Vulnerabilities of the Online Public Square to Manipulation

Anonymous

Keywords: social media, agent-based simulation, misinformation, platform policy, moderation

Extended Abstract

Social media, 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 ethical challenges posed by 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 adversarial manipulation tactics on the quality of content. We find that social media features such as high information load, limited attention, and the presence of influentials exacerbate the vulnerabilities of online communities. Infiltrating a community is the most harmful tactic that bad actors can exploit and the most likely to make low-quality content go viral. The harm is further compounded by inauthentic agents flooding the network with engaging low-quality content, but is mitigated when influential or vulnerable individuals are targeted. These insights suggest countermeasures that platforms could employ to increase the resilience of social media users to manipulation.

Method

We model information diffusion in a social media platform such as Twitter or Instagram. The information system is a directed network with nodes representing accounts and links representing follower relations (Fig. 1). Similar to real-world platforms, agents can post new messages or reshare messages from their feeds. Messages shared by an account in turn appear on the news feed of their followers in the order in which they were shared; old messages are discarded. The probability each agent introduces a new message into the system — μ models *information load* of the system, and the size of news feeds — α represents the *limited attention* of an agent.

Even though people prefer quality content [1], their sharing behavior is mediated by other factors such as engagement and beliefs. Our model accounts for this by assigning two distinct attributes to each message m: *engagement* e_m — the likelihood that the message is actually shared by agents, and *quality* q_m — the objective, desirable properties such as originality or truthfulness. An agent does not know the real quality of the messages in their feed; a message is reshared with probability proportional to its engagement. Deceptive posts may have low quality and high engagement. 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 be even more likely to spread virally than high-quality information [4]. Low-quality content may be novel, clickbait, ripped from headlines, and/or may appeal to people's political, emotional, or conspiratorial bias.

Unlike authentic agents whose intention is to consume and share high-quality information, we define inauthentic agents (henceforth "bots") as accounts that are controlled by adversarial actors with the goal of spreading low-quality content among authentic agents. To amplify the spread of their messages, bots get authentic accounts to follow them. Two parameters β and γ model the *prevalence* of bots and their *infiltration* into the social network, respectively (Fig. 2). We assume that bots manipulate information in the network by having an engagement advantage over other agents. While the engagement of messages from authentic accounts reflects their quality ($q_m = e_m$ for any message m), bot messages have low quality but are deceptively

engaging ($q_m = 0$, $e_m > 0$ for any bot message m). The engagement differential of bot-amplified content is modeled by a *deception* parameter ϕ, the probability that bot content is irresistible ($e_m = 1$). When there is no deception ($\phi = 0$), bot messages have engagement drawn from the same distribution as those from authentic accounts. *Flooding*—captured by θ—is another way bots can amplify their influence by crowding out high-quality information [2, 3]; a bot generates θ times as many messages as an authentic account.

We explore human vulnerabilities by measuring the impact of information load μ, agent attention α, and network structural features on the system's *average quality*. Similarly, we measure how average quality deteriorates in response to bot tactics modeled by network infiltration (γ), deceptive content (ϕ), and flooding (θ); and to bot strategies targeting specific types of accounts. The system's average quality is the mean quality of the messages in the feeds of authentic agents once the system reaches a *steady state*. For each set of parameters, the measurement is averaged across multiple simulations that start with different random seeds. When bots are included in the model, we report on the *relative quality*, defined as the ratio of the average quality of authentic agents to that of a baseline without bots.

Results and Contributions

Here we introduce *SimSoM*, a minimal model of a generic social media platform. The model allows a multi-faceted exploration of 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 strategies. We present results from simulations of the model on online communities derived from an empirical follower network ($N \approx 10^4$ Twitter accounts).

We find that the cognitive limits of social media users contribute to the prevalence of low-quality information. The system's quality decreases significantly as the information load μ increases. In contrast, quality increases with α; as agents can evaluate more messages in their feeds, they are more likely to reshare high-quality information. Structural features of the empirical network also matters: while the presence of community structure does not significantly affect the overall quality, having hubs makes a network more vulnerable to manipulation.

Infiltrating a community is the most harmful manipulation tactic: when authentic agents have a $\gamma = 10\%$ probability to follow each bot, the average quality in the system is reduced to about a third. Flooding and generating attention-grabbing content have smaller effects (Fig.4).

We also quantify the harm done by bot *selective targeting*, in which bad actors focus their efforts on specific groups instead of making humans follow them uniformly at random. Counterintuitively, strategies that target influentials or politically-active agents tend to result in significantly higher overall quality ($p < 10^{-3}$). Due to space, we include only one explanation. When bots make politically left-, or right-learning agents follow them, messages get shared and forgotten rapidly within densely connected partisan communities, sparing the rest of the network from low-quality content (Fig. 3, bottom network). Random targeting which allows bots to spread low-quality broadly across the network is the most disruptive.

[1] Gordon Pennycook et al. "Shifting attention to accuracy can reduce misinformation online". In: *Nature* 592.7855 (2021), pp. 590–595.

[2] Chengcheng Shao et al. "The spread of low-credibility content by social bots". In: *Nature Communications* 9 (2018), p. 4787.

[3] Christopher Torres-Lugo et al. "Manipulating Twitter through Deletions". In: *Proc. Intl. AAAI Conf. on Web and Social Media (ICWSM)*. Vol. 16. 2022, pp. 1029–1039. DOI: 10.1609/icwsm.v16i1.19355.

[4] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151. DOI: 10.1126/science.aap9559.

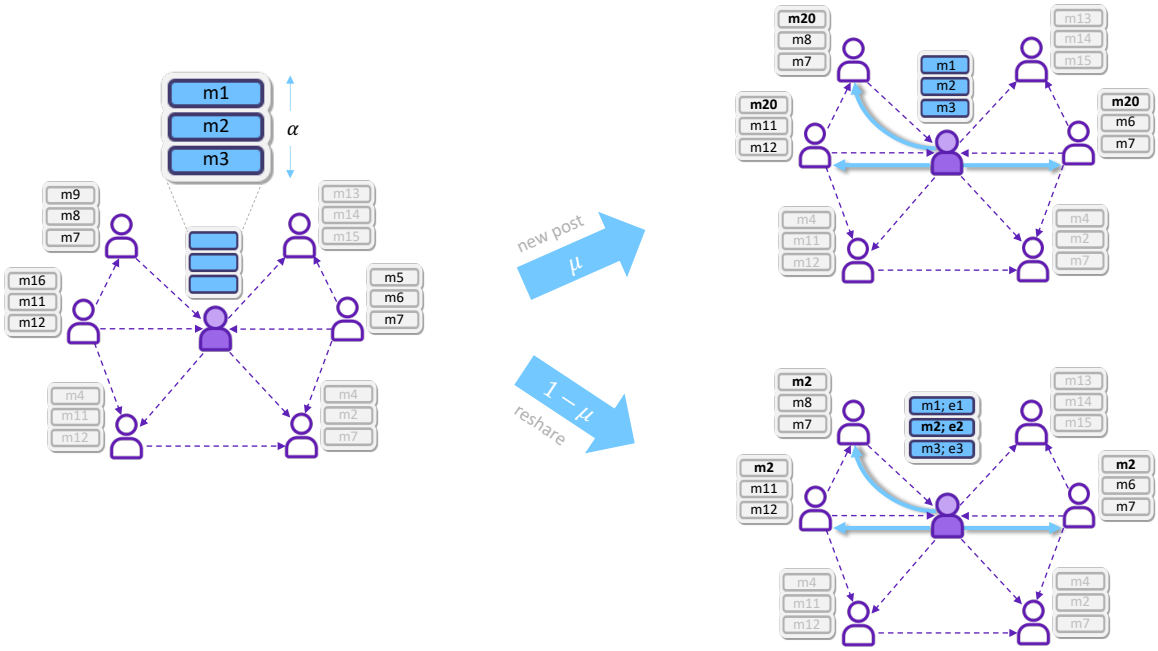

Figure 1: Illustration of the *SimSoM* model. Each node has a news feed of size α, containing the most recent messages posted or reposted by friends. Dashed arrows represent follower links; messages propagate from agents to their followers along solid links. At each time step, an agent is considered (colored node). With probability μ, the node posts a new message (here, `m20`). Otherwise, with probability $1-\mu$, the agent reposts one of the existing messages in their feed, selected with probability proportional to their engagement (here, `m2` is selected with probability proportional to `e2`). The message spreads to the node's followers and shows up at the top of their feeds.

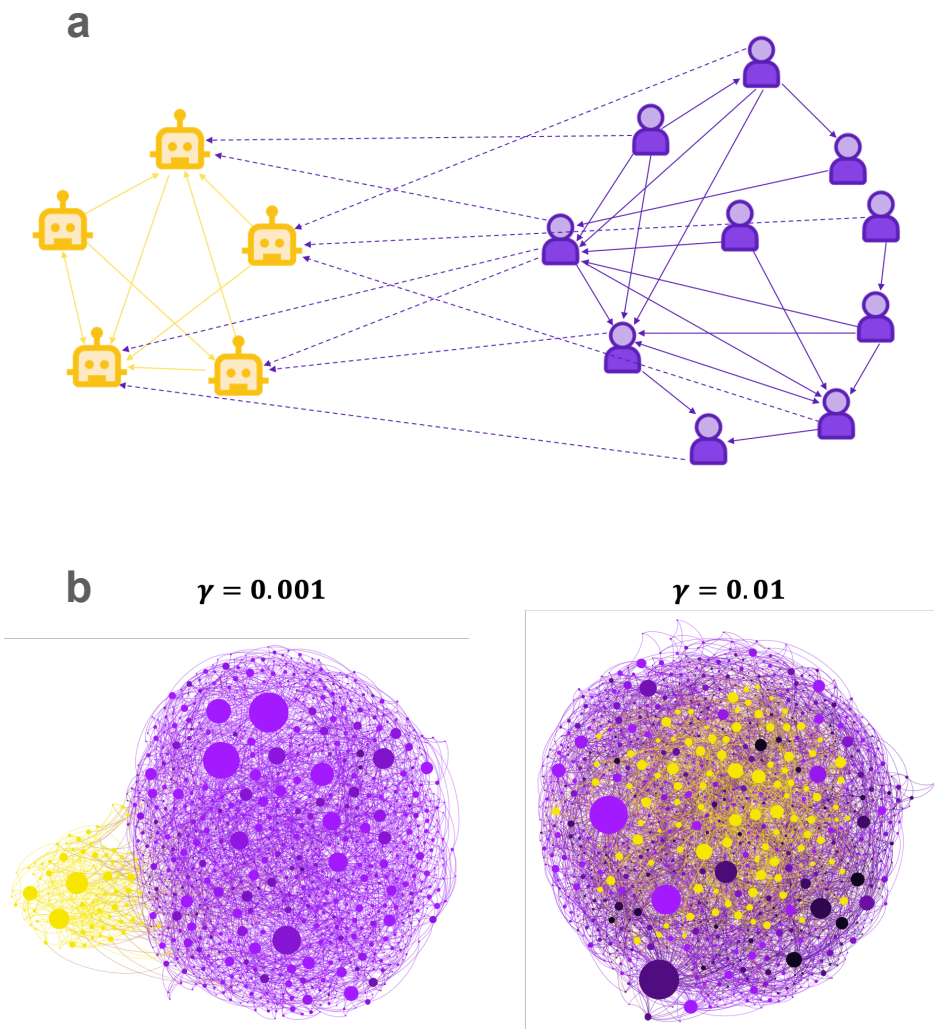

Figure 2: (a) Subnetworks modeling authentic accounts (purple nodes) and social bots (yellow nodes). (a) Follower link structure. Solid links indicate follower relations within each subnetwork. Both subnetworks have hub and clustering structure that mimics or derives from online social networks. Dashed links represent authentic accounts following bots, according to the infiltration parameter γ, which represents the probability that an authentic node follows any given bot. When $\gamma = 0$ there is no infiltration and bots are isolated, therefore harmless; the opposite extreme $\gamma = 1$ indicates complete infiltration, such that bots dominate the network. (b) Effects of bot infiltration γ on the quality of messages in synthetic networks with $N = 10^3$ nodes and $\beta = 0.1$. The parameter β is the ratio between the number of bots and authentic accounts (see Methods). Node size represents the number of followers. The darker an authentic agent node, the lower the quality of messages in their feed; black corresponds to all messages in the agent's feed having minimal quality.

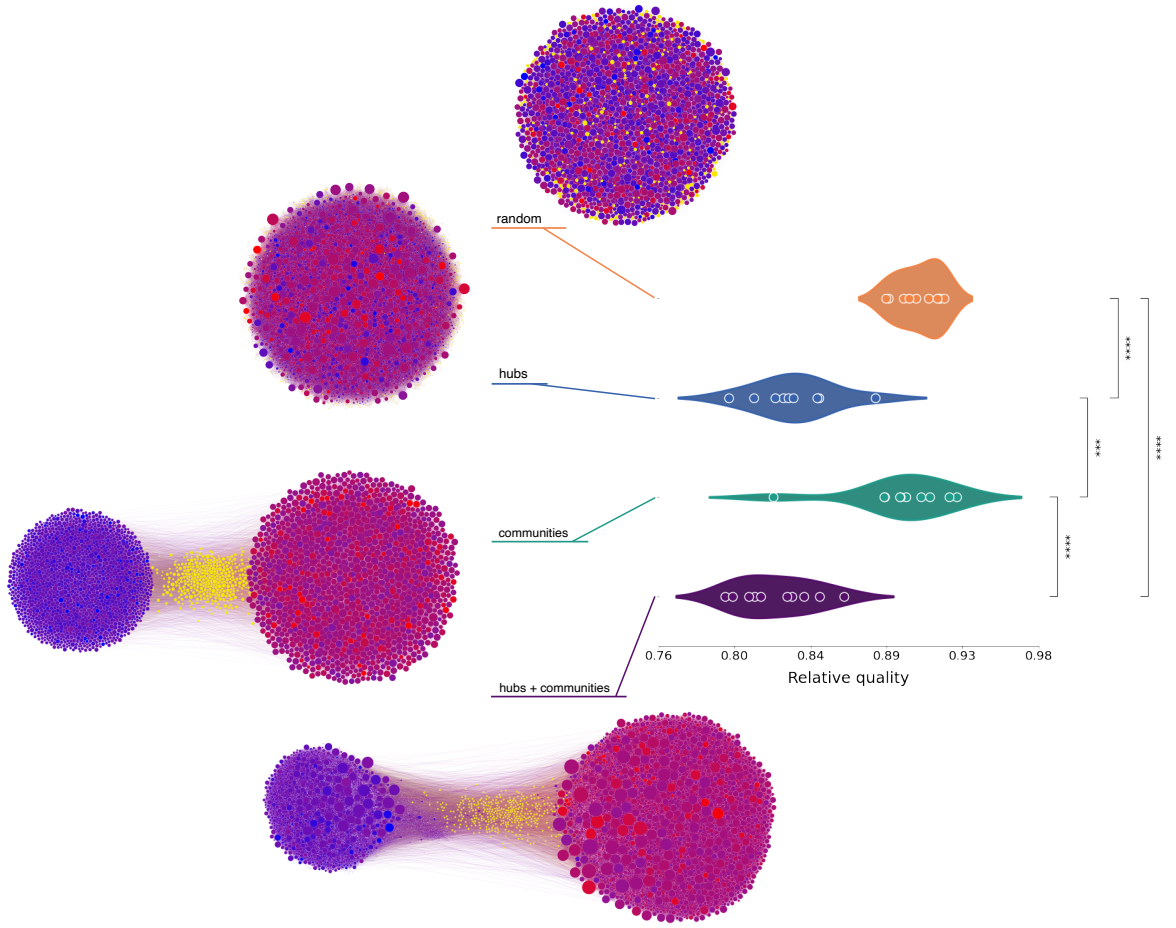

Figure 3: Impacts of different network structural features on the average information quality, relative to the scenario without bots. The original network ("hubs + communities") is visualized along with versions in which the links are shuffled while preserving communities, hubs, or neither ("random"). 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 bots. Pairwise statistical significance is calculated using Welch's two-sided t-test (*** for $p < 10^{-3}$ and **** for $p < 10^{-4}$); only significant differences are reported.

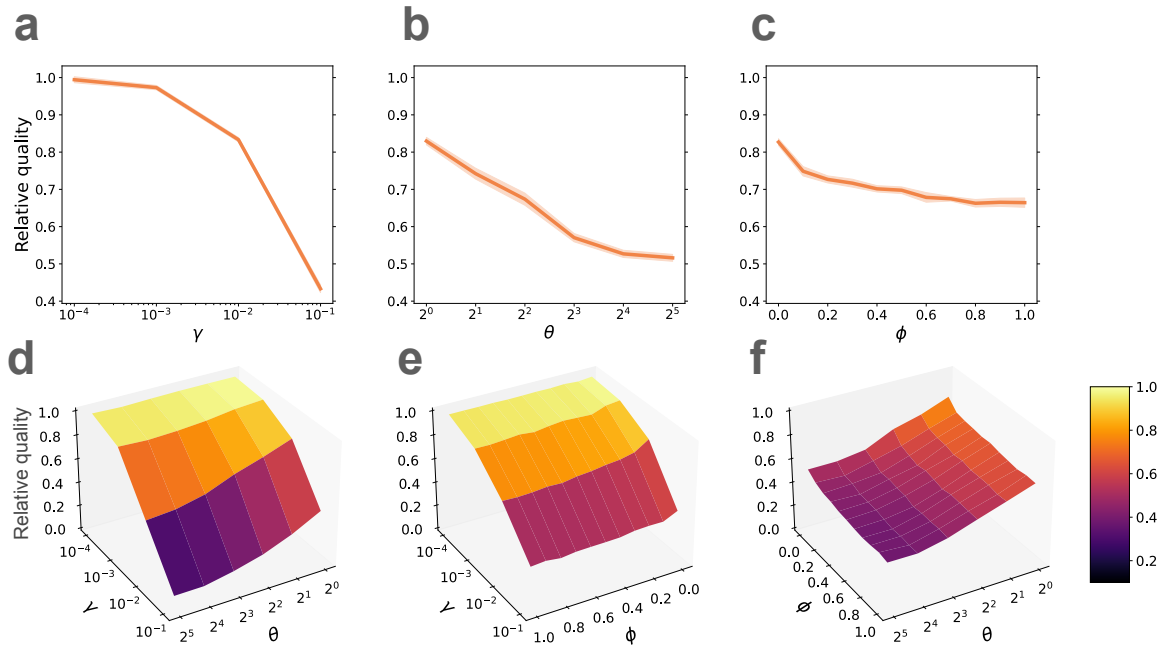

Figure 4: Effects of individual and combined bot tactics on the system's message quality, relative to the scenario without bots. (a) Varying infiltration γ, without flooding ($\theta = 1$) or deception (ϕ=0). Shading represents 95% confidence intervals across runs in panels a–c. (b) Varying flooding 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 $\gamma = 0.01$.