Quantifying the Vulnerabilities of the Online Public Square to Adversarial Manipulation Tactics

Bao Tran Truong, Xiaodan Lou, Alessandro Flammini, Filippo Menczer

Observatory on Social Media

Indiana University, Bloomington

Abstract

Social media, seen by some as the modern public square, is vulnerable to manipulation. By controlling inauthentic accounts impersonating humans, malicious actors can amplify disinformation within target communities. The consequences of such operations are difficult to evaluate due to the challenges posed by collecting data and carrying out ethical experiments that would influence online communities. Here we use a social media model that simulates information diffusion in an empirical network to quantify the impacts of several adversarial manipulation tactics on the quality of content. We find that the presence of influential accounts, a hallmark of social media, exacerbates the vulnerabilities of online communities to manipulation. Among the explored tactics that bad actors can employ, infiltrating a community is the most likely to make low-quality content go viral. Such harm can be further compounded by inauthentic agents flooding the network with low-quality, yet appealing content, but is mitigated when bad actors focus on specific targets, such as influential or vulnerable individuals. These insights suggest countermeasures that platforms could employ to increase the resilience of social media users to manipulation.

Significance Statement: We show that social media users are vulnerable to adversarial manipulation tactics, through which bad actors can amplify exposure to content that potentially threatens democratic elections and public health. While tactics such as flooding the network with low-quality yet appealing content are damaging, getting users to follow inauthentic accounts has the most detrimental impact. Bad actors can increase harm by maximizing coverage rather than targeting particular individuals, such as influential ones. The varying degrees of harm associated with these tactics highlight tradeoffs and specific areas on which platforms could focus as they develop deterrence strategies against manipulation.

1 Introduction

The vision of social media as the modern *public square* has been challenged as users have become victims of manipulation by astroturf (1, 2), trolling (3), impersonation (4), and misinformation (5–7). False news have been reported to spread virally — similarly to reliable information (8) or even more (9) depending on operational definitions. These kinds of manipulation exploit a complex interplay of socio-cognitive (10, 11), ideological (7), and algorithmic (12, 13) biases. The exploitation is enabled or greatly facilitated by inauthentic accounts that impersonate people with malicious intent. Many such accounts can be coordinated by a single entity (14), either manually or through software applications commonly known as *social bots* (15, 16). Inauthentic and/or coordinated accounts have been observed to amplify disinformation (8), influence public opinion (14, 17–19), commit financial fraud (14, 20), infiltrate vulnerable communities (3, 21, 22), and disrupt communication (23, 24).

Can social media be manipulated to the point that they no longer function as a public square? Under what conditions? It is difficult to carry out empirical experiments and analyses in the real world to explore these questions (25, 26). One challenge is the limited size of experiments in the wild, stemming from both costs and ethical concerns about the potentially harmful nature of content from bad actors. A second difficulty is the limited data from social media platforms available to researchers (27), exacerbated by recent events such as the acquisition of Twitter by Elon Musk. These difficulties have led, for example, to conflicting accounts about whether disinformation campaigns on social media can sway elections (28–33). Evidence suggests that these operations mainly impact specific vulnerable communities (34, 35). However, we lack a comprehensive quantitative understanding of how coordinated inauthentic tactics can disrupt online communities. This prevents the informed design of moderation or regulatory policies to protect the online public square from adversarial manipulation.

Here we introduce *SimSoM*, a minimalistic model of a generic social media platform. The model allows us to explore scenarios in which an information-sharing network is manipulated by malicious actors controlling inauthentic accounts, and to measure the consequences of such information operations. We assume that bad actors aim to spread low-quality information. While there are different kinds of low-quality content in reality — disinformation, conspiracy theories, malware, or other harmful messages — our model uses an abstract definition of low-quality content that encompasses these different types. The impact of the manipulation is measured in terms of the quality of information to which users are exposed in the network.

We find that the presence of manipulation is sufficient to suppress quality information, driving low-quality content to spread virally in the network. We also examine network vulnerabilities that may amplify the effects of manipulation, and evaluate the overall information quality as the result of different malicious

tactics, such as infiltrating a community, generating attention-grabbing content, flooding the network, and targeting specific individuals. Insights from these analyses are instrumental in developing countermeasures to increase the resilience of social media and their users against manipulation. We discuss mitigation steps that platforms could take and the issues that arise from regulations aimed at protecting human speech from suppression.

Results

We model information diffusion in a social media platform such as Twitter/X, Instagram, or Mastodon. The information system is a directed network with nodes representing accounts and links representing follower relations. Similar to real-world platforms, content circulates through messages that appear in news feeds. Agents can post new messages or reshare messages from their feeds, generated by their friends, i.e., the accounts they follow. The information diffusion process is illustrated in Fig. 1. Messages represent information that could take the form of text, links, hashtags, images, or other media. An agent can introduce a new message into the system or, alternatively, select a message from their news feed to reshare. Messages created and reshared by an agent then appear on the news feeds of their followers.

Even though people prefer quality content (36), their sharing behavior is mediated by other factors such as laziness (37) and message appeal. To account for this, each message m in the model has two intrinsic and independent attributes. The *appeal* a_m models the likelihood that the message is actually reshared by agents (Fig. 1). The *quality* q_m, on the other hand, represents objective, desirable properties of content such as the originality of an idea or the accuracy of a claim. Here we naively represent quality as a scalar value. Deceptive posts may have low quality yet high appeal. For example, false news and junk science articles have low quality — most people would not share them knowingly. Yet such low-quality content may be even more likely to spread virally than high-quality information (9). Low-quality content may be novel, clickbait, ripped from headlines, and/or may appeal to people's political, emotional, or conspiratorial bias. Worse yet, bad actors can employ generative AI to produce such content at scale (38).

The model captures bias towards appeal as well as two other ingredients that are prioritized by the ranking algorithms of social media platforms, namely social engagement and recency (39). An agent selects a message to reshare from the news feed, which is an inventory of distinct messages recently shared by the agent's friends. The message is selected with probability proportional to: (i) its appeal; (ii) its social engagement, defined as the number of times it has been shared by the agent's friends; and (iii) its recency, which decreases with time in the feed (see details in Methods).

Unlike authentic agents, whose intention is to consume and share high-quality information, we define

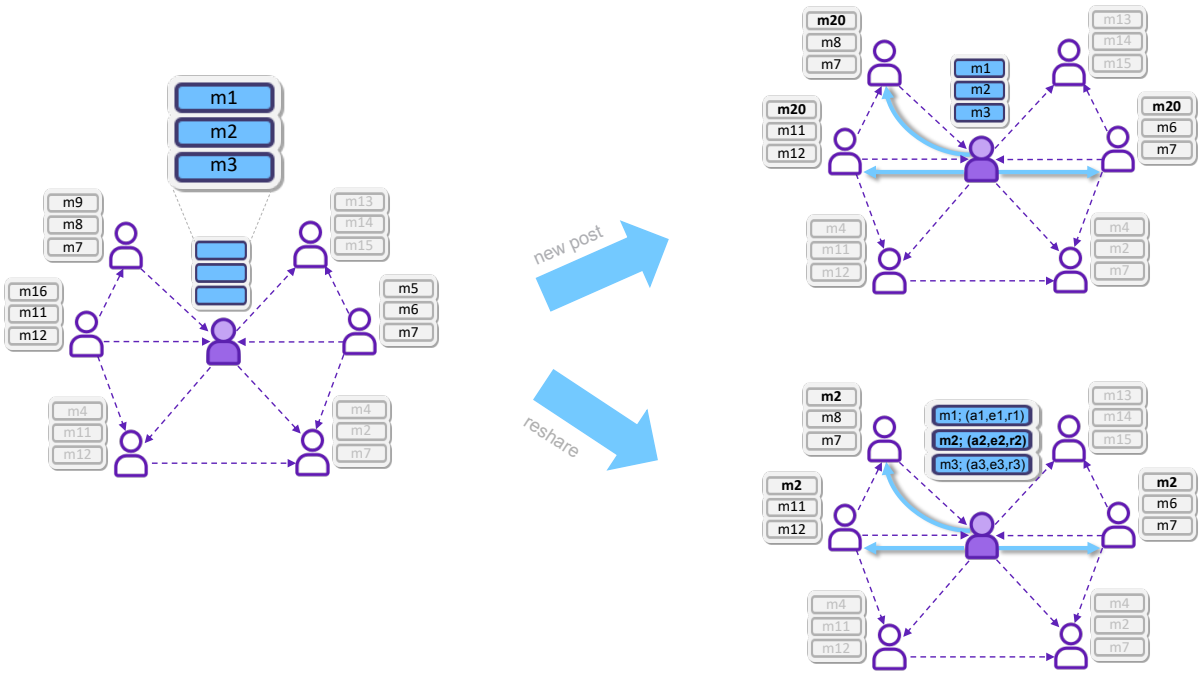

Figure 1: Illustration of the *SimSoM* model. Each agent has a limited-size news feed, containing messages posted or reposted by friends. Dashed arrows represent follower links; messages propagate from agents to their followers along solid links. At each time step, an active agent (colored node) either posts a new message (here, `m20`) or reposts one of the existing messages in their feed, selected with probability proportional to their appeal a, social engagement e, and recency r (here, `m2` is selected). The message spreads to the node's followers and shows up on their feeds.

inauthentic agents as accounts that are controlled by bad (adversarial) actors to spread low-quality content among authentic agents. We refer to these accounts as "bad actors" or "inauthentic agents" throughout this paper. Such accounts may be controlled by humans (trolls), software (social bots), or a mixture (cyborgs). The model has three parameters to model manipulation tactics by bad actors: *infiltration*, *deception*, and *flooding*, explained next.

Infiltration describes how bad actors amplify exposure to their messages by getting authentic accounts to follow them (Fig. 2(a)). Bad actor infiltration into the social network is modeled by a parameter γ, the probability that a bad actor is followed by an authentic agent. Unless otherwise stated, we assume that authentic agents follow bad actors uniformly at random. Fig. 2(b) illustrates the effective suppression of information quality when γ is high.

The quality q and appeal a of messages originating from authentic accounts are drawn independently from two distinct distributions, reflecting empirical evidence that these messages tend to have high quality and low appeal (see Supplementary Material). In contrast, we assume that bad actors can manipulate information in the network by creating messages with low quality ($q = 0$) and deceptively high appeal. The appeal

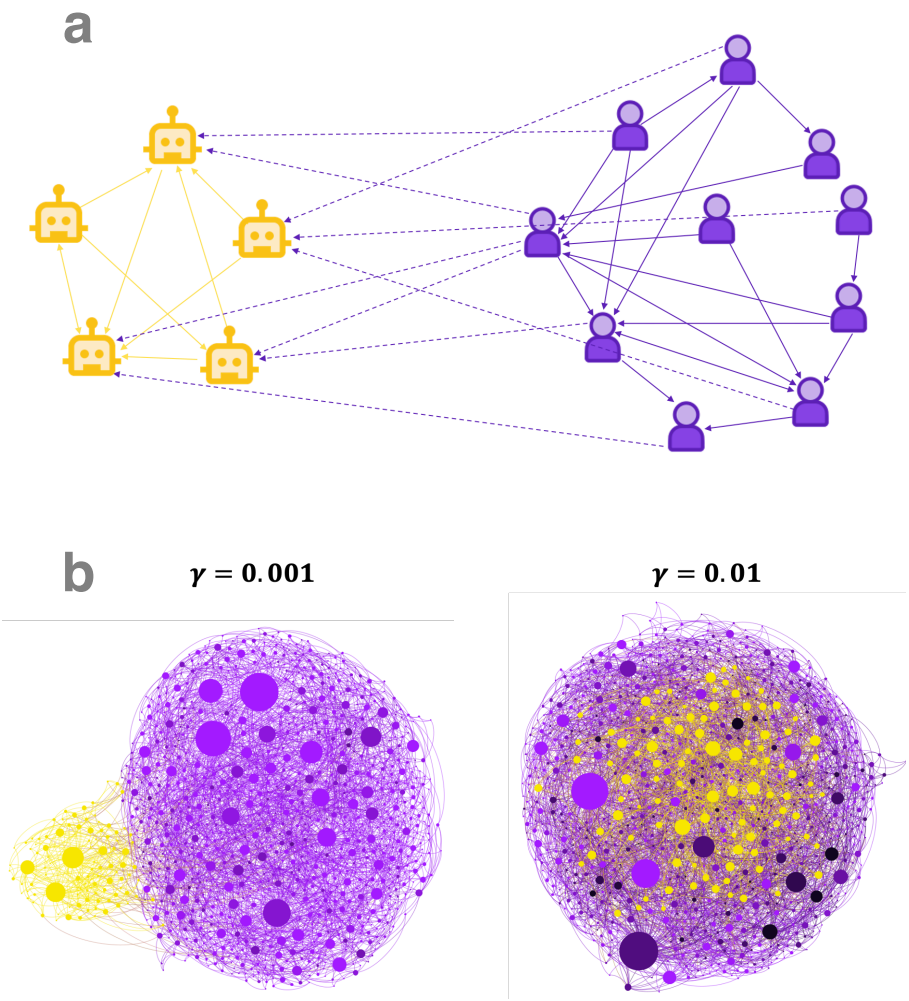

Figure 2: Subnetworks modeling authentic accounts (purple nodes) and bad actors (yellow nodes). (a) Illustration of the follower link structure. Solid links indicate follower relations within each subnetwork. Both subnetworks have hub and clustering structure that mimics or derives from online social networks. Dashed links represent authentic accounts following bad actors, according to the infiltration parameter γ, which represents the probability that an authentic node follows any given bad actor. When $\gamma = 0$ there is no infiltration and bad actors are isolated, therefore harmless; the opposite extreme $\gamma = 1$ indicates complete infiltration, such that bad actors dominate the network. (b) Effects of bad actor infiltration γ on the quality of messages in synthetic networks with 10^3 authentic agents and 100 inauthentic agents. For illustration purposes, both the authentic and inauthentic subnetworks in this panel are generated with the same method used for the inauthentic subnetworks in our experiments (see Methods). Node size represents the number of followers. The darker an authentic agent node, the lower the quality of messages in their feed.

differential of content from bad actors is modeled by the deception parameter ϕ, defined as the probability that bad actor content is irresistible ($a = 1$). In the absence of deception ($\phi = 0$), the appeal of bad actor messages is drawn from the same distribution as those from authentic accounts (see details in Methods).

Flooding is another tactic inauthentic accounts can use to amplify their influence, by crowding out high-quality information. To model this, the parameter θ is defined as the ratio between the exposure of bad actor content and authentic content (see Methods). Bad actors can achieve high exposure in several ways, including posting at high frequency/volume and artificially inflating popularity/engagement indicators (8, 38, 40).

SimSoM lets us explore information diffusion on social media, including the properties of authentic accounts that might render them vulnerable to adversarial attacks and the effects of different manipulation tactics. In the next sections, we present the results from simulations of the model on online communities derived from an empirical follower network ($N \approx 10^4$ Twitter accounts). The network has both scale-free structure (hubs) and high clustering (triangles), structural characteristics that are ubiquitous in socio-technical networks. It also has a realistic community structure, with two well-separated groups of accounts capturing political polarization (see Methods).

Once the system reaches a *steady state*, in which the message quality across the network has stabilized, we record the mean quality of the messages in the feeds of authentic agents. These measurements are further averaged across simulation runs with the same parameters but different random seeds. We simulate the information diffusion process in networks with different structures to explore whether social network features render communities vulnerable to bad actor tactics. Similarly, we evaluate the impact of these tactics by evaluating the model with varying levels of bad actor infiltration (γ), deception (ϕ), flooding (θ), and targeting specific types of accounts. We report on the *relative quality*, defined as the ratio of the average quality of authentic agents to that of a baseline without bad actors (see Methods).

Network Vulnerabilities

Key structural features of the social network may play a role in amplifying our vulnerability to manipulation by bad actors. The empirical network has two features that are ubiquitous in social media: the presence of hubs and highly clustered communities. We can explore how overall quality is affected by these features through three alternative networks constructed by shuffling links while preserving hubs, community structure, or neither (see Methods). Fig. 3 shows that community structure does not significantly affect the overall quality (purple vs. blue) because we assume that agents in each community are equally likely to follow bad actors. On the other hand, having hubs makes a network more vulnerable to manipulation: the relative system quality is significantly lower in the presence of hubs, both when there are communities (purple vs.

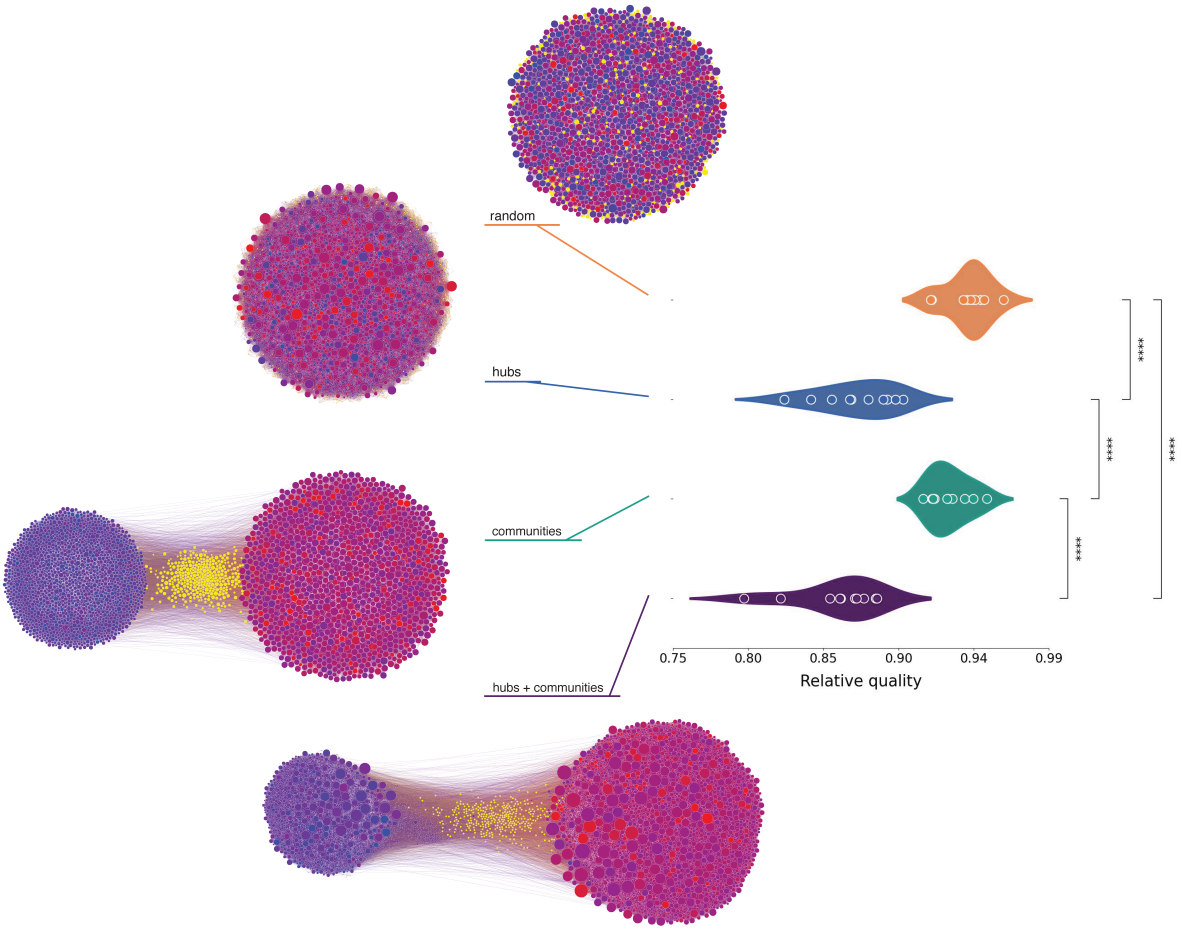

Figure 3: Impacts of different network structural features on the average information quality, relative to the scenario without bad actors. The original network ("hubs + communities") is visualized along with shuffled networks in which links from the original network are rewired while preserving communities, hubs, or neither ("random"). Node size and color represent, respectively, the number of followers of an account and their political leaning ranging from liberal to conservative (red to blue, see Methods). Yellow nodes are bad actors. Pairwise statistical significance is calculated using Welch's two-sided t-test (**** for $p < 10^{-4}$); only significant differences are reported.

green, $p < 10^{-4}$) and not (blue vs. orange, $p < 10^{-4}$). This is because node in-degrees and out-degrees are highly correlated (Spearman correlation 0.9, $p < 10^{-4}$) in the empirical network. Therefore, when authentic followers are concentrated among hubs, high-quality content is also concentrated among those hubs. This implies that high-quality content has more competition and becomes obsolete more quickly compared to the case in which this content is uniformly distributed among authentic nodes. The same does not apply to content from bad actors because this content spreads uniformly among authentic users. (The scenario in which content from bad actors mostly concentrates among hubs is explored later.)

Infiltration, Deception, and Flooding Tactics

Bad actors may maximize their message spread by combining various manipulation tactics. We systematically quantify the effects of these tactics through simulations varying the parameters for infiltration ($10^{-4} \leqslant \gamma \leqslant 10^{-1}$), deception ($0 \leqslant \phi \leqslant 1$), and flooding ($1 \leqslant \theta \leqslant 32$). See Methods for further details. Fig. 4(a,b,c) illustrates the effects of individual malicious tactics on the overall quality of information spreading through the network, compared to a baseline scenario without bad actors. We observe that infiltration is the most harmful manipulation tactic: when authentic agents have a $\gamma = 10\%$ probability of following each bad actor, the average quality in the system is reduced to less than half. Flooding and deception have smaller effects. When low-quality content has $\theta = 32$ times more exposure than authentic content, quality is reduced to less than 70%. When bad actors generate content with maximum appeal exclusively ($\phi = 1$), quality is reduced to about 70%.

Similarly, Fig. 4(d,e,f) shows the effects of pairs of tactics combined. Infiltration is dominant, but more harm can be done in combination with flooding or deception: the average quality is reduced to 40% when $\gamma = 0.1$ and $\theta = 32$ (Fig. 4(d)) or $\phi = 1$ (Fig. 4(e)). Combining flooding and deception ($\theta = 32, \phi = 1$, Fig. 4(f)) only results in marginal loss of quality (below 70%). With all three tactics combined ($\gamma = 0.1, \theta = 32, \phi = 1$, not shown), bad actors can further reduce the quality to 30%.

Reshare and Exposure Cascades

Empirical evidence has shown that among fact-checked claims, low-quality content (debunked claims) tends to have larger retweet cascades than high-quality content (confirmed claims) (9, 41). *SimSoM* allows us to examine how different factors may contribute to such a virality pattern. Fig. 5 illustrates the effects of bad actor tactics on the size of reshare cascades for content generated by both authentic agents ("high-quality") and inauthentic agents ("low-quality"). We observe that low-quality content can be boosted most effectively through high bad actor infiltration. This appears to be at odds with the hypothesis that attributes the

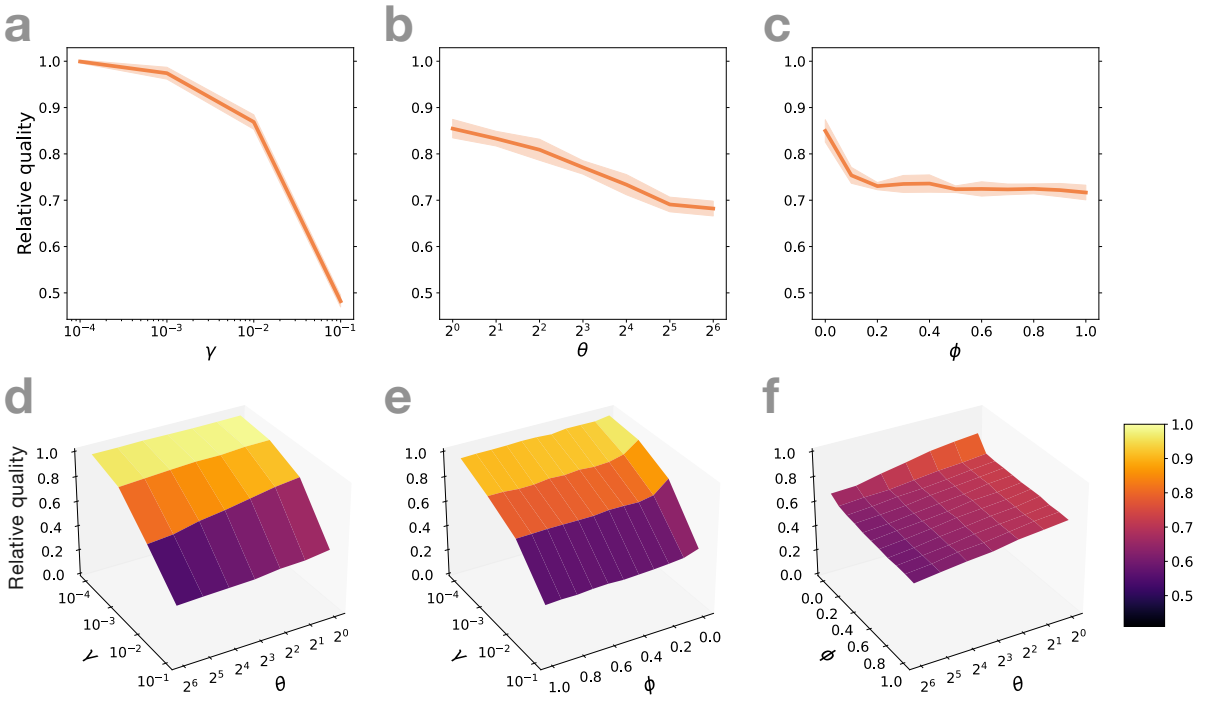

Figure 4: Effects of individual and combined tactics by bad actors on the system's message quality, relative to the scenario without bad actors. (a) Varying infiltration γ, without flooding ($\theta = 1$) or deception (ϕ=0). Shading represents 95% confidence intervals across runs in panels a–c. (b) Varying flooding θ with infiltration $\gamma = 0.01$ and no deception ($\phi = 0$). (c) Varying deception ϕ with infiltration $\gamma = 0.01$ and no flooding ($\theta = 1$). (d) Joint infiltration and flooding with no deception. (e) Joint infiltration and deception with no flooding. (f) Joint deception and flooding with infiltration $\gamma = 0.01$.

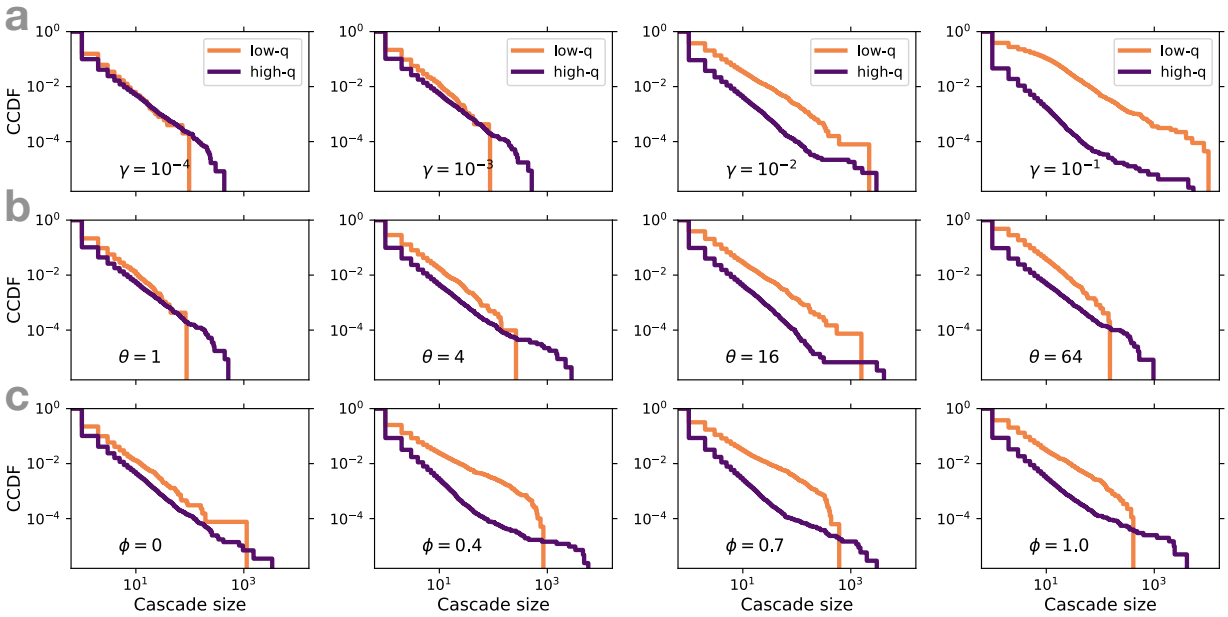

Figure 5: Complementary cumulative distributions of reshare cascade sizes for low- and high-quality content, generated by inauthentic and authentic agents, respectively. The plots are based on single runs of the model. (a) Effect of bad actor infiltration γ, with no flooding ($\theta = 1$) or deception ($\phi = 0$). (b) Effect of flooding θ, with low infiltration ($\gamma = 10^{-3}$) and no deception ($\phi = 0$). (c) Effect of deception ϕ, with low infiltration ($\gamma = 10^{-3}$) and no flooding ($\theta = 1$).

empirical difference in virality to factors like novelty, which make false news more appealing, rather than to inauthentic accounts such as social bots (9). In our model, even the highest boost in appeal ($\phi = 1$) does not grant low-quality cascades a virality comparable to those under higher bad actor amplification. A couple of observations reconcile the apparent contradiction. First, multiple factors may contribute to the virality of low-quality information, including inauthentic accounts and deceptively appealing content. Second, data from Twitter does not allow for the reconstruction of actual cascades. Therefore, empirical analyses may underestimate intermediary amplification by inauthentic accounts even when those accounts are removed. On the other hand, our model lets us reconstruct the likely reshare cascades that include intermediary amplification by inauthentic accounts.

To date, empirical data from social media platforms has allowed researchers to measure reshare cascades, but not exposure. Using the *SimSoM* model, it is possible to reconstruct not only likely reshare networks but also exposure networks, thus estimating how many accounts are exposed to (i.e., view) a message even if they do not reshare it (see Methods). Fig. 6 compares the sizes of reshare and exposure cascades. In general, reshares underestimate exposures by roughly one order of magnitude. Excluding the smallest and largest cascades, we observe that exposure networks grow sub-linearly with reshare networks: $s_v \sim s_r^{\nu}$ where s_v and

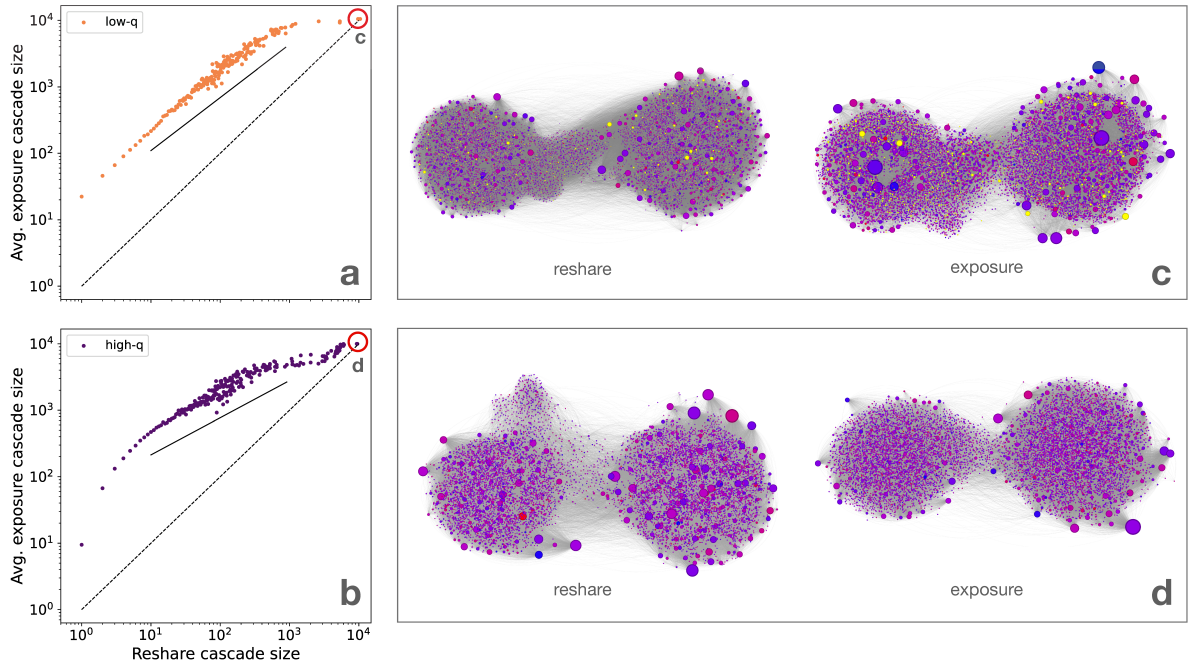

Figure 6: Scaling between reshare and exposure cascade sizes. (a) Scaling for low-quality messages (posted by inauthentic agents). (b) Scaling for high-quality messages (posted by authentic agents). The exposure cascade size is averaged across messages with the same reshare cascade size, based on 10 simulations. The dashed lines provide a linear scaling reference, while the solid lines show the slopes (exponents) ν of power-law fits for reshare cascades of size between 10 and 1,000, yielding $\nu = 0.80 \pm 0.01$ (low-quality messages) and $\nu = 0.56 \pm 0.01$ (high-quality messages). The largest reshare and exposure cascades (corresponding to the circles in panels a and b) are also visualized for (c) low-quality and (d) high-quality messages, based on one simulation. Node colors are the same as in Fig. 3; node size represents out-degree, or influence. Here we use $\theta = 1, \phi = 0, \gamma = 10^{-2}$; the results are similar for other γ values.

s_r are the exposure and reshare cascade sizes, respectively, and $\nu < 1$ is the scaling exponent. This means that as messages go viral, exposures grow more slowly than reshares. The exponent is higher for low-quality ($\nu \approx 0.8$) than for high-quality content ($\nu \approx 0.6$), suggesting that for each extra reshare, messages posted by inauthentic agents gain more views.

Targeting Tactics

The above results show that bad actors can use inauthentic accounts to infiltrate and disrupt an online public square. We have thus far assumed that all authentic agents have the same probability of following inauthentic ones, reflecting a scenario in which bad actors do not focus their efforts on specific potential followers. However, those interested in manipulating the network may want to maximize the spread of low-quality content through the community by targeting certain groups of accounts.

As an example, an adversary might target influentials based on the assumption that having such followers can multiply their impact — a message reshared by an influential account has a higher chance of going viral. Targeting influentials is well within the capability of bad actors and even automated accounts; the number of followers, often used as a proxy for influence (42), is public information on all social media platforms. A bad actor can easily interact with accounts having many followers by mentioning and/or following them (43, 44); other ploys include retweeting, quoting, and/or liking their tweets. There is empirical evidence of preferential targeting by bad actors that spread misinformation (1, 8).

Targeting politically active accounts or habitual misinformation spreaders are also conceivable tactics. An important question, then, is whether tactics targeting specific authentic accounts do in fact increase the manipulative power of bad actors. To explore this question, we introduce a preferential targeting tactic for adversarial actors in the model. Authentic agents have different probabilities of following inauthentic ones, proportional to one of five features that are available in the empirical data: number of followers (*hubs* tactic), propensity to share misinformation (*misinformation* tactic), political partisanship (*partisanship* tactic), or specific political leaning (*liberal* and *conservative* tactic). See Methods for details.

Fig. 7 shows the impact of these targeting tactics on information quality. Counterintuitively, preferential targeting is less harmful than random targeting: the distribution of quality is uneven so that the targeted population is worse off, but other parts of the community are spared. Targeting tactics therefore tend to backfire if we assume that bad actors intend to maximize the spread of their content across the full community.

Targeting hubs, for example, results in a network with significantly higher average quality ($p < 10^{-4}$). The amplification power of an influential is counterbalanced by the concentration of low-quality content, which has fewer chances to be reshared; the majority of other agents are left relatively free from manipulation. Note that this is analog to the reason why the presence of hubs leads to high-quality content being forgotten more quickly when hubs are not targeted by bad actors, as seen earlier.

The harm of manipulation through the network also diminishes when bad actors target accounts sharing a lot of misinformation ($p < 10^{-3}$) or with specific political leaning ($p < 10^{-4}$). The echo-chamber structure of the empirical network (Fig. 7(b)) helps interpret the latter finding: low-quality messages get shared and become obsolete rapidly within one densely connected partisan sub-community, sparing the rest of the network.

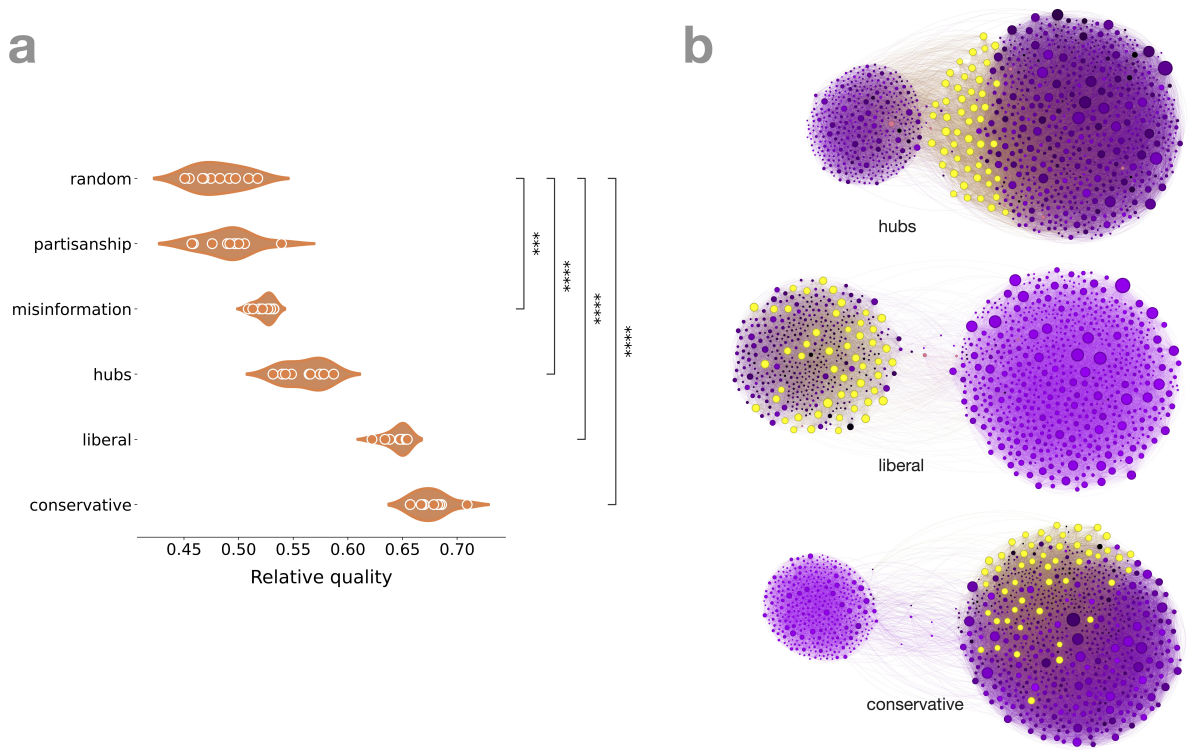

Figure 7: Effects of targeting tactics. (a) Average information quality resulting from each tactic, as well as the default random targeting, relative to the scenario without bad actors. We highlight significant differences calculated using Welch's two-sided t-test (*** for $p < 10^{-3}$ and **** for $p < 10^{-4}$). (b) Suppression of quality in the empirical network when bad actors specifically target influential accounts (hubs), and when they target politically left- (liberal) and right-leaning (conservative) accounts. The network has 10^3 authentic agents (purple nodes) and 50 inauthentic agents. Node size represents the number of followers. The darker an authentic agent node, the lower the quality of messages in their feed. Significant changes due to targeting tactics are only observed when bad actor infiltration is sufficiently high, therefore we use $\gamma = 10^{-1}$ in experiments for both panels.

Discussion

Social media platforms have enhanced the so-called *attention economy*, in which abundant content must compete for our scarce attention (45). But how to ensure that accurate, relevant, timely information wins this competition? To date, the policies that govern social media have been mainly guided by the concept of a *free marketplace of ideas*, rooted in John Milton's centuries-old reasoning that truth prevails in a free and open encounter of opinions (46). In an ideal world, the *wisdom of the crowd* (47) would realize this vision by combining the opinions of many users (48). Unfortunately, several aspects of modern social media challenge the illusion of a public square or marketplace in which wise crowds access and select quality information (49, 50). First, information production is affected by information consumption (51), creating incentives to produce appealing but not necessarily high-quality content. Second, social influence undermines the wisdom of crowds

because the information and opinions to which we are exposed online may be inauthentic, correlated through coordination (52, 53), or dominated by few influential individuals (54). Third, confirmation bias (55) can increase vulnerability to misinformation in social media (11, 56). Finally, the structure of information flow networks can distort perceptions and increase vulnerability to malicious actors (57, 58).

Exploration using *SimSoM* quantifies how manipulation by bad actors can prevent social media from functioning as a public town square. Simulations of the model reveal that making low-quality content highly appealing plays a lesser role compared to other harmful tactics by inauthentic accounts. This suggests that novelty, for example, may not provide the primary explanation for the virality of fake news, as previously hypothesized (9). Tactics that inflate engagement indicators, such as flooding, can erode the system's quality just as well. More importantly, we find that infiltrating the network is a dominant harmful tactic available to bad actors — these accounts can drastically suppress quality by inducing only a small fraction of the community into following them.

A wealth of previous models and experiments have focused on information diffusion and popularity. Several studies have investigated the role played by network mechanisms affecting the popularity of individual posts, including exogenous and endogenous bursts of attention (59, 60), memory (61), novelty (62, 63), and position bias (64, 65). This literature considers the popularity of pieces of information in isolation. Market-like environments in which *many* messages compete for limited attention have received less consideration. Exceptions have considered the cost of learning about quality (66), distortions of quality assessments that result from aggregate knowledge of peer choices (52), and confirmation bias in the spread of misinformation (67). The *SimSoM* model proposed here extends the model of Weng *et al.* (68), who demonstrated that some posts inevitably achieve viral popularity irrespective of quality in the presence of competition among networked agents with limited attention. The model was formalized as a critical branching process and studied analytically, predicting that the popularity of posts follows a power-law distribution with heavy tails (69–71).

The current *SimSoM* model has several limitations. First, it assumes that all authentic agents are similar, when in reality users have diverse behaviors. For example, in the U.S., conservative social media users are more likely to share low-credibility content (7, 44, 72). Large language models have recently been proposed as a way to model agents with more realistic and heterogeneous behaviors (73). Second, *SimSoM* models information diffusion through simple contagion. While this approach is supported by empirical evidence (41, 74), complex contagion (25) can also play a role in the spread of certain types of harmful information (26). Finally, *SimSoM* neglects many mechanisms of actual online social media that may contribute to message exposure, such as search and the secret algorithmic details of each platform. Yet the model captures several universal ingredients of socio-technical networks, including selection criteria (appeal,

social engagement, and recency) and limited individual attention. The empirical follower network also accounts for structural features (hubs, clustering, polarized communities) that play a key role in information diffusion (72, 75–78). The model's predictions are consistent with empirical findings about the difference in virality or lack thereof between low- and high-quality information on Twitter (8, 9). This suggests a reasonable balance between model realism and generality.

We can think of inauthentic accounts as zealots, a minority of agents committed to a particular view. Opinion dynamics models have shown that a critical minority of active zealots can quickly drive a system to consensus toward their opinion (79–82). In our social media model, we only explore the capacity to suppress information quality rather than to drive consensus to a particular opinion.

Our results suggest that, surprisingly, inauthentic accounts do not need to target hubs (influentials); they can do more damage by connecting to random accounts. Future research should focus on whether the strategic placement of bad actors within/across polarized online communities can sway users toward a particular outcome, for example by distorting popularity perceptions (57, 58). Further work is also needed to characterize the effects of strategic targeting when the attacker has limited resources.

The insights gained from the present findings suggest several countermeasures to increase the resilience of social media to manipulation. The first lesson is that we must make it more difficult for bad actors to infiltrate the network. Platform efforts to detect and take down deceptive accounts must be strengthened, especially as tactics to hack follower networks get more sophisticated (43).

Recent models show that caps on the depth and/or breadth of diffusion networks can decrease the ratio of distorted messages received by social media users (83). While these models assume that distortions occur randomly, *SimSoM* demonstrates how adversarial actors can exploit vulnerabilities by flooding our news feeds, thereby crowding out quality information. A countermeasure would be to challenge accounts that post at very high rates to prove that they are human. Users could also be warned when they follow accounts that post low-quality content and/or when friend accounts are suspended or take suspicious actions, such as changing names/handles (14) or posting and deleting large volumes of content (40).

Our findings suggest that a winning tactic for malicious accounts is to target all social media users, and not only the most influential, if their goal is to spread harm widely across the network. Literacy programs may provide some protection against disinformation. This is also suggested by a binary agreement model that assumes each agent may be committed to truth ordisinformation (84). Social media platforms could lead by educating users about their vulnerability to manipulation and deception. The simplest version of this is through accuracy reminders, which can improve content quality during both posting (85) and resharing (86). Flooding reduces the effectiveness of such accuracy interventions (87). Our model could be extended to explore ways to combine friction and accuracy nudges by, e.g., lowering the probability that

low-quality messages are reshared by agents after they are exposed to warnings.

Given that inauthentic accounts can be used to suppress human speech, granting them (or any entity that controls them) unlimited free-speech rights would seem to lead to a logical contradiction (88). Yet, ironically, efforts by social media platforms to moderate abusive accounts and even research on social bot detection have been assailed by some with charges of censorship (89). These questions may have significant repercussions on regulations designed to protect the online public square (49).

Methods

Social Media Diffusion Model

SimSoM[1] is a parsimonious agent-based model inspired by the long tradition of representing the spread of ideas as an epidemic process where messages are passed along the edges of a network (90). The model simulates a directed follower network, as in Twitter/X, Mastodon, or Threads. Nodes represent agents (users) and links represent follower relations, which may or may not be reciprocal. The direction of a link goes from the follower to the followed (friend) account, capturing the flow of attention; when a friend's post is reshared by a follower, information spreads in the opposite direction. In line with previous work (41, 74, 91), the diffusion process is modeled as simple contagion, where each exposure to a message results in the same resharing probability. We assume that the structure of the network is static (no unfollowing/blocking of accounts) during the information-spreading process. In contrast to classical epidemiological models, new messages are continuously introduced into the system in an exogenous fashion.

At each time step, an agent i produces a new message with probability μ or chooses one of the messages in their news feed to be reshared with probability $1-\mu$ (Fig. 1). The new or reshared message is then added to the news feeds of i's followers. We set $\mu = 0.5$, reflecting the empirical average ratio for English-language tweets (92). Based on empirical data, agents are assumed to have limited-size inventories: only the most recent $\sigma = 15$ messages are retained in each news feed (see Supplementary Material). The results are robust for different values of μ and σ (see Fig. S1 in Supplementary Material).

We assume authentic agents prefer to reshare messages posted by their friends that are appealing, recent, and popular. This is based on empirical evidence that users are more likely to share popular content according to engagement signals (93). The probability that an agent shares a message from their news feed, allowing it to spread, is proportional to the message's appeal, social engagement, and recency. More explicitly, let M_i be the feed of i ($|M_i| = \sigma$). The probability of message $m \in M_i$ being selected is $P(m) = a_m e_m r_m / \sum_{j \in M_i} a_j e_j r_j$

[1] Code and data to implement the model and reproduce results are available at `github.com/osome-iu/SimSoM`

where a_m is the appeal of message m, e_m is the social engagement, i.e., the number of times it was (re)shared by i's friends, and r_m is the recency. Message recency decays with time as a stretched exponential function $r_m(t) = e^{-0.4t^{0.4}}$, where t is the "age," or the time passed since m was first introduced to the agent's news feed. This decay function is based on empirical online news engagement data (63). To model flooding, content from bad actors has exposure that is θ times higher than authentic content.

Quality and Appeal

Both the quality q and the appeal a of new messages are defined in the unit interval. Informed by empirical data, q and a for authentic accounts are assumed to be independent (see Fig. S2 in Supplementary Material). For authentic accounts, quality q is drawn from an exponential distribution $P(q) = Ce^{-\tau q}$ where high-quality information is more common than low-quality information. The term $C = \frac{\tau}{1-e^{-\tau}}$ is a normalizing constant such that $\int_0^1 P(q)dq = 1$; the exponent $\tau = -10$ is estimated empirically (see Fig. S3 in Supplementary Material). We independently draw appeal from the distribution $P(a) = (1 + \alpha)(1 - a)^\alpha$, where $\alpha > 1$ captures the rarity of appealing messages (Fig. S4(a) in Supplementary Material). We set $\alpha = 4$. This choice for the appeal probability density function reproduces a broad distribution of reshares comparable to that observed in real-world information diffusion networks: few messages go viral while the majority do not (Fig. S4(b) in Supplementary Material).

On the other hand, we assume that bad actors strictly generate low-quality messages ($q = 0$). The potentially deceptive nature of this content is modeled by the deception parameter ϕ ($0 \leqslant \phi \leqslant 1$), the probability that a bad actor message is irresistibly appealing. With probability ϕ, we set $a = 1$, and with probability $1-\phi$ we draw appeal from the same distribution as for authentic accounts, $P(a) = (1+\alpha)(1-a)^\alpha$. If $\phi = 0$, bad actors and authentic accounts generate messages with appeal drawn from the same distribution. If $\phi > 0$, bad actor messages are more likely to have high appeal; the larger ϕ, the greater the potential virality of low-quality content by bad actors.

Empirical Follow Network

We run simulations on a follower network derived from empirical Twitter data. This network was constructed from a 10% random sample of public tweets between June 1–30, 2017 (94). The data includes users (excluding likely automated accounts) that shared at least ten links to news sources, at least one of which was to a source labeled as low-quality. Only news sources with known political valence were considered. This sampling procedure captures a community of accounts that are both active and vulnerable to misinformation. Based on the shared links, most accounts in the dataset have an associated *partisanship score* (from -1 for left-

leaning to +1 for right-leaning) defined as the average political bias of the news sources they share; and a *misinformation score* defined as the fraction of posts linking to low-credibility sources. Both partisanship and misinformation scores, included in the original network dataset, were based on news source labels from third-party fact-checkers (72).

From the original dataset, we select nodes with both partisanship and misinformation attributes. We further reduce the size of the network to speed up our simulations. We apply k-core decomposition to select $N = 10{,}006$ nodes forming the $k = 94$ core. Finally, we remove a random sample of edges to decrease the density of this core while preserving the average in/out-degree (number of friends/followers) of the original network ($k = 0$ core). This results in $E = 1{,}809{,}798$ edges; each node has on average approximately 180 friends/followers (Fig. 3, "hubs+communities" network).

Bad Actor Subnetwork

Since the empirical network described above does not include likely inauthentic accounts, we use it to model the subnetwork of authentic accounts. We then add a subnetwork representing inauthentic accounts that infiltrate the system. The ratio between the sizes of the two subnetworks is described by β, i.e., for an authentic agent subnetwork of N nodes, the inauthentic subnetwork is composed of βN nodes (Fig. 2). Since there are many types of inauthentic accounts (trolls, social bots, cyborgs), estimating the percentage of these on social media is a very difficult task. We thus set $\beta = 0.05$ following a rough estimation of the prevalence of bots on Twitter.[2] Note that from the perspective of information quality in the model, increasing the prevalence of bad actors is equivalent to increasing their infiltration. Therefore we focus on the effects of varying γ rather than β.

The model assumes that inauthentic accounts follow each other to amplify low-quality content. To capture the characteristic presence of hubs and clustering (directed triads), the inauthentic subnetwork is created using a directed variant of the random-walk growth model (95). Specifically, the network is initialized with four fully connected nodes. We then add one new node at a time, assuming that each has fixed out-degree $k_{out} = 3$. Once a new node i comes into the network, it links to (follows) a randomly selected target node (friend) j. Each of the remaining $k_{out} - 1 = 2$ friends are selected as follows: with probability expressed by a parameter p, i follows a random friend of j's; with probability $1 - p$, i follows another randomly selected node. Following friends of a friend has the effect of generating closed, directed triads and approximates a preferential attachment process, giving rise to hub nodes with high in-degree. The parameter p thus models both hubs and clustering. We use $p = 0.5$.

In addition, the bad actor subnetwork is designed to manipulate information flow and collective attention

[2] theconversation.com/how-many-bots-are-on-twitter-the-question-is-difficult-to-answer-and-misses-the-point-183425

by spreading certain messages. Therefore we assume that bad actors get random authentic accounts to follow them: we add a directed link from each authentic node to each bad actor node with probability γ. (In the next subsection we present alternatives to this random targeting tactic.) The parameter γ models the degree of infiltration of the network by bad actors (Fig. 2). When $\gamma = 0$, there is no infiltration and bad actors are isolated, therefore harmless; the opposite extreme $\gamma = 1$ indicates complete infiltration such that bad actors dominate the network. Because we are not concerned with the quality of messages on bad actor news feeds, they do not follow or reshare content from authentic agents for the analyses in this paper.

Bad Actor Targeting Tactics

In the scheme described above, the authentic agents that follow bad actors do so randomly. We also study scenarios in which bad actors target certain accounts as potential followers (Fig. 7). In all of these scenarios, each bad actor is still followed by γN authentic accounts on average, but these targets are selected according to some criterion, such as whether they are politically active or have many followers.

Each targeting tactic is modeled by making the probability that an authentic agent i follows bad actors proportional to some feature $f(i)$. In the hubs tactic, we set $f(i) = k_{in}(i)$, the number of followers of i. Misinformation and partisanship attributes of authentic accounts in the empirical network allow us to model other targeting scenarios. In the misinformation tactic, $f(i)$ is set to i's misinformation score, i.e., their propensity to share misinformation. The partisanship score is used in the liberal and conservative tactics, while its absolute value is used in the partisanship tactic.

Overall Quality

The effects of manipulation on authentic agents are quantified by the quality of all content in circulation. The overall quality of the information system at time t is measured by the *average quality* across all the messages visible through the feeds of the authentic agents:

$$Q_t = \frac{1}{\sigma N} \sum_{i=1}^{N} \sum_{m \in M_i} q_{i,m,t},$$

where $q_{i,m,t}$ is the quality of the message in the m-th position in authentic user i's feed at time t.

The system's quality at each time step t is calculated with an exponential moving average $\bar{Q}_t = \rho \bar{Q}_{t-1} + (1 - \rho) Q_t$. As the simulation takes place, some of the messages become obsolete quickly, while others live longer and infect a larger fraction of the network. The simulation ends once the system reaches a *steady state* in which the difference between the average quality in two consecutive time steps is smaller than a

threshold, i.e., $|\bar{Q}_t - \bar{Q}_{t-1}|/\bar{Q}_{t-1} < \epsilon$. The average quality reported for all analyses is calculated at the end of the simulation.

A *reshare cascade* is a tree that begins when an agent i (root) posts a new message m. After that, a new link is created when a follower j of i reshares m. We say that i is the parent of j in the tree. Similarly, an agent's exposure to a message is defined as having that message on their feed while being activated — we assume that an agent who is about to share or reshare a post has seen the content on their feed. A link in the *exposure cascade* for message m is created between an activated agent with m in their feed and their friend who had shared m. When constructing both reshare and exposure cascades, m may have been reshared by more than one of an agent's friends. When this occurs, one of them is selected at random to be the parent in the cascade tree. This choice does not affect a cascade tree's size, but might affect its structure. Other alternatives, such as selecting as parent the friend who most recently shared m, can be explored if one desires to analyze the structure of the reshare or exposure cascades. By definition, reshare cascades can be as small as one and exposure cascades can be as small as zero.

To capture the complete diffusion cascades, the size distributions plotted in Fig. 5 include only cascades of "extinct" messages, those that have become obsolete and are no longer present in any news feed by the end of simulations.

Simulation Framework and Parameters

For each set of parameters, we run 10 simulations starting from random conditions; the reported average quality is the average across many runs.

In scenarios where there are bad actors in the system, the default bad actor parameters are $\gamma = 0.01$, $\phi = 0$, and $\theta = 1$. Except for the simulations exploring inauthentic targeting tactics, authentic followers of the bad actors are selected at random.

The parameters $\rho = 0.8$ and $\epsilon = 0.0001$ were tested to ensure that the system's quality stabilizes at the steady state.

We shuffle the empirical follower network in various ways to derive the scenarios reported in Fig. 3. The "hubs+communities" network is the original one. In the "hubs" network, we shuffle the original network while preserving the degree distribution and destroying any community clustering. In the "communities" network, the outgoing endpoint of each edge in the original network is rewired to another node within the same community; this preserves the community structure and destroys the hub structure. In the "random" shuffle, all edges are rewired at random with uniform probability, destroying all hub and clustering structure.

Acknowledgments

We are grateful to Marshall Van Alstyne, David Axelrod, Rachith Aiyappa and Erfan Samieyan for useful discussion and suggestions; to Dimitar Nikolov, Chengcheng Shao, Giovanni Ciampaglia, and Pik-Mai Hui for the data collection used to construct the empirical network; to Kai-Cheng Yang and Christopher Torres-Lugo for the COVID-19 data collection; and to Alireza Sahami Shirazi for data about scrolling session depth on a social media mobile app. This work was carried out in part while XL visited the Observatory on Social Media at Indiana University, with support by the China Scholarship Council. BT, AF and FM were supported in part by The Knight Foundation. BT was funded in part by the Ostrom Workshop. AF and FM were funded in part by DARPA (grants W911NF-17-C-0094 and HR001121C0169). FM and BT were supported in part by the Swiss National Science Foundation (Sinergia grant CRSII5_209250). FM was supported in part by Democracy Fund and Craig Newmark Philanthropies. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Detecting and tracking political abuse in social media. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[2] Panagiotis T Metaxas and Eni Mustafaraj. Social media and the elections. *Science*, 338(6106):472–473, 2012.

[3] Leo G Stewart, Ahmer Arif, and Kate Starbird. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM Workshop on Misinformation and Misbehavior Mining on the Web*, 2018.

[4] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. Acting the part: Examining information operations within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):20, 2018.

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

[6] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Anatomy of an online misinformation network. *PLoS ONE*, 13(4):e0196087, 2018.

[7] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378, 2019.

[8] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kaicheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9:4787, 2018.

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

[10] Herbert Lin. The existential threat from cyber-enabled information warfare. *Bulletin of the Atomic Scientists*, 75(4):187–196, 2019.

[11] Filippo Menczer and Thomas Hills. The attention economy. *Scientific American*, 323(6):54–61, Dec 2020.

[12] Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*, 8:15951, 2018.

[13] Dimitar Nikolov, Mounia Lalmas, Alessandro Flammini, and Filippo Menczer. Quantifying biases in online information exposure. *Journal of the Association for Information Science and Technology*, 70(3):218–229, 2019.

[14] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. Uncovering coordinated networks on social media: Methods and case studies. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 455–466, 2021.

[15] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Comm. ACM*, 59(7):96–104, 2016.

[16] Kai-Cheng Yang, Onur Varol, Clayton A. Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019.

[17] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, 21(11), 2016.

[18] Massimo Stella, Marco Cristoforetti, and Manlio De Domenico. Influence of augmented humans in online interactions during voting events. *PLOS ONE*, 14(5):1–16, 2019.

[19] Emilio Ferrara. Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *First Monday*, 22(8), 2017.

[20] Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Identifying and analyzing cryptocurrency manipulations in social media. Preprint 1902.03110, arXiv, 2019.

[21] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440, 2018.

[22] Guido Caldarelli, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco. The role of bot squads in the political propaganda on twitter. *Communications Physics*, 3:81, 2020.

[23] Norah Abokhodair, Daisy Yoo, and David W McDonald. Dissecting a social botnet: Growth, content and influence in twitter. In *Proc. 18th ACM Conf. on Computer Supported Cooperative Work & Social Computing (CSCW)*, pages 839–851, 2015.

[24] Pablo Suárez-Serrato, Margaret E. Roberts, Clayton Davis, and Filippo Menczer. On the influence of social bots in online protests. In Emma Spiro and Yong-Yeol Ahn, editors, *Social Informatics: Proc. 8th International Conference (SocInfo), Part II*, volume 10047 of *Lecture Notes in Computer Science*, pages 269–278, 2016.

[25] Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119, 2018.

[26] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. Evidence of complex contagion of information in social media: An experiment using twitter bots. *PloS one*, 12(9):e0184148, 2017.

[27] Irene V. Pasquetto, Briony Swire-Thompson, et al. Tackling misinformation: What researchers could do with social media data. *HKS Misinformation Review*, 1(8), 2020.

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

[29] Kathleen Hall Jamieson. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President.* Oxford University Press, 2018.

[30] A. Badawy, E. Ferrara, and K. Lerman. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265, 2018.

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

[32] R. Kelly Garrett. Social media's contribution to political misperceptions in U.S. Presidential elections. *PLOS ONE*, 14(3):1–16, 2019.

[33] Damian J. Ruck, Natalie Manaeva Rice, Joshua Borycz, and R. Alexander Bentley. Internet Research Agency Twitter activity predicted 2016 U.S. election polls. *First Monday*, 24(7), 2019.

[34] Christopher A. Bail, Brian Guay, Emily Maloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences*, 117(1):243–250, 2020.

[35] Gregory Eady, Tom Paskhalis, Jan Zilinsky, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14(1):62, 2023.

[36] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, 2021.

[37] Gordon Pennycook and David G Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019.

[38] Kai-Cheng Yang and Filippo Menczer. Anatomy of an ai-powered malicious social botnet. Preprint 2307.16336, arXiv, 2023.

[39] Twitter. Twitter's recommendation algorithm. `blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm`, 2023.

[40] Christopher Torres-Lugo, Manita Pote, Alexander Nwala, and Filippo Menczer. Manipulating Twitter through Deletions. In *Proc. Intl. AAAI Conf. on Web and Social Media (ICWSM)*, volume 16, pages 1029–1039, 2022.

[41] Jonas L. Juul and Johan Ugander. Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences*, 118(46):e2100786118, 2021.

[42] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[43] Christopher Torres-Lugo, Kai-Cheng Yang, and Filippo Menczer. The manufacture of partisan echo chambers by follow train abuse on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1017–1028, 2022.

[44] Wen Chen, Diogo Pacheco, Kai-Cheng Yang, and Filippo Menczer. Neutral bots probe political bias on social media. *Nature Communications*, 12:5580, 2021.

[45] H. Simon. Designing organizations for an information-rich world. In Martin Greenberger, editor, *Computers, Communication, and the Public Interest*, pages 37–52. The Johns Hopkins Press, Baltimore, 1971.

[46] John Milton. Areopagitica. Dartmouth's Milton Reading room. Accessed online at www.dartmouth.edu/˜milton/reading_room/areopagitica/text.shtml, 1644.

[47] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[48] Scott E Page. *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press, 2008.

[49] Herbert Lin. On the Organization of the U.S. Government for Responding to Adversarial Information Warfare and Influence Operations. *I/S: A Journal of Law and Policy for the Information Society*, 15:1–43, 2019.

[50] Giancarlo Ruffo, Alfonso Semeraro, Anastasia Giachanou, and Paolo Rosso. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer Science Review*, 47:100531, 2023.

[51] Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. The production of information in the attention economy. *Scientific Reports*, 5:9452, 2015.

[52] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

[53] J Lorenz, H Rauhut, F Schweitzer, and D Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011.

[54] Joshua Becker, Devon Brackbill, and Damon Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 114(26):E5070–E5076, 2017.

[55] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

[56] Thomas T. Hills. The dark side of information proliferation. *Perspectives on Psychological Science*, 14(3):323–330, 2019.

[57] Alexander J. Stewart, Mohsen Mosleh, Marina Diakonova, Antonio A. Arechar, David G. Rand, and Joshua B. Plotkin. Information gerrymandering and undemocratic decisions. *Nature*, 573(7772):117–121, 2019.

[58] Nazanin Alipourfard, Buddhika Nettasinghe, Andrés Abeliuk, Vikram Krishnamurthy, and Kristina Lerman. Friendship paradox biases perceptions in directed networks. *Nature Communications*, 11(1):707, 2020.

[59] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[60] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.*, 105(15):158701, 2010.

[61] Haluk Bingol. Fame emerges as a result of small memory. *Physical Review E*, 77(3):036118, 2008.

[62] Bernardo A Huberman. Social computing and the attention economy. *Journal of Statistical Physics*, 151(1–2):329–339, 2013.

[63] Fang Wu and Bernardo A Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

[64] Nathan O. Hodas and Kristina Lerman. How limited visibility and divided attention constrain social contagion. In *Proc. ASE/IEEE International Conference on Social Computing*, 2012.

[65] Jeon-Hyung Kang and Kristina Lerman. VIP: Incorporating Human Cognitive Biases in a Probabilistic Model of Retweeting. In *Proc. International Conference on Social Computing, Behavioral Modeling and Prediction*, 2015.

[66] Moshe Adler. Stardom and talent. *American Economic Review*, 75(1):208–12, 1985.

[67] 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.

[68] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Sci. Rep.*, 2(335), 2012.

[69] James P Gleeson, Jonathan A Ward, Kevin P O'Sullivan, and William T Lee. Competition-induced criticality in a model of meme popularity. *Physical Review Letters*, 112(4):048701, 2014.

[70] James P. Gleeson, Kevin P. O'Sullivan, Raquel A. Baños, and Yamir Moreno. Effects of network structure, competition and memory time on social spreading phenomena. *Phys. Rev. X*, 6(2):021019, 2016.

[71] Daniele Notarmuzi and Claudio Castellano. Analytical study of quality-biased competition dynamics for memes in social media. *Europhysics Letters*, 122(2):28002, 2018.

[72] Dimitar Nikolov, Alessandro Flammini, and Filippo Menczer. Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *HKS Misinformation Review*, 1(7), 2021.

[73] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms. Preprint 2310.05984, arXiv, 2023.

[74] Nathan O Hodas and Kristina Lerman. The simple rules of social contagion. *Scientific reports*, 4(1):4343, 2014.

[75] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Political polarization on twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[76] Michael D Conover, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Partisan asymmetries in online political activity. *EPJ Data Science*, 1:6, 2012.

[77] L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Sci. Rep.*, 3(2522), 2013.

[78] Azadeh Nematzadeh, Emilio Ferrara, Alessandro Flammini, and Yong-Yeol Ahn. Optimal network modularity for information diffusion. *Physical review letters*, 113(8):088701, 2014.

[79] Serge Galam and Frans Jacobs. The role of inflexible minorities in the breaking of democratic opinion dynamics. *Physica A: Statistical Mechanics and its Applications*, 381:366–376, 2007.

[80] Alex Waagen, Gunjan Verma, Kevin Chan, Ananthram Swami, and Raissa D'Souza. Effect of zealotry in high-dimensional opinion dynamics models. *Phys. Rev. E*, 91:022811, 2015.

[81] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski. Social consensus through the influence of committed minorities. *Phys. Rev. E*, 84:011130, 2011.

[82] Dina Mistry, Qian Zhang, Nicola Perra, and Andrea Baronchelli. Committed activists and the reshaping of status-quo social consensus. *Phys. Rev. E*, 92:042805, 2015.

[83] Matthew O. Jackson, Suraj Malladi, and David McAdams. Learning through the grapevine and the impact of the breadth and depth of social networks. *Proceedings of the National Academy of Sciences*, 119(34):e2205549119, 2022.

[84] David J Butts, Sam A Bollman, and Michael S Murillo. Mathematical modeling of disinformation and effectiveness of mitigation policies. *Scientific Reports*, 13(1):18735, 2023.

[85] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 477–487, 2022.

[86] 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.

[87] L. K. Fazio. Pausing to consider why a headline is true or false can help reduce the sharing of false news. *The Harvard Kennedy School Misinformation Review*, 1(2), 2020.

[88] Marshall W. Van Alstyne. A Response to Fake News as a Response to Citizens United. *Comm. ACM*, 62(8):26–29, 2019.

[89] Jeffrey Mervis. An internet research project draws conservative ire. *Science*, 346(6210):686–687, 2014.

[90] Daryl J Daley and David G Kendall. Epidemics and rumours. *Nature*, 204:1118, 1964.

[91] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 6(10):1372–1380, 2022.

[92] Thayer Alshaabi, David Rushing Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. *EPJ Data Science*, 10:15, 2021.

[93] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. Exposure to social engagement metrics increases vulnerability to misinformation. *HKS Misinformation Review*, 1(5), 2020.

[94] Dimitar Nikolov, Alessandro Flammini, and Filippo Menczer. Replication Data for: Right and left, partisanship predicts vulnerability to misinformation. Harvard Dataverse, 2020. doi:10.7910/DVN/6CZHH5.

[95] Alexei Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E*, 67:056104, 2003.

[96] Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak. In *Proc. ICWSM Intl. Workshop on Cyber Social Threats (CySoc)*, 2020.

[97] C A Davis, G L Ciampaglia, L M Aiello, K Chung, M D Conover, E Ferrara, A Flammini, G C Fox, X Gao, B Gonçalves, P A Grabowicz, K Hong, P Hui, S McCaulay, K McKelvey, M R Meiss, S Patil, C Peli Kankanamalage, V Pentchev, J Qiu, J Ratkiewicz, A Rudnick, B Serrette, P Shiralkar, O Varol, L Weng, T Wu, A J Younge, and F Menczer. OSoMe: The IUNI Observatory on Social Media. *PeerJ Computer Science*, 2:e87, 2016.

Supplementary Material

S1 News feed size and information load

We use a news feed size $\sigma = 15$. This value is approximated from the average depth of approximately 107 mobile scrolling sessions measured on Tumblr during two weeks in 2016. The feed interface of this app was similar to those of other social media platforms. We considered a session to have ended when there was no interaction for 30 minutes or longer. The depth of a session was recorded as the number of times that a user scrolled at least 500 pixels through the feed and then stopped for at least one second.

Results are robust for different values of news feed size σ as well as posting activity μ (Fig. S1).

S2 Relationship between quality and engagement

To estimate the relationship between quality and engagement, we start from a dataset of tweets about COVID-19 (96). The dataset consists of tweets containing the hashtags #coronavirus or #covid19 collected in 2020, selected from a 10% random sample of public tweets (97). From this set of tweets, we analyze only posts that shared links to news sources from March 9–29, 2020. Linked sources were extracted from URLs in the tweet metadata. URLs shortened with 70 popular URL shortening services were also expanded to reveal the actual sources. Posts with links to Twitter and other social media platforms were excluded. The remaining links were matched against sources with ratings obtained from NewsGuard[3] in April 2021. The final dataset contains 110,224 original Twitter posts that share at least one link with a NewsGuard rating.

[3] `newsguardtech.com`

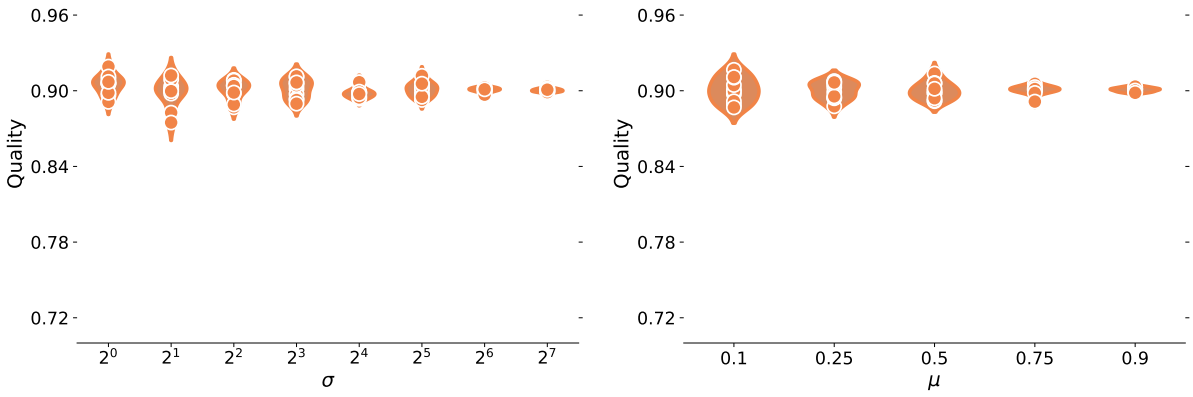

Figure S1: Effects of cognitive features of authentic agents on information quality. Left: News feed size σ. Right: Information load μ. The simulations are run on the original network without bad actors. The results for $\mu = 0.5$ and $\sigma = 15$ are used as a baseline to calculate relative quality in the main text.

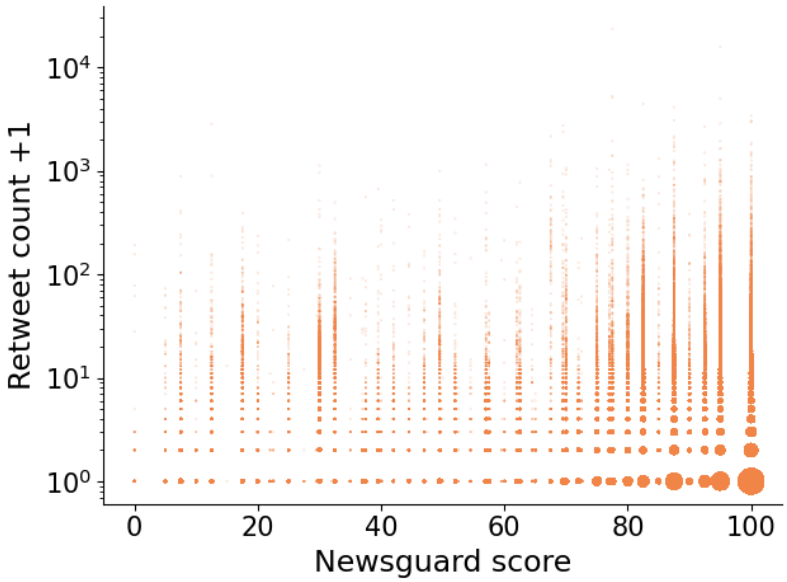

Figure S2: Scatter plot of retweet counts and Newsguard scores of original tweets with links to news articles. The size of each point is proportional to the number of tweets having a given retweet count and Newsguard score. We add one to the retweet count so that the tweets with zero retweets are visible on the log scale. The slightly negative correlation is driven by the majority of tweets linking to high-quality sources and having no retweets.

For each post we have both the rating of the shared source, used as a proxy for post quality, and the number of retweets. The latter is extracted from the metadata of the latest retweet of each post.

Fig. S2 shows that the engagement of a post, as measured by its retweet count, is weakly correlated with its quality (Spearman coefficient -0.13). Based on this observation, we assume that appeal is independent of quality in the model.

S3 Empirical estimation of quality distribution

The same empirical data described above is used to estimate the distribution of authentic message quality. The quality of a post is the average Newsguard score of the news sources shared in the post, normalized to be in the unit interval. As shown in Fig. S3, we fit the data to the exponential probability density function $P(q) = \frac{\tau}{1-e^{-\tau}} e^{-\tau q}$, with $\tau = -10$. The normalization constant is obtained by setting $\int_0^1 P(q)dq = 1$.

S4 Empirical estimation of appeal distribution

The appeal distribution of authentic content is modeled by the probability density function $P(a) = (1 + \alpha)(1 - a)^\alpha$. The normalization constant is obtained by setting $\int_0^1 P(a)da = 1$. This captures the intuition that most messages have low appeal. As illustrated in Fig. S4(a), high appeal values are much more rare for

$\alpha = 10$ than $\alpha = 1$. While there is no empirical proxy data for appeal, a reasonable choice should give rise to a distribution of engagement consistent with empirical reshare data. Fig. S4(b) shows that values $\alpha \geqslant 1$ result in distributions of engagement in the model that roughly capture the broad distribution of reshares in the same empirical data described above.

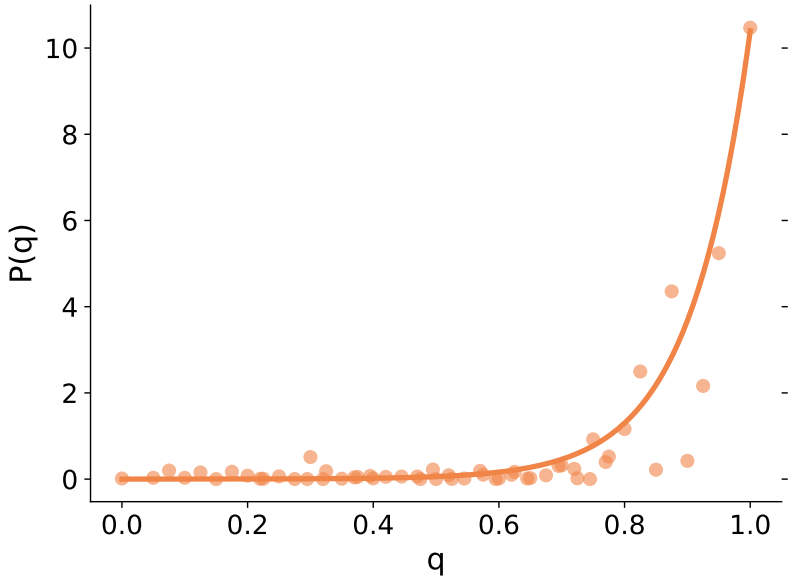

Figure S3: Distribution of quality for authentic agent messages (line) fitted to empirical Twitter data (points).

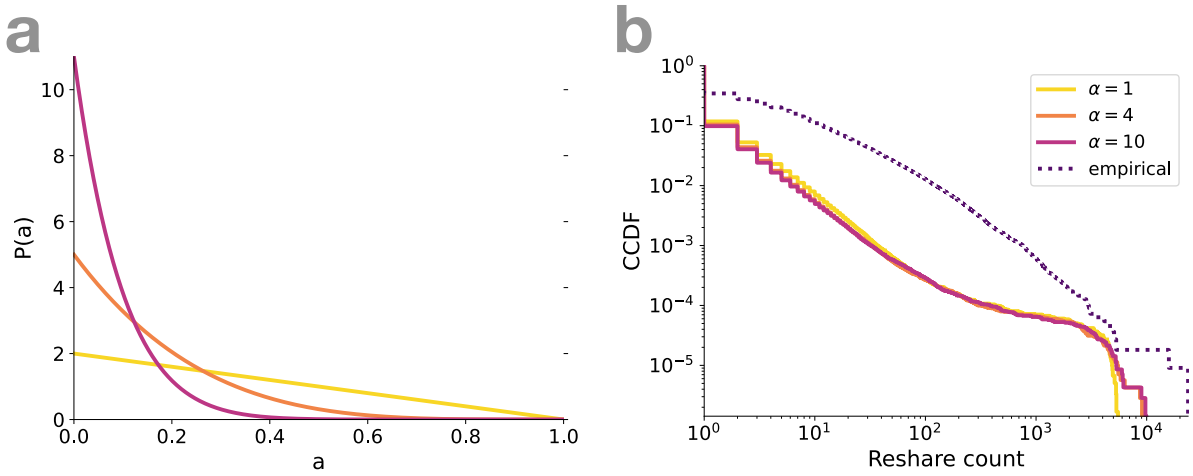

Figure S4: Appeal distribution. (a) Distribution of appeal for authentic agent messages, for different values of the parameter α. (b) Reshare count distribution of empirical posts, and of messages in the baseline simulation with only authentic accounts. The message reshare counts in the model are measured at the steady state.