strategy to extinction. Similarly, when there is a sufficient fact-checker presence, the risk of punishment for spreading misinformation is too great and the entire population eventually converges to sharing real news. However, there is a wide range of fact-checker densities where neither strategy is driven to extinction quickly and instead we see the spontaneous formation of echo chambers in our simulations. These echo chambers are not deliberately seeded, instead emerging on their own from the

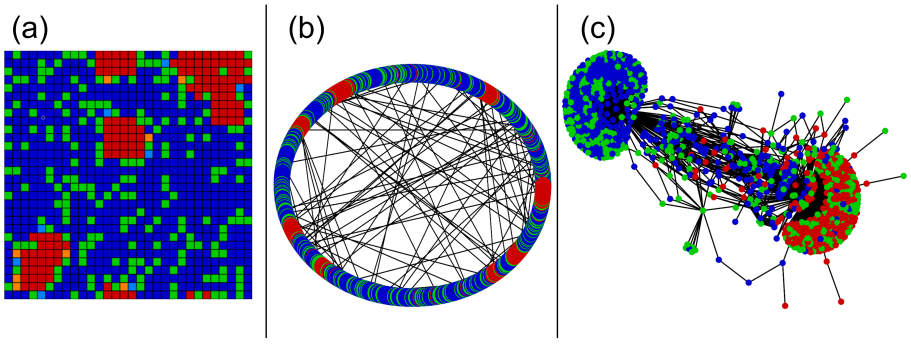

Figure 2: Echo chambers of fake news spreaders in a majority real news-spreading population that are isolated from the rest of the population. In (a), the lightly-colored individuals are those that have changed strategy recently. The network in (b) is a Watts-Strogatz small world network, and (c) is a small breadth-first subgraph of the Twitter network of approximately 1,000 vertices, but the simulation was run on the entire $\approx 400,000$ vertex network (see Methods & Model).

noisy initial state where real and fake news sharers each make up approximately 50% of the population.

We define echo chambers by their longevity, as either real or fake news goes extinct unless the minority strategy manages to form small, highly interconnected communities that are secure from invasion by the majority strategy. For additional discussion about the longevity of these pseudo-steady states, see the Supplementary Information. Figure 2 shows examples of these echo chambers on the three different network topologies we studied.

Once they form, these echo chambers are incredibly resistant to invasion, resulting in a *pseudo-steady state* that cannot last forever, but will take an extremely long time to break down. After forming in relatively few (¡100) time steps, these echo chambers remained largely unchanged for over one million time steps in our longest simulations. There will be small variation in the pseudo-steady state when specific individuals change behavior, but as a whole the echo chamber remains unchanged. Observe in Figure 2a that the only individuals changing strategy are on the borders of the echo chambers in the system. It is very unlikely that a small perturbation on the border will result in any change to the interior of the echo chamber. Individuals on the periphery of the echo chamber are exposed to both real and fake news and may change strategy occasionally, but those in the interior are surrounded by like-minded individuals and have high fitness, which allows them to reinforce minority behavior by the more exposed peripheral individuals.

Our comprehensive simulations confirm that the formation of echo cham-

bers occurs across a wide range of payoff values and selection strengths. Local variation in fact-checker density means in some areas there are no fact-checkers (leaving room for a fake news echo chamber) and in others they make a fact-checking wall which becomes more and more difficult for fake news sharers to penetrate as selection strength grows.

3.2 Critical Fact-checker Density

These echo chambers can made up of fake news sharers, as in Figure 2, or real news sharers if fact-checker density is low enough. The critical fact-checker density is the tipping point at which real news sharers are more likely to be in the majority than in the minority. Figure 3 shows how the probability that real news becomes the dominant strategy changes as the fraction of fact-checkers increases. It is clear that the critical fact-checker density varies depending on the topology of the social network: $p_c \approx 0.235$ on the square lattice, $p_C \approx 0.2$ for small-worlds, and $p_C \approx 0.275$ for the Twitter network.

We can compare these results to the simple case of an infinite, well-mixed population evolving according to replicator dynamics [39]. After initializing with some fraction p_C of fact-checkers and the rest of the population evenly divided between real and fake news (so $p_A = p_B = \frac{1-p_C}{2}$), we consider the relative payoffs of A and B players when choosing a random opponent under the payoff matrix (1).

The expected payoff for an A player at $t = 0$ is

$$f_A(0) = 1(p_A) + 1(p_C) = \frac{1 - p_C}{2} + p_C = \frac{1 + p_C}{2} \tag{2}$$

and the expected payoff for a B player at $t = 0$ is

$$f_B(0) = 2(p_B) - 4(p_C) = 2\frac{1 - p_C}{2} - 4p_C = 1 - 5p_C \tag{3}$$

Because this is a coordination game, if A has a higher initial fitness, the proportion of A players will grow and $f_A(t)$ will get only get larger while $f_B(t)$ gets smaller, until B becomes functionally extinct. Therefore, the fixation of A is favored over B if $f_A(0) > f_B(0)$, which can be solving for p_C using the equations above. We get the critical threshold for p_C, that is

$$p_C > \frac{1}{11} \approx 0.091. \tag{4}$$

We conclude that the network structure of the spatial game makes containing fake news significantly more challenging. In fact, between two to three times as many fact-checkers are needed to contain the sharing of fake news in small, isolated echo chambers, and even more fact-checkers are needed to have a good chance of driving fake news sharing behavior to total extinction.

As a note, observe that for very high values of p_C, the probability that real news dominates actually decreases. This seemingly paradoxical result can be explained by noting that for such high values of p_C, the population of real and fake

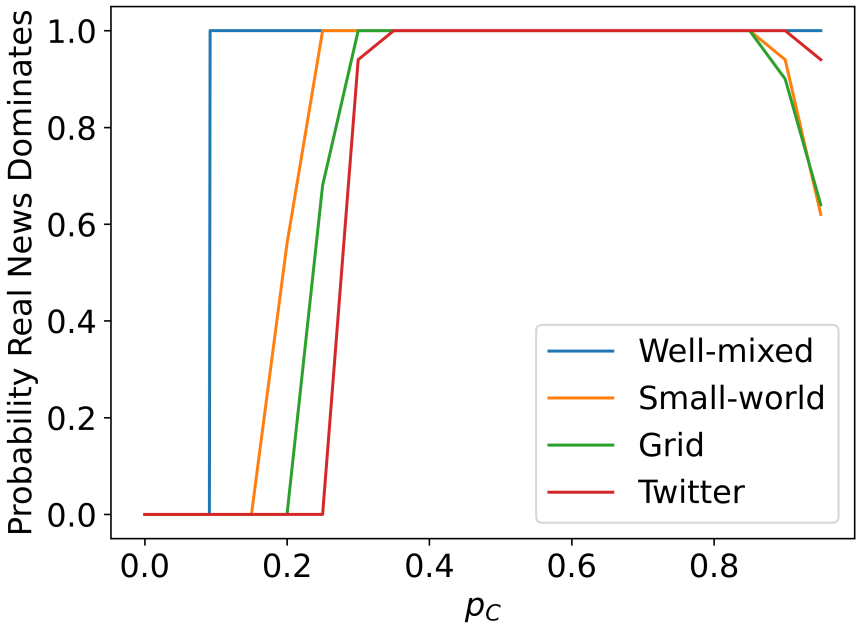

Figure 3: The probability that over half the viable population ends up sharing real news as a function of fact-checker density for different network topologies. The well-mixed result comes from analysis of replicator dynamics, and the rest come from simulations. For very high values of p_C, the spreader layer breaks apart into isolated individuals, at which point the dominant strategy is determined more by the random initialization than selection. The simulations consisted 50 populations at 20 evenly spaced fact-checker densities. At 5,000 time steps, a pseudo-steady state was declared and the simulation ended, except the Twitter simulations which ended at 500 time steps for computational reasons.

news sharers has completely broken down into small, disconnected components by the large number of static fact-checkers. These components typically have only one or two individuals, and are therefore completely constrained by their initial conditions. Selection cannot help individuals copy more beneficial strategies if there are no neighbors to copy. Fortunately, when selecting fact-checkers randomly, this only occurs for unrealistically high values of p_C.

3.3 Targeted Fact-checking

So far, we have only considered populations where fact-checkers are placed randomly. However, in almost all networks, some vertices are more centrally located than others, and this effect is particularly pronounced in naturally formed social networks. To improve the effectiveness of crowdsourced fact-checking with limited resources, it is vitally important to study targeted intervention algorithms by selecting the individuals that will have the most impact. Our results, shown in Figure 4, focus on two measures of network centrality, degree (the number of edges attached to a vertex) and betweenness [40], but there are many more centrality measures and the problem of selecting individuals for optimal fact-checking remains an open problem. Since all vertices in an infinite square lattice have the same centrality, our work here is restricted to small-world networks and the Twitter network.

Intuitively, selecting vertices with the largest degree will maximize the number of chances fact-checkers will have to punish fake news, because they will play more games against more opponents that the vertices with low degree. Betweenness centrality, on the other hand, will be selecting vertices that are most critical to transfering information between vertices. Thus, selecting by betweenness centrality could theoretically remove important pathways fake news needs to spread from one part of the population to another.

Figure 4 has several interesting features. First, we see that in small-worlds, using the degree and betweenness centralities have virtually the same performance. This is unsurprising in small-worlds as the additional shortcut edges are what create short path lengths and therefore give those individuals a high betweenness value, so the two centralities are highly correlated. More surprising is the fact that targeted fact-checking is only marginally more successful than random fact-checker placement, which can be seen by comparing Figures 3 and 4. This may be due to the relatively uniform nature of small-world networks, where there is little variation from vertex to vertex.

However, the Twitter network has much more diversity in its degree distribution and here we see a large change between random and targeted fact-checking. By targeting high degree or betweenness centrality individuals to be fact-checkers, we quickly separate the non-fact-checkers in the network into disconnected singletons and pairs, as these types of networks become disconnected very quickly when vertices with high degree are removed from the network [41]. Therefore, it is about equally likely that the initial random distribution will have more fake or real news sharers, so the probability that real news "dominates" by being present in over half the viable population hovers around 0.5 for almost

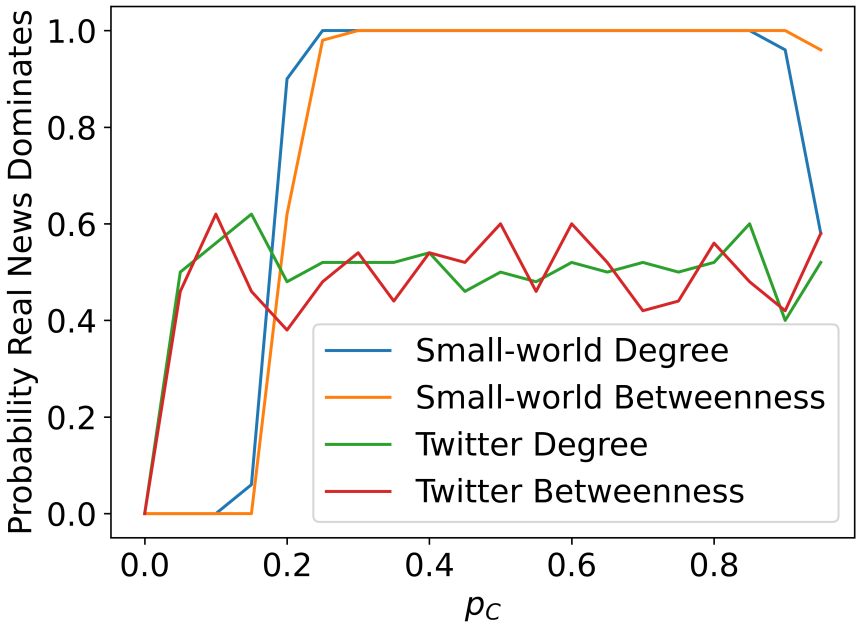

Figure 4: The probability of real news dominating on small-world networks and the Twitter network using the degree and betweenness centralities to place fact-checkers. Once again, we run simulations with 50 iterations, 20 density values, and a limit of 5,000 (or 500) time steps.

all values of fact-checker density. We observed a similar effect for high values of p_C in Figure 3.

The good news here is that a very small percentage of inoculated individuals (¡5%) is needed to break the paths of information transfer that fake news needs to spread. This suggests that in real world networks, a targeted crowdsourced fact-checking effort where fact-checkers are **also** encouraged to share real news with their neighbors could be highly effective with relatively little collective effort. In this scenario, the network structure will actually benefit real news instead of fake news by removing important vertices that fake news needs to move through to get to the rest of the population, while still allowing real news to spread. Enhancing our model by allowing fact-checkers to "pass along" real news between neighbors while still stopping fake news is one future path to more effectively study targeted fact-checking algorithms.

3.4 Analytic Results under Weak Selection

The selection strength β determines the effect payoff from the fake news game has on reproductive success. As β approaches zero [27, 42], the evolution of the system comes to resemble *neutral drift*, in which individuals choose strategy with no regard for payoff. In this domain, the pseudo-steady state with its echo chambers becomes short-lived, and the system quickly converges to all possible individuals sharing the same type of news. In this final section, we derive analytical results in this limit of weak selection.

Assuming a k-regular network structure like the square lattice, we will use an extended pair approximation method [43] to study the emergence and spread of real and fake news. The fixation probability of A is the probability that a population with some initial condition evolves so that the entire viable population eventually evolves to play A, and we present a closed-form expression for this probability in this section. Our aim here is to study the effects of changing the payoffs for real news, fake news, and fact-checkers, so we will begin with a general payoff matrix:

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} a & b & \alpha \\ c & d & \gamma \\ 0 & 0 & 0 \end{pmatrix} \end{array} \tag{5}$$

In the limit of weak selection $\beta \ll 1$, we will obtain conditions for the fixation probabilities of A and B as functions of these payoff values.

When we suppose that we begin with a fraction p of A individuals, we can calculate the expected value $m_A(p)$ and variance $v_A(p)$ of the change in abundance of A during the asynchronous update step where a single random individual considers changing strategy. The fixation probability of A for an initial fraction p of A players, denoted $\rho_A(p)$, satisfies the diffusion approximation

equation for large populations (see [27] for details):

$$m_A(p)\frac{d}{dp}\rho_A(p) + \left(\frac{v_A(p)}{2}\right)\frac{d^2}{dp^2}\rho_A(p) = 0 \tag{6}$$

with the boundary conditions $\rho_A(0) = 0$ and $\rho_A(1) = 1$. This equation has closed-form solution, and thus we can obtain an exact formula for ρ_A.

Our derivation of the following explicit expressions for the fixation probabilities in terms of the payoff values, lattice degree k, and fact-checker density p_c, is detailed in the Supplementary Information. The final result is that, for small values of p,

$$\rho_A(p) \approx p + \frac{\beta N p(1-p)}{6k}(-u_1 - 3u_2) \tag{7}$$

$$\rho_B(p) \approx p + \frac{\beta N p(1-p)}{6k}(-w_1 - 3w_2) \tag{8}$$

where $u_1 = (a-b-c+d)\Big(1 - k^2 - \frac{1+k}{(p_C-1)(1-p_C)}\Big)$, $u_2 = -a+b+c-d-ak+bk-bk^2+dk^2+(k-1)\Big(c+(b-\alpha+\gamma)k-d(1+k)\Big)p_C$, $w_1 = u_1$, and $w_2 = -(u_1+u_2)$.

In particular, we may be interested in the emergence of new behavior in a previously homogeneous population. We calculate the fixation probability ρ_A when beginning with a single initial A player, called the invasion probability, and derive the conditions for truthful behavior to be favored, that is, when $\rho_A > 1/N$ where N is the size of the population. We also repeat the process for a single B player. Using Equations (7) and (8), we show the effect p_C and γ, the punishment defectors suffer from fact-checkers, have on the invasion probabilities of real and fake news in Figure 5a.

This allows us to determine the conditions under which fact-checking will be effective at stemming misinformation and quantify how steep the penalty γ needs to be for a given proportion of fact-checkers, p_C, in the system. In Figure 5a, we see that for strong penalties, $\gamma < -4$, only a fifth of the population or less needs to be fact-checkers for selection to favor real news. However, if fact-checkers are less willing to publicly shame fake news spreaders and γ gets closer to zero, the number of fact-checkers need goes up to about half the population. The green region of the $p_C - \gamma$ plane shows where selection favors fake news; this only happens when there are very few fact-checkers. Notice that there is a wide region in orange where selection does not favor invasion by real or fake news. This is because the fake news game is a coordination game that tends to put minorities (like a single invading mutant) at a disadvantage. These analytic approximations closely match simulation testing, as shown in Figure 5b.

4 Discussion and Conclusion

This work adds to the growing body of research surrounding fake news, echo chambers, and fact-checking and we believe that it has immediate implications

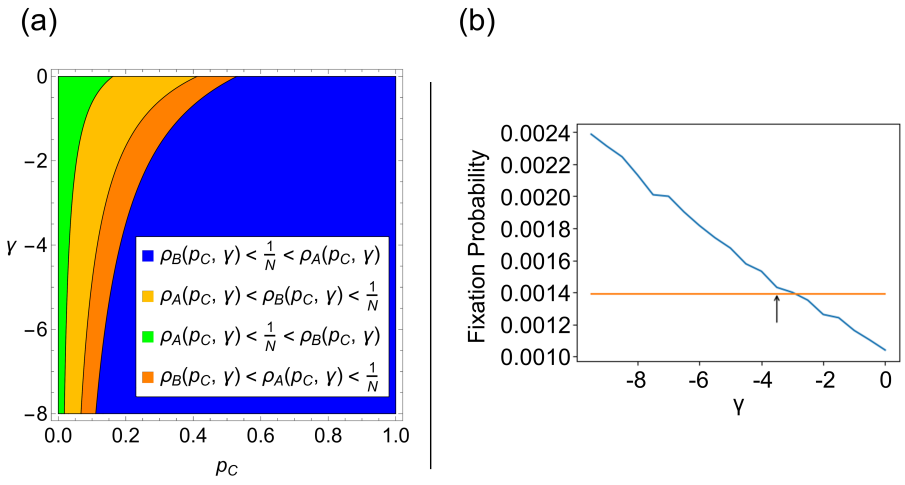

Figure 5: The invasion probabilities of real and fake news spreaders in the limit of weak selection using payoff values from (1), except for γ which varies from 0 to -8. In (a), we see what regions of the $p_C - \gamma$ plane give true stories an advantage (blue region), fake news an advantage (green region), or neither (orange regions). In (b), we see an approximation of the invasion probability for a single real news sharer from simulations, when $p_C = 0.2$ and $\beta = 0.0001$. These simulation results intersect the threshold line $\frac{1}{N} \approx 0.0014$ close to where it was predicted by the analytic results, indicated by the arrow.

for the study of misinformation. We have shown that the spatial structure of social networks tends to favor the spread of fake news, but by carefully selecting fact-checkers, that same structure can be used to combat misinformation by amplifying fact-checking efforts where they are most needed.

Our analytic results allow us to easily test potential combinations of reward and punishment and use both "carrots and sticks" to encourage real news and dampen fake news. Like previous work studying public goods games, we see that a strong punishment of defectors is effective at stopping bad behavior [44, 45, 46].

Future work combining potential experimental behavior data [36] with our present model will help incorporate relevant social network and psychological factors in our research. In particular, the constants in the payoff matrix and the selection strength were chosen fairly arbitrarily. Analyzing real-world data may allow us better estimates of some of these values, which in turn can give better actionable advice about how to control the spread of fake news. We would also like to analyze preexisting data sets or create new empirical studies to confirm our predictions regarding the effects that the rewards and punishments of sharing real and fake news have on the ability of fake news to spread through a population. As an example, perhaps placing fact-checking comments at the top of any fake news threads would sufficiently increase the punishment suffered by fake news's sharers to prevent its spread.

Recent theoretical research has demonstrated that partisan bias [47] and information cascades [48] are two possible explanations for the formation of echo chambers. Our work here shows that the spatial distribution of fact-checkers can also contribute to echo chamber creation, but this work only represents the first steps towards understanding how fact-checkers impact echo chamber formation. These echo chambers require certain conditions to form, including an appropriate selection strength, but there is much we still do not understand. Preliminary results show that the formation of resilient echo chambers is dependent on the type of network used. While social media sites do resemble lattices or small worlds in some respects, there are other properties of social networks that may be more or less conducive to echo chamber formation.

Extensions of our present work on targeted fact-checking efforts will likely lead to useful insights for optimizing field deployment of crowdsourcing fact-checking. There will be a good deal of further work to do, for example, on using other network topologies and other targeting centralities. In addition, the use of larger network data sets will give us more realistic behavior as there may be large-scale social network features essential to the development of echo chambers that are not captured in any of the network models we used.

Last but not least, our present work will help stimulate future work extending targeting algorithms to multiplex networks that take into account the fact that the interconnected ecosystems of social media platforms enable multi-channel communication and spillover from one platform to the other. In doing so, we hope to develop mechanistic models that allow us to explore realistic extensions incorporating social psychological factors such as heterogeneity of social influence, repeated exposure, and pre-existing beliefs.

5 References

References

[1] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[2] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287, 2018.

[3] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[4] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M. Mason. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon bombing. *iConference 2014 Proceedings*, page 654–662, 2014.

[5] John Woodrow Cox. 'we're going to put a bullet in your head': #PizzaGate threats terrorize D.C. shop owners, Dec 2016.

[6] Areeb Mian and Shujhat Khan. Coronavirus: The spread of misinformation. *BMC Medicine*, 18(1):89, 2020.

[7] Leonardo Bursztyn, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott. Misinformation during a pandemic. Working Paper 27417, National Bureau of Economic Research, 2020.

[8] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780, 2020.

[9] Gordon Pennycook and David G. Rand. Research note: Examining false beliefs about voter fraud in the wake of the 2020 presidential election. *Harvard Kennedy School Misinformation Review*, 2021.

[10] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

[11] Caitlin Dewey. Facebook fake-news writer: 'I think Donald Trump is in the White House because of me'. *The Washington Post*, Nov 2016.

[12] Rasmus Kleis Nielson. How to respond to disinformation while protecting free speech, Feb 2021.

[13] Lee Rainie, Janna Anderson, and Jonathan Albright. The future of free speech, trolls, anonymity and fake news online, Mar 2017.

[14] Alexander J. Stewart, Antonio A. Arechar, David G. Rand, and Joshua B. Plotkin. The coercive logic of fake news. *arXiv preprint abs/2108.13687*, 2021.

[15] Gordon Pennycook and David G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.

[16] Jon Roozenbeek and Sander van der Linden. The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5):570–580, 2019.

[17] Jon Roozenbeek and Sander van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1):65, 2019.

[18] William J. McGuire and Demetrios Papageorgis. Effectiveness of forewarning in developing resistance to persuasion. *Public Opinion Quarterly*, 26(1):24–34, 1962.

[19] John A. Banas and Stephen A. Rains. A meta-analysis of research on inoculation theory. *Communication Monographs*, 77(3):281–311, 2010.

[20] Shaoli Wang, Libin Rong, and Jianhong Wu. Bistability and multistability in opinion dynamics models. *Applied Mathematics and Computation*, 289:388–395, 2016.

[21] Xin Wang, Antonio D. Sirianni, Shaoting Tang, Zhiming Zheng, and Feng Fu. Public discourse and social network echo chambers driven by sociocognitive biases. *Physical Review X*, 10(4):041042, 2020.

[22] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[23] Tucker Evans and Feng Fu. Opinion formation on dynamic networks: Identifying conditions for the emergence of partisan echo chambers. *Royal Society Open Science*, 5(10):181122, 2018.

[24] Pew Research Center. Political polarization in the American public. Technical report, Pew Research Center, June 2014.

[25] Pew Research Center. Partisanship and political animosity in 2016. Technical report, Pew Research Center, June 2016.

[26] Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. Polarization of the vaccination debate on Facebook. *Vaccine*, 36(25):3606–3612, 2018.

[27] Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A. Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505, 2006.

[28] Corina E. Tarnita, Hisashi Ohtsuki, Tibor Antal, Feng Fu, and Martin A. Nowak. Strategy selection in structured populations. *Journal of Theoretical Biology*, 259(3):570–581, 2009.

[29] Martin A. Nowak and Robert M. May. Evolutionary games and spatial chaos. *Nature*, 359(6398):826–829, 1992.

[30] Feng Fu and Long Wang. Coevolutionary dynamics of opinions and networks: From diversity to uniformity. *Physical Review E*, 78(1):016104, 2008.

[31] Petter Holme and M. E. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):056108, 2006.

[32] Damián H. Zanette and Santiago Gil. Opinion spreading and agent segregation on evolving networks. *Physica D: Nonlinear Phenomena*, 224(1-2):156–165, 2006.

[33] Cecilia Nardini, Balázs Kozma, and Alain Barrat. Who's talking first? Consensus or lack thereof in coevolving opinion formation models. *Physical Review Letters*, 100(15):158701, 2008.

[34] Noah E. Friedkin, Anton V. Proskurnikov, Roberto Tempo, and Sergey E. Parsegov. Network science on belief system dynamics under logic constraints. *Science*, 354(6310):321–326, 2016.

[35] Chris G. Antonopoulos and Yilun Shang. Opinion formation in multiplex networks with general initial distributions. *Scientific Reports*, 8(1):2852, 2018.

[36] Gordon Pennycook, Tyrone D. Cannon, and David G. Rand. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12):1865–1880, 2018.

[37] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[38] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 4292–4293. AAAI Press, 2015.

[39] Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519, 2003.

[40] Marc Barthelemy. Betweenness centrality in large complex networks. *The European Physical Journal B - Condensed Matter*, 38(2):163–168, 2004.

[41] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[42] Martin A. Nowak, Akira Sasaki, Christine Taylor, and Drew Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428(6983):646–650, 2004.

[43] Tommy Khoo, Feng Fu, and Scott Pauls. Spillover modes in multiplex games: Double-edged effects on cooperation and their coevolution. *Scientific Reports*, 8(1):6922, 2018.

[44] K. Sigmund, C. Hauert, and M. A. Nowak. Reward and punishment. *Proceedings of the National Academy of Sciences*, 98(19):10757–10762, 2001.

[45] Karl Sigmund, Hannelore De Silva, Arne Traulsen, and Christoph Hauert. Social learning promotes institutions for governing the commons. *Nature*, 466(7308):861–863, 2010.

[46] Dirk Helbing, Attila Szolnoki, Matjaž Perc, and György Szabó. Punish, but not too hard: How costly punishment spreads in the spatial public goods game. *New Journal of Physics*, 12(8):083005, 2010.

[47] Mari Kawakatsu, Yphtach Lelkes, Simon A. Levin, and Corina E. Tarnita. Interindividual cooperation mediated by partisanship complicates Madison's cure for "mischiefs of faction". *Proceedings of the National Academy of Sciences*, 118(50):e2102148118, 2021.

[48] Christopher K. Tokita, Andrew M. Guess, and Corina E. Tarnita. Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*, 118(50):e2102147118, 2021.