Information Pollution by Social Bots

Xiaodan Lou Alessandro Flammini
Filippo Menczer
Center for Complex Networks and Systems Research
Indiana University, Bloomington

Abstract

Social media are vulnerable to deceptive social bots, which can impersonate humans to amplify misinformation and manipulate opinions. Little is known about the large-scale consequences of such pollution operations. Here we introduce an agent-based model of information spreading with quality preference and limited individual attention to evaluate the impact of different strategies that bots can exploit to pollute the network and degrade the overall quality of the information ecosystem. We find that penetrating a critical fraction of the network is more important than generating attention-grabbing content and that targeting random users is more damaging than targeting hub nodes. The model is able to reproduce empirical patterns about exposure amplification and virality of low-quality information. We discuss insights provided by our analysis, with a focus on the development of countermeasures to increase the resilience of social media users to manipulation.

1 Introduction

It has become increasingly clear in recent years that social media and their users are vulnerable to manipulation by various means such as astroturf [1, 2], trolling [3], impersonation [4], and misinformation [5, 6, 7, 8]. The vulnerabilities of people and platforms to these kinds of manipulation can be attributed to a complex interplay of socio-cognitive [9, 10], political [8], and algorithmic [11, 12] biases. In addition to interacting and reinforcing each other, these biases are exploited by social media accounts controlled in part by software, commonly called social bots [13].

Some social bots are designed with benign intentions and serve useful purposes [14, 15]. However, social bots also have many harmful applications; by impersonating human users they can deceptively pollute our information ecosystem to influence people and systems. Such inauthentic accounts are used to amplify misinformation [16], create the appearance that people or opinions are more popular than they are [1], influence public opinion [17, 18, 2, 19], commit financial fraud [20], infiltrate vulnerable communities [21, 3, 22], or disrupt communication [23, 24]. This paper focuses on such malicious applications of social bots.

There is a growing research literature on machine learning algorithms to detect social bots [25, 26, 27, 13, 28, 29, 30, 31, 32]. This is a difficult adversarial problem, as more sophisticated bots are designed to counter detection tools [33]. Some social bots behave in ways that cannot be distinguished by those of humans on an individual basis because automation is mixed with human behavior, or because their behaviors are driven by (or copied from) humans; automation is used mainly to replicate a behavior across many inauthentic accounts. Supervised learning cannot recognize these so-called *cyborgs* [34]. Unsupervised learning can be used to cluster accounts based on coordinated patters in timing, content, or network features [35, 36, 37]. But these approaches have limited scalability and therefore a significant amount of social media manipulation remains undetected. Despite steps reportedly taken by social media platforms to curb malicious accounts [38, 39], we expect social bots to continue to pollute our information ecosystem.

We can measure the effectiveness of social bots in certain pollution operations and in specific domains, such as the spread of low-credibility information [16]. Yet, due to both ethical and data availability limitations, it is difficult to carry out empirical experiments and analyses in the real world to measure the large-scale consequences of such pollution operations on human behavior. There are conflicting accounts, for example, about whether disinformation campaigns on social media may have swayed elections [40, 41, 42, 43, 44, 45]. To the extent that humans are influenced by the information to which they are exposed, it is reasonable to theorize that their behavior can be manipulated or controlled if a malicious coordinated attack is able to control the flow of information on an online social network. Nothing is known, however, about the power of social bots to disrupt information ecosystem on such a global scale.

In this paper we introduce a *social media model* to explore scenarios in which social bots manipulate an information sharing network and to evaluate the impact of different bot strategies to pollute and degrade the overall quality of the information ecosystem. Each piece of information that can be shared and reshared, such as an image, link, hashtag, or phrase, is referred to as a *meme* [46]. The underlying agent-based model accounts for limited attention by agents, which can give rise to heavy-tailed distributions of meme popularity and lifetime consistent with empirical data [47]. We assign a *quality* value to each meme and assume that agents select and reshare memes shared by their friends with probability proportional to their *perceived* quality. This approach allows us to evaluate the impact of different strategies that bots can exploit to pollute the network and degrade the overall quality of the information ecosystem.

The information-spreading model with limited attention and quality preference was introduced in a prior paper, where we found that the correlation between meme quality and popularity degrades with increasing information load and/or decreasing individual attention [48]. In that paper we also explored whether the model was able to explain the empirical finding that low-quality information is as likely to spread in a viral fashion as high-quality information.

Our initial analysis suggested this to be the case. However, we later discovered an error in the analysis; the model was in fact unable to generate equal distributions of popularity for low- and high-quality memes, therefore we retracted the paper [49]. In the present paper, we revisit this question and find that the presence of manipulation by social bots is sufficient to explain the empirical virality patterns of low-quality information.

The social media model presented here suggests that bots can be used to pollute the information ecosystem, amplifying exposure to low-quality information as observed in empirical data [16]. Our analysis sheds light on effective manipulation strategies. In particular, we show that infiltrating the network is more important than generating attention-grabbing (e.g., novel or compelling) content. Contrary to intuition, we also find that targeting hubs is a less destructive strategy than targeting random users. These insights can help develop countermeasures to increase the resilience of social media and their users to manipulation. We discuss mitigation steps that could be taken by social media platforms and the legal issues that arise from government regulations aimed at protecting human speech from suppression by bots.

2 Methods

We aim to examine the conditions in which bots can displace quality information. To this end, we propose a simple 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 [50, 51, 52, 53]. We model a social media platform for information sharing, such as Twitter or Instagram, as a directed network. Nodes represent agents (users) and links represent follower relations, which may or may not be reciprocal. The direction of the link goes from the follower to the followed (friend) account, capturing attention; when a friend's post is reshared by a follower, information spreads in the opposite direction.

2.1 Network Model

To model the presence of both humans and social bots, we assume that the system consists of two subnetworks, with a parameter describing the ratio between their sizes. Let the human subnetwork be composed of N nodes and the bot subnetwork of βN nodes.

Both human and bot subnetworks have a structure that mimics online social networks. To capture the characteristic presence of hubs and clustering (directed triads), we construct the two subnetworks using a directed variant of the random-walk growth model [54]. Each network is initialized with four fully connected nodes. Then we add one new node at a time, and assume that each has fixed out-degree k_{out}. Once a new node i comes into the network, it links to a randomly selected target node j. Each of the remaining $k_{out} - 1$ friends are selected as follows: with probability expressed by a parameter p, i follows a

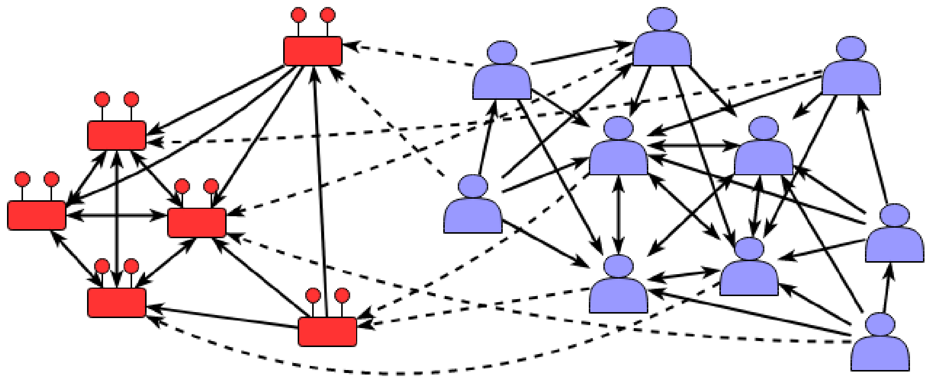

Figure 1: Schematic illustration of the network structure. Human (blue) and bot (red) nodes are depicted by different icons. Solid links indicate follower relations within each subnetwork. Dashed links represent humans following bots, according to the infiltration parameter γ.

random friend of j's; with probability $1-p$, i follows another randomly selected node. Following friends of a friend is akin to copying links; this has the effect of generating closed directed triads and approximates a preferential attachment process, giving rise to hub nodes with high in-degree.

In addition, the bot subnetwork is designed to manipulate information flow and collective attention by amplifying the spread of certain messages and the influence of certain accounts. Therefore we assume that bots get human accounts to follow them: we add a directed link from each human node to each bot node with probability γ. The parameter γ models the *infiltration* of the human network by bots: when $\gamma = 0$ there is no infiltration and bots are isolated, therefore harmless; the opposite extreme $\gamma = 1$ indicates complete penetration such that bots dominate the network. Fig. 1 illustrates the network structure.

In the scheme described above, irrespective of the infiltration parameter γ, the humans that follow the bots are selected by *random wiring*. To study whether bots can manipulate the network more efficiently by targeting influential human accounts, we also explore a *preferential wiring* strategy. Each bot is still followed by γN humans on average, but these human targets are selected with probability proportional to their in-degree k_{in}, i.e., their number of followers.

2.2 Social Media Model

Fig. 2 illustrates the dynamics of the social media model. In contrast with classical epidemiological models, messages carrying new memes are continuously introduced into the system in an exogenous fashion. At each time step one agent i is chosen at random. With probability μ, i produces a message carrying a new meme. Alternatively, with probability $1-\mu$, i selects one of the messages in

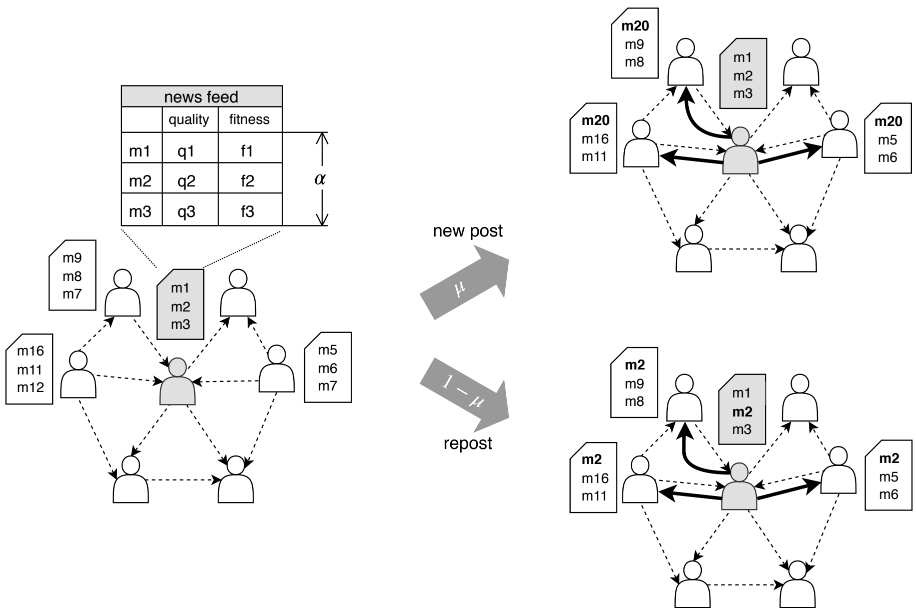

Figure 2: Illustration of information diffusion in the social media model. Each node has a news feed of limited size α, containing messages posted or reposted by friends. Messages propagate along follower links (dashed arrows). At each time step, an agent is considered (gray node). With probability μ, the node posts a new message (here, `m20`). Otherwise, with probability $1 - \mu$, the agent scans its feed and reposts one of the messages according to their fitness (here, `m2`). The message propagate to the node's followers along the solid arrows and shows up at the top of their feeds.

its news feed to be reposted. The parameter μ measures the average number of memes received by an agent per unit time, and represents the *information load* of the agents in the network.

The new or reposted message is added to the news feeds of i's followers. Each agent's feed contains the α most recent messages posted or reposted by their friends; if a feed exceeds α messages, the oldest is forgotten. Although social media platforms do not rank posts in strict reverse chronological order, this is a realistic simplifying assumption because all platforms give high priority to recent posts. The parameter α models the number of memes viewed in a news feed during a session, and represents the *finite attention* of the agents.

To measure how bot activity can manipulate information diffusion, we assume that each meme is characterized by two properties: *quality* and *fitness*. Quality may represent desirable properties that make the meme more likely to be shared, depending on the situation being modeled: the originality of an idea, the beauty of a picture, and the truthfulness of a claim are valid examples. Fitness, on the other hand, represents quality as perceived by human agents; it is the likelihood that a meme is actually shared. In an ideal world, fitness would equal quality. In the real world, however, deceptive messages may have low quality and high fitness. For example, false news and junk science articles have low quality — we would not share them if we knew their true nature. Yet they may go viral because of deceptive content that is novel, click-bait, ripped from headlines, and/or that appeals to people's political, emotional, or conspiratorial bias. In fact, these deceptive features may make low-quality content even more likely to spread virally than high-quality content [6].

We model the dichotomy between meme quality q and fitness f by assuming that memes originating from humans have fitness reflecting their quality ($q = f$) whereas bot memes have low quality ($q = 0$) but deceptively high fitness. Both q and f are defined in the unit interval. The potential difference in fitness due to the deceptive nature of bot-amplified content is modeled by drawing the fitness of human and bot memes from the two respective distributions:

$$P_h(f) = \frac{(1-f)^{\phi}}{\int_0^1 (1-f)^{\phi} \mathrm{d}f} = (1+\phi)(1-f)^{\phi} \quad (1)$$

$$P_b(f) = \frac{(1-f)^{1/\phi}}{\int_0^1 (1-f)^{1/\phi} \mathrm{d}f} = (1+1/\phi)(1-f)^{1/\phi} \quad (2)$$

where $\phi \geq 1$ is a *deception* parameter that captures the differential in fitness. As shown in Fig. 3, if $\phi = 1$, bots and humans generate memes with fitness drawn from the same distribution $P(f) = 2(1-f)$. If $\phi > 1$, bot memes are more likely to have high fitness; the larger ϕ, the greater the potential virality of low-quality content amplified by bots.

Humans pay attention to memes shared by their friends. An agent does not know the real quality of the memes in its feed; we assume that the probability that it shares one of these memes, allowing it to spread, is proportional to the meme's fitness. More explicitly, let M_i be the feed of i ($|M_i| = \alpha$). The

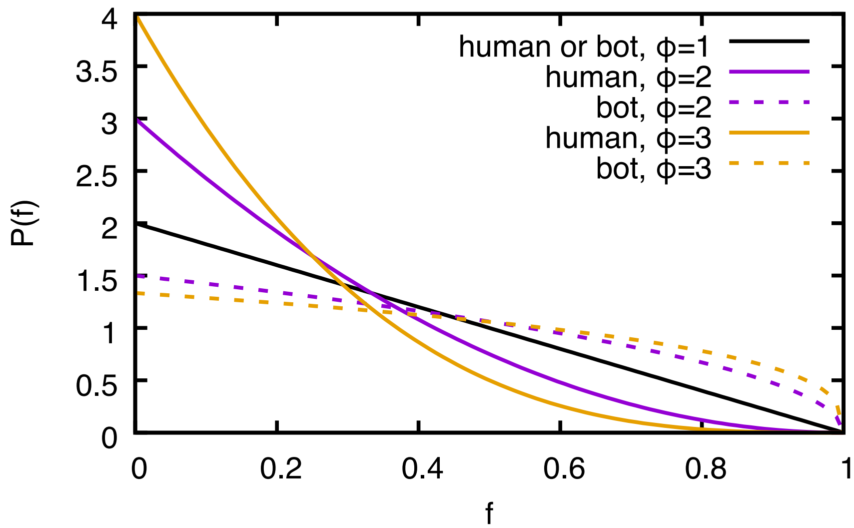

Figure 3: Distributions of fitness for human and bot memes, regulated by the deception parameter ϕ.

probability of message $m \in M_i$ being selected is $P(m) = f(m)/\sum_{j \in M_i} f(j)$ where $f(m)$ is the fitness of the meme carried by m.

2.3 Descriptive Metrics

We wish to quantify three desirable properties of the information ecosystem: the average quality and diversity of the memes in the network, and the system's capacity to discriminate information on the basis of its quality. Next we define three metrics that capture these properties. Our subsequent analyses aim to evaluate information ecosystem pollution by measuring the changes in these metrics in response to bot strategies.

Average quality measures how the injection of low-quality memes by bots affects the overall quality of the information ecosystem. We calculate it by averaging quality across all the memes visible through the feeds of all the human agents:

$$Q = \frac{1}{\alpha N} \sum_{i=1}^{N} \sum_{m \in M_i} q_{im},$$

where q_{im} is the quality of the meme in the mth message of user i's feed.

Information diversity is another desirable property of an ideal communication system, because without diversity of ideas there can be no room for informed discussions and decisions. Due to people's limited attention, the injection and amplification of low-quality memes by bots could suppress the diversity of quality information in the system. We use entropy to measure the amount of diversity in the human network:

$$D = \sum_m P(m) \log P(m),$$

where $P(m)$ is the fraction of messages containing meme m across all the feeds of users in the human network. The minimum diversity is zero, which means there is a single meme dominating the entire ecosystem.

There is a trade-off between average quality and diversity in the network, because the maximum quality may be obtained by having the best meme take over the system, which would kill the diversity. Ideally, in the presence of both high quality and diversity, one would want the system to discriminate against low-quality memes by reaching a strong correlation between quality and popularity: the higher the quality of a meme, the more widely it should be shared among human users. We capture this *discriminative power* by measuring Kendall's rank correlation between popularity and quality of human-generated memes. The popularity of a meme is measured by counting its occurrences in the feeds of human agents, i.e., the number of times it is shared or reshared. Kendall's correlation coefficient is computed by ranking memes according to the two criteria and then counting the number of meme pairs for which the two rankings are concordant or discordant, properly accounting for ties [55]:

$$\tau = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_t^q)(n_c + n_d + n_t^p)}}$$

where n_c is the number of concordant pairs, n_d the number of discordant pairs, n_t^q the number of ties only in quality, and n_t^p the number of ties only in popularity. High τ indicates that high-quality memes are more likely to go viral, granting the system discriminative power; in the extreme case $\tau = 1$ the two rankings are completely concordant. Small τ signifies a lack of quality discrimination by the network.

2.4 Simulation Framework

The parameters μ and α that regulate information load and limited attention have been explored in previous work [48] and shown to decrease the correlation between quality and popularity of memes (Fig. 4). We use $\alpha = 15$, which is the approximate average scrolling session depth measured on a social media mobile app; and we draw μ from an empirical distribution of tweet (vs. retweet) rates on Twitter. The distribution has average $\langle\mu\rangle \approx 0.75$ and is peaked at $\mu = 1$. We set the value of $\beta = 0.1$ based on the empirical finding that 9–15% of Twitter accounts use automation [30]. For the remaining parameters, we use $N = 10^4$ nodes, $k_{out} = 3$, and $p = 0.5$. Unless otherwise stated, the human followers of the bots are selected at random.

We wish to study the characteristics of bot behaviors that pollute social media platforms, as modeled by network infiltration and content deception. Therefore we explore how desirable properties of the information ecosystem deteriorate in response to different values of the γ and ϕ parameters.

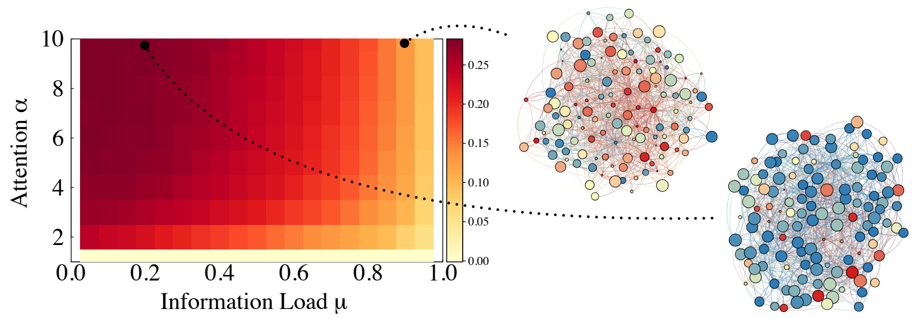

Figure 4: Map of discriminative power τ as a function of information load μ and finite attention α. This data is based on a simplified version of the model with no bots ($\beta = 0$), $N = 1000$ agents in an undirected network constructed by preferential attachment [56], $\langle k \rangle = 20$, and constant μ [48]. The two small networks ($N = 128$) illustrate how the discriminative power decreases with larger μ ($\alpha = 10$, $\mu = 0.2, 0.9$). Node color and size represent the last shared meme and its quality, respectively.

For each set of parameters, we analyze the behavior of the social media model by simulating the information diffusion process on a directed network generated through the above method. As the competition takes place, some of the memes die fast while others live longer and infect a large fraction of the network. Such a process continues until the systems reaches a *steady state* in which the average number of distinct memes remains roughly constant (this number depends on μ). In our analysis, we only consider the memes that are introduced after the system reaches the steady state. Then we monitor 10^6 memes from their birth until they completely disappear from the human system. We record each meme's quality and popularity during this steady-state stage.

Each simulation is run 20 times, starting from random conditions, for a total of 2×10^7 memes that are included in our analysis. The measurements of average quality, diversity, and discriminative power are averaged across simulation runs.

3 Results

The social media model allows us to explore the effects of different bot pollution strategies: infiltration of the network, deceptive content, and targeting of influential human users. In this section we present the results of these analyses and then compare the predictions of the model with some empirical observations.

3.1 Effects of Pollution

To quantify the effects of bot strategies on pollution, we measured average quality, diversity, and discriminative power after running simulations of our information ecosystem model for different values of the infiltration parameter ($10^{-3} \leq \gamma \leq 1$) and the deception parameter ($1 \leq \phi \leq 10$). The results are shown in Fig.5.

Not surprisingly, the system is most robust to pollution when both the infiltration of bots into the human social network (γ) and their deception (ϕ) are minimal. Both bot infiltration and deception have a clear negative impact on the average quality of information in the system; the higher γ and ϕ, the lower the average information quality. We also note an abrupt, non-linear transition toward low quality when humans follow 1% or more of the bots. For γ above this level, quality is suppressed so much that bot deception makes little difference.

One might expect highly deceptive bot memes to displace human memes and therefore suppress information diversity. However, we observe that the diversity does not depend in a significant way on the deception level ϕ. Bot infiltration does have a strong negative effect on diversity, but only when γ is sufficiently high — again we note a sharp transition around 1%.

Similar to diversity, the discriminative power of the human network degrades mainly as a result of the increase in bot infiltration γ. When the deception ϕ is large, the capacity to distinguish good memes decreases gradually even as bots infiltrate the network very marginally. When ϕ is small, on the other hand, the discriminative power remains high for small γ and then drops abruptly as bot infiltration passes the critical threshold $\gamma \approx 1\%$.

3.2 Targeting Influential Users

The above results show that bots can destroy the health of the information ecosystem if they are able to penetrate a sufficient fraction of the human network. This infiltration assumes that the human followers of the bots are selected randomly with equal probability. But those interested in manipulating the network through bots may be able to target influential nodes. This is well within the capability of automated accounts; for example, the number of followers is often used as a proxy for influence [57] and is public information for most Twitter and Instagram accounts, as well as Facebook pages. A bot can easily interact with accounts having many followers by mentioning and/or following them as well as retweeting, quoting, and/or liking their tweets.

There is empirical evidence of preferential targeting in bots that spread misinformation [1, 16]. Presumably this strategy is based on the assumption that it can multiply the impact a bot, because a retweet by an influential human follower can reach many nodes. An important question, then, is whether targeting influential human accounts by bots can in fact increase the harm of pollution.

Let us explore this question through a variation of our model in which links from humans to bots are created in such a way that people with more followers have a higher probability of following bots — the probability of a human node being selected is proportional to the node's in-degree k_{in}. The model is otherwise identical to the one discussed so far. To compare the pollution effects of the two targeting strategies, we plot in Fig. 6 the ratio $Q_{\text{pref}}/Q_{\text{rand}}$ between the average quality of the preferential and random wiring models.

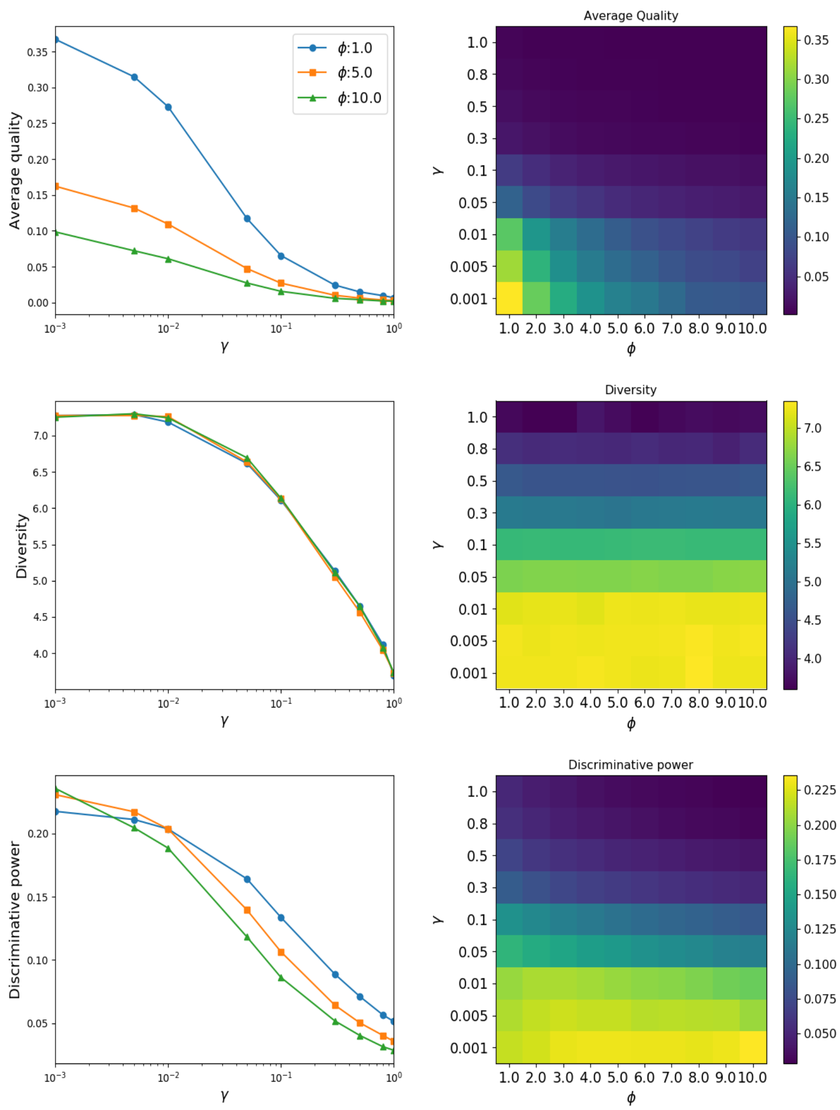

Figure 5: Pollution effects measured by average meme quality Q (top), meme diversity D (middle), and discriminative power τ (bottom) as a function of infiltration probability γ ($10^{-3} \leq \gamma \leq 1$) and deception level ϕ. The charts on the left focus on three values of ϕ, while the density plots on the right map the range $1 \leq \phi \leq 10$ and use colors to represent the metric values. Only memes created by humans and their spread in the human system are considered.

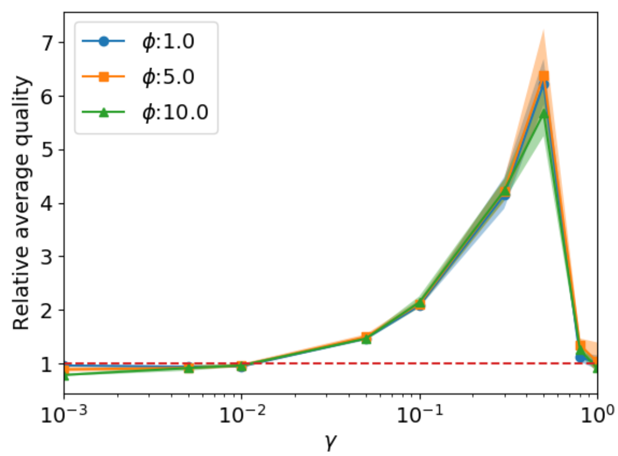

Figure 6: Ratio of average quality in preferential targeting vs. random wiring strategies, $Q_{\text{pref}}/Q_{\text{rand}}$, for different values of infiltration γ and deception ϕ. The dashed line is the baseline where the two strategies produce equal harm. Colored bands around the curves represent 95% confidence intervals based on 20 simulations.

For extremely low bot infiltration, preferential targeting is slightly more destructive than random wiring, but the wiring strategy has little influence on the pollution outcome when $\gamma \leq 1\%$. When the infiltration is sufficiently high ($\gamma > 1\%$), surprisingly, we observe the opposite effect $Q_{\text{pref}} > Q_{\text{rand}}$: the preferential targeting model is significantly less effective at suppressing the quality of the information ecosystem compared to random wiring. At the extreme $\gamma = 1$, naturally, the wiring strategy is irrelevant because all humans follow bots. Different deception levels do not affect these results.

These results mean that a (destructive) advantage of targeting hub nodes can only be seen if very few humans follow the bots. Otherwise, targeting hubs actually mitigates the harm to the information ecosystem. Further inspection to interpret this counterintuitive finding reveals that targeting hubs leads to a greater concentration of low-quality content in the network, which is reflected in a higher average quality. For an illustration, consider the networks visualized in Fig. 7. When the infiltration is low ($\gamma = 1\%$), preferential targeting pollutes a few hubs. For high infiltration ($\gamma = 10\%$), random wiring results in a spread of the low-quality information across the network. Preferential targeting instead concentrates the low-quality memes among the hubs, leaving many low-degree nodes free from pollution. This concentration makes the targeting of influential accounts less deleterious to the ecosystem.

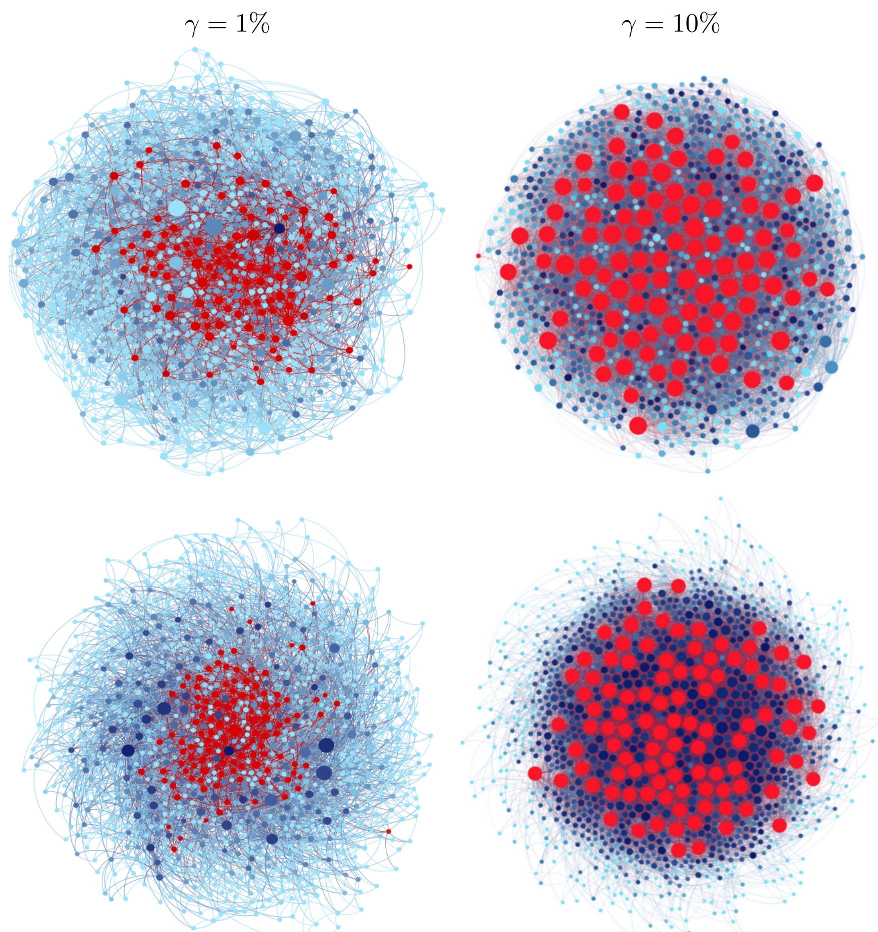

Figure 7: Visualizations of network pollution by bots under random (top) and preferential (bottom) wiring strategies and different infiltration levels γ, with deception $\phi = 5$. Red and blue nodes represent bots and humans, respectively. The darker a blue node, the more bad ($q = 0$) memes in their feed. Node size represents the number of followers.

3.3 Empirical Virality and Amplification

Our model is an extremely simplified representation of the real, complex social media platforms we use everyday. A key question, then, is whether the stylized facts presented above provide a realistic picture of how social bot behaviors can pollute our actual information ecosystems. While it is hard to imagine ethical experiments that could validate our model predictions, we can compare some of our findings with existing empirical data to see whether our results are consistent with patterns observed in the real world.

One set of empirical observations concerns the comparison between the virality of high- and low-quality information. Our previous analyses [48, 16] showed that the popularity distributions of true and false news on Facebook, and of articles from low-quality and fact-checking sources on Twitter, are practically indistinguishable, as shown in Fig. 8(a,b). Without bots, the model is unable to generate equal distributions of popularity for low- and high-quality memes [49]. In the presence of low-infiltration and low-deception bots, however, the model can reproduce equal distributions (Fig. 8(c)). Using higher values of the bot infiltration and deception parameters γ and ϕ, the model predicts that low-quality information can be even more likely to go viral than high-quality information, as illustrated in Fig. 8(d–h). These results are consistent with empirical analyses comparing the virality of accurate and debunked news on Twitter [6].

For a second empirical observation, let us consider how human exposure to low-quality information can be amplified by the sharing activity of social bots, based on our prior analysis of the spread of low-credibility articles on Twitter [16]. Fig. 9(a) plots the estimated numbers of tweets linking to articles from low-credibility sources by likely humans vs. bots, separated using a threshold on bot scores calculated by the Botometer tool [30, 33]. Irrespective of the threshold, the volume of human tweets per article V_h grows faster than the volume of bot tweets for article V_b; fitting a power law $V_h = V_b^{\eta}$, a super-linear relationship is indicated by the exponent $\eta > 1$: bots amplify the reach of articles from low-credibility sources. No amplification effect was found for fact-checking articles ($\eta = 1$). Fig. 9(b) shows that a similar super-linear relationship is predicted by our model for low-quality memes. To estimate the exponent, we plot $\eta = \frac{\log V_h}{\log V_b}$ in Fig. 9(c). We observe $\eta > 1$, consistent with the empirical results: bots amplify the spread of low-quality information.

4 Discussion

Simulations of our social media model in the presence of deceptive social bots reveal that both infiltrating the network and generating attention-grabbing content are effective bot strategies, but infiltration is the dominant factor in suppressing quality. When bots can induce as little as 1% of the humans in the network to follow them, there is a drastic transition toward suppression of quality. Surprisingly, unless infiltration is below this critical value, bots do not even need to target influential users; they can do more damage by connecting to ran-

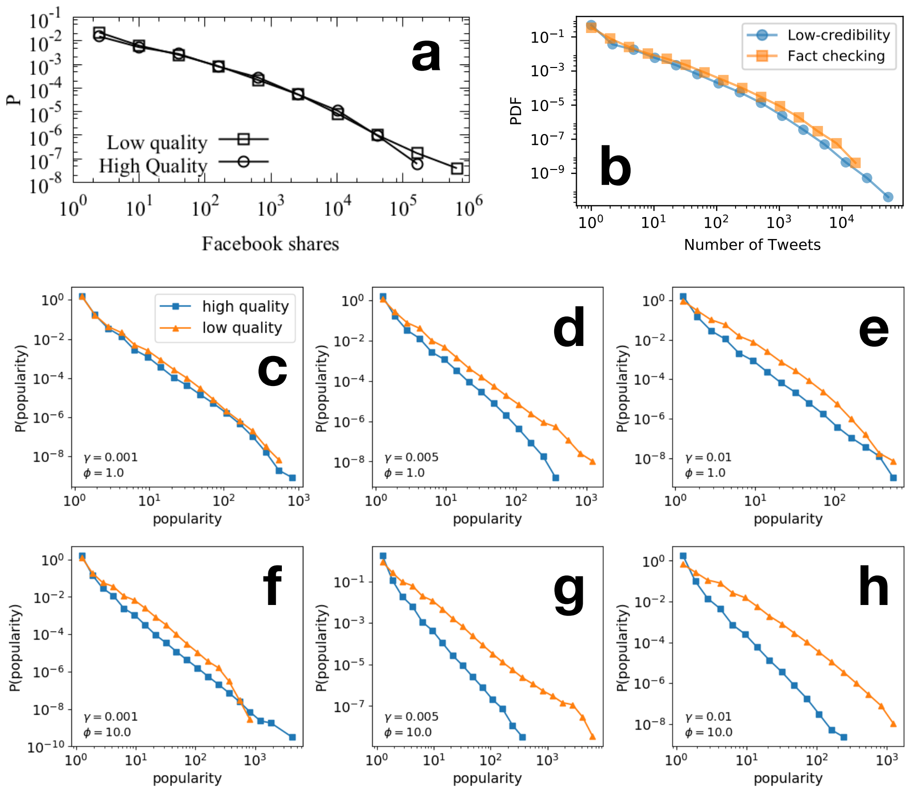

Figure 8: Probability distributions of meme popularity. (a) Facebook shares of news articles that support false claims or undermine true claims (low quality) and articles that fact-check false claim or support true claims (high quality); data from Emergent [48]. (b) Tweets of articles from low-credibility and fact-checking sources (the numbers of article-sharing accounts have very similar distributions); data from Hoaxy [16]. (c–h) Posts of low-quality ($q = 0$) and high-quality ($q > 0$) memes in model simulations with different γ and ϕ values.

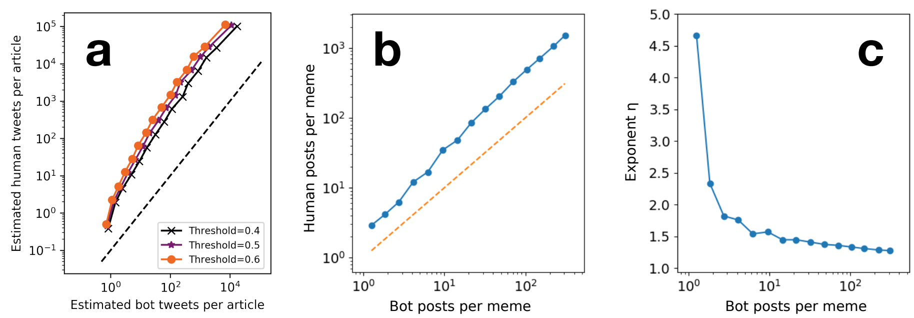

Figure 9: Relationship between human and bot sharing of low-quality information. The dashed lines are guides to the eye, showing linear growth; a superlinear relationship is a signature of amplification by bots. (a) Volume of tweets with links to articles from low-credibility sources by accounts with bot score above a threshold versus the volume of tweets by accounts with bot score below the threshold [16]. (b) Number of low-quality ($q = 0$) memes posted by bots versus humans in simulations of our model with $\gamma = 0.5$ and $\phi = 1$. (c) Exponent of the power-law relationship estimated by the model numbers in panel (b).

dom accounts. The model's predictions are consistent with empirical findings about the relative virality of low- and high-quality information on Twitter and Facebook, and about amplification of low-quality information spread by social bots on Twitter.

The results are robust to changes in the network size ($10^3 \leq N \leq 10^4$) and structure. We experimented with different values of the link copy probability parameter p, yielding networks with different clustering coefficients. We also rule out the possibility that the results are affected by a bias in the directed version of the random-walk model used to build the subnetworks. Early nodes are more likely to become hubs with high in-degree, while later nodes tend to follow older nodes without being followed by them (low or zero in-degree). We repeated our simulations on networks generated with an additional rewiring mechanism that eliminates these "dead-end" nodes. If a node i has $k_{in} = 0$, we randomly choose a node j with $k_{in} > 2$ and rewire one of its outgoing links toward i. In this way, the total in-degree and out-degree of the system remain the same, but every node has at least one follower and is thus able to propagate memes. Finally, we considered networks without clustering, generated with a directed version of the preferential attachment model [56]. All of these network structure variations yield results similar to those reported above and support the same conclusions.

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 memes, including exogenous and endogenous bursts of attention [58, 59], memory [60], novelty [61, 62], and position bias [63, 64]. This literature considers the popularity of pieces of information in isolation. Markets in which *many* memes compete for limited attention have received scarce consideration. Exceptions have considered the cost of learning about quality [65], distortions of quality assessments that result from aggregate knowledge of peer choices [66], and confirmation bias in the spread of misinformation [67]. The social media model extends the model of Weng *et. al.* [47], who demonstrated that some memes inevitably achieve viral popularity irrespective of quality in the presence of competition among networked agents with limited attention. Gleeson *et. al.* [68, 69] formalized this model as a critical branching process, predicting that the popularity of memes follows a power-law distribution with very heavy tails.

We can think of bots as zealots, a minority of agents committed to a particular view. Opinion dynamics models, such as the majority rule, voter, and naming game models, have shown that even a small minority of zealots can drive a system to consensus toward their opinion [70]. This also holds in high-dimensional opinion spaces [71]. When the committed fraction grows beyond a critical value, there is a dramatic decrease in the time taken for the entire population to adopt the committed opinion [72]. Furthermore, when zealots are more active in pushing their message, their efforts require an order of magnitude fewer individuals to bring the population over to a new consensus, compared to a randomly selected committed minority [73]. Although in the social media model we only explore the bots' capacity to pollute the network rather than to drive consensus to a particular meme/opinion, these results suggest future work in the direction of evaluating whether bots can sway social media users toward a particular narrative.

The insights gained from the present findings suggest several countermeasures to increase the resilience of social media users to manipulation and protect information ecosystems from pollution. The first lesson is that we must make it more difficult for social bots to infiltrate the network. Social media platforms are already stepping up their efforts in detecting and taking down abusive and deceptive accounts [38, 39], but could further strengthen these efforts. This could be accomplished by adding friction to following/friending actions, for example warning users who attempt to link to accounts that lack a strong record of human-like activity. Users should also be warned when friend accounts are suspended or take suspicious actions, such as changing names/handles and/or deleting their prior posts *en masse*. Expanding account verification programs and bolstering bot detection research will also help.

Our findings suggest that all social media users, and not only the most influential, may be targeted by malicious bots. Literacy programs should therefore be supported and expanded. Social media platforms could lead by educating users about their vulnerabilities to manipulation by inauthentic accounts and pages. Such literacy efforts could also cover deception — how to spot and ignore content from low-credibility sources.

In our model, social bots exploit and amplify vulnerabilities created by information overload and finite attention: bots can easily flood the system, crowding out quality information in our feed. One way to counter this manipulation is to reduce information load by adding friction to automated posts; accounts posting at very high rates should be challenged to prove that they are human. By aggressively curbing this kind of abuse, social media platforms could improve the overall quality of information to which we are exposed.

Ironically, efforts by social media platforms to suspend abusive and deceptive bots have been criticized by some as censorship [74, 75]; even research on social bot detection has been assailed with charges of free speech suppression [76]. Following a recent California law requiring disclosure when bots are used to promote, say, political candidates [77], questions are being raised as to whether bots should enjoy First Amendment rights [78] or journalistic protections [79].

The fundamental argument of critics of bot countermeasures — whether by platforms or governments — is 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 ideas [80]. Unfortunately, several aspects of modern social media challenge the assumptions of a free marketplace in which participants can access and adopt better ideas [81]. Platforms like Facebook and Twitter have given rise to the so-called *attention economy* [82, 83, 84] in which information production is affected by information consumption [85]. Furthermore, while the *wisdom of the crowd* enabled by social media [86] should facilitate the discrimination of information based on quality by combining the opinions of many individuals [87], we cannot assume that the opinions to which we are exposed online are independent. This leads to higher confidence and lower accuracy [88]. Similarly, strategies such as accepting new information if it comes from multiple sources [89] are vulnerable because a single entity can use bots to create the appearance of popularity.

The findings presented here suggest that social bots can be used to suppress human speech. Therefore, in the United States, a reading of the First Amendment that grants bots (or any entity that controls them) unlimited free-speech rights would seem to lead to a logical contradiction [90]. These questions are yet to be addressed by the courts, and may have huge repercussions on regulations designed to protect the information ecosystem from manipulation [81].

Acknowledgments

We are grateful Marshall Van Alstyne for useful discussion; to Chengcheng Shao and Giovanni Ciampaglia for the Hoaxy data collection; to Alireza Sahami Shirazi for data about scrolling session depth on a social media mobile app; to Onur Varol, Emilio Ferrara, and Clayton Davis for their inspiring work on Botometer; and to Craig Silverman for sharing data from Emergent.info. This work was carried out while XL visited the Center for Complex Networks and Systems Research at Indiana University, with support by the China Scholarship Council. AF and FM were funded in part by DARPA contract W911NF-17-C-0094. 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] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[7] 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.

[8] 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.

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

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

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

[12] 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.

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

[14] Tetyana Lokot and Nicholas Diakopoulos. News bots: Automating news and information dissemination on twitter. *Digital Journalism*, 4(6):682–699, 2016.

[15] 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):1–12, 9 2017.

[16] 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.

[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. Preprint 1905.12687, arXiv, 2019.

[23] Norah Abokhodair, Daisy Yoo, and David W McDonald. Dissecting a social botnet: Growth, content and influence in twitter. In *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] Kyumin Lee, Brian David Eoff, and James Caverlee. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *Proc. AAAI Intl. Conf. on Web and Social Media (ICWSM)*, 2011.

[26] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. Key challenges in defending against malicious socialbots. In *Proc. 5th USENIX Workshop on Large-Scale Exploits and Emergent Threats*, 2012.

[27] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: efficient detection of fake twitter followers. *Decision Support Systems*, 80:56–71, 2015.

[28] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 273–274, 2016.

[29] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, Filippo Menczer, Andrew Stevens, Alexander Dekhtyar, Shuyang Gao, Tad Hogg, Farshad Kooti, Yan Liu, Onur Varol, Prashant Shiralkar, Vinod Vydiswaran, Qiaozhu Mei, and Tim Hwang. The DARPA Twitter Bot Challenge. *IEEE Computer*, 49(6):38–46, 2016.

[30] Onur Varol, Emilio Ferrara, Clayton Allen Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proc. Intl. AAAI Conf. on Web and Social Media (ICWSM)*, pages 280—289, 2017.

[31] Sneha Kudugunta and Emilio Ferrara. Deep neural networks for bot detection. *Information Sciences*, 467(October):312–322, 2018.

[32] Stefano Cresci, Marinella Petrocchi, Angelo Spognardi, and Stefano Tognazzi. On the capability of evolved spambots to evade detection via genetic engineering. *Online Social Networks and Media*, 9:1–16, 2019.

[33] 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.

[34] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 21–30, 2010.

[35] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. DeBot: Twitter Bot Detection via Warped Correlation. In *Proc. Intl. Conf. on Data Mining (ICDM)*, pages 817–822, 2016.

[36] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. DNA-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 31(5):58–64, 2016.

[37] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 963–972, 2017.

[38] Craig Timberg and Elizabeth Dwoskin. Twitter is sweeping out fake accounts like never before, putting user growth at risk. The Washington Post, 2018.

[39] Rob Lever. Fake Facebook accounts: the never-ending battle against bots. AFP/Phys.org, 2019.

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

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

[42] 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.

[43] 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.

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

[45] 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.

[46] R. Dawkins. *The selfish gene*. Oxford University Press, 1989.

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

[48] Xiaoyan Qiu, Diego F. M. Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 1:0132, 2017.

[49] Xiaoyan Qiu, Diego F. M. Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. Retraction note: Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 3:102, 2019.

[50] William Goffman and V A Newill. Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204(4955):225–228, 1964.

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

[52] Stephen Morris. Contagion. *The Review of Economic Studies*, 67(1):57–78, 2000.

[53] Michaela Goetz, Jure Leskovec, Mary McGlohon, and Christos Faloutsos. Modeling blog dynamics. In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.

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

[55] Maurice G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

[56] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97, 2002.

[57] 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.

[58] 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.

[59] 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.

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

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

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

[63] 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.

[64] 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.

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

[66] 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.

[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] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] Cecilia Kang and Sheera Frenkel. Republicans Accuse Twitter of Bias Against Conservatives. The New York Times, 2018.

[75] Dan Tynan. Facebook accused of censorship after hundreds of US political pages purged. The Guardian, 2018.

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

[77] Noam Cohen. Will California's New Bot Law Strengthen Democracy? The New Yorker, 2019.

[78] Laurent Sacharoff. Do Bots Have First Amendment Rights? Politico, 2018.

[79] Jared Schroeder. Are bots entitled to free speech? Columbia Journalism Review, 2018.

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

[81] 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.

[82] 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.

[83] Michael H. Goldhaber. The attention economy and the net. *First Monday*, 2(4), 1997.

[84] Josef Falkinger. Attention economies. *Journal of Economic Theory*, 133(1):266–294, 2007.

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

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

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

[88] 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.

[89] D Centola and M Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734, 2007.

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