Social Network Analysis and Mining (2024) 14:145
https://doi.org/10.1007/s13278-024-01305-x

ORIGINAL ARTICLE

An agent-based model of cross-platform information diffusion and moderation

Isabel Murdock[1] · Kathleen M. Carley[2] · Osman Yağan[1]

Received: 2 March 2024 / Revised: 17 June 2024 / Accepted: 10 July 2024

Abstract
Social media platforms are highly interconnected because many users maintain a presence across multiple platforms. Consequently, efforts to limit the spread of misinformation taken by individual platforms can have complex consequences on misinformation diffusion across the social media ecosystem. This is further complicated by the diverse social structures, platform standards, and moderation mechanisms provided on each platform. We study this issue by extending our previous model of Reddit interactions and community-specific moderation measures. By adding a followership-based model of Twitter interactions and facilitating cross-platform user participation, we simulate information diffusion across heterogeneous social media platforms. While incorporating platform-specific moderation mechanisms, we simulate interactions at the user level and specify user-specific attributes. This allows practitioners to conduct experiments with various types of actors and different combinations of moderation. We show how the model can simulate the impacts of such features on discussions facilitated by Reddit and Twitter and the cross-platform spread of misinformation. To validate this model, we use a combination of empirical datasets from three U.S. political events and prior findings from user surveys and studies.

Keywords Cross-platform · Agent-based model · Moderation · Misinformation · Twitter · Reddit

1 Introduction

As social media use has grown over the past two decades, it has become a popular medium for conducting disinformation campaigns and a fertile environment for misinformation diffusion. Social media users often use multiple platforms and, in doing so, can spread misinformation across their various social networks on different platforms (Gottfried 2024; Papakyriakopoulos et al. 2020). Additionally, the actors behind disinformation campaigns have intentionally leveraged multiple platforms to conduct their operations, often capitalizing on the unique characteristics of each platform (Starbird and Wilson 2020; Lukito 2020). This cross-platform dynamic complicates efforts to limit misinformation diffusion and design effective countermeasures to prevent the spread of harmful content (Gatta et al. 2023).

The desire to predict and understand the spread of rumors and misinformation online, along with the ethical concerns of performing real-world experiments involving misinformation, has led to the use of models to simulate the impact of interventions. Many models of misinformation diffusion have focused on single mainstream platforms, such as Twitter and Facebook. Less attention has been paid to simulating diffusion over alternative, decentralized platforms like Reddit. However, recent studies highlight the importance of understanding information diffusion on such platforms. Investigations have led to the finding that the Internet Research Agency used Reddit, in coordination with other platforms, to heighten political tensions during the 2016 U.S. presidential election (Lukito 2020; Zannettou et al. 2019), while other work has shown how narratives from pro-Russian propaganda websites regarding the invasion of Ukraine infiltrated political communities on Reddit during the initial stages of the war (Hanley et al. 2023). In addition to political misinformation, community-based platforms,

✉ Isabel Murdock
iem@andrew.cmu.edu

Kathleen M. Carley
kathleen.carley@cs.cmu.edu

Osman Yağan
oyagan@ece.cmu.edu

[1] Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2] Software and Societal Systems, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Published online: 26 July 2024

including Reddit, have hosted “alternative health” communities and served as forums for sharing health-related misinformation (Zimdars et al. 2023; Du 2021; Kumar et al. 2022).

In this work, we aim to address two challenges in modeling the spread of misinformation and the impacts of moderation: i) the diversity of the individual platforms involved in the spread and ii) the cross-platform dynamics of user behavior. The first challenge impacts how information spreads within a given platform, through the social structures and newsfeed algorithms specific to each platform, as well as the types of moderation that can be employed. The second challenge influences the global effectiveness of countermeasures in an environment where information can flow between platforms and users can respond to moderation by migrating to alternative social media platforms. We address these issues by extending a paper where we introduced and validated an agent-based model of Reddit interactions (Murdock et al. 2023) to include a model of Twitter interactions and simulating cross-platform user engagement.

Among popular social media platforms, Reddit stands out for its decentralized approach to moderation and community management. Due to this and its community-based network structure, it provides a useful environment to simulate information diffusion in contrast to platforms with centralized moderation systems and direct user-to-user social networks. On the other hand, Twitter is a beneficial platform to model as it has been well-studied and has known trends of content diffusion and user engagement. These make it possible to validate the general processes implemented in models of Twitter. It also supports a different social structure and moderation approach from Reddit, allowing us to simulate information diffusion across two diverse platforms.

In this paper, we develop a Twitter-based model with two forms of moderation and validate it using a combination of face and empirical validation. Then, we integrate it with our existing Reddit model to produce an agent-based model that simulates user interactions within multiple heterogeneous platforms and facilitates cross-platform user activity. Both models contain platform-specific moderation mechanisms that can be used to simulate content removal and user banning. We demonstrate how the cross-platform model can be used to study the impacts of moderation on the spread of misinformation. This is the first model of cross-platform information diffusion that, to the best of our knowledge, incorporates all of the following key features:

- Users can read posts from platform-specific newsfeeds based on the network structures of each platform.
- Users update their knowledge and beliefs based on the posts they read and can spread knowledge and beliefs between their communities on each platform.
- Moderating actions are modifiable and unique to each platform based on the respective platform’s structure.
- User, moderator, and platform attributes can be tuned on each platform to study how different types of users and moderation may impact misinformation diffusion across the platforms.

The paper is organized as follows: Section 2 covers the motivation for studying cross-platform information diffusion and provides an overview of existing social media models. Section 3 lists the main characteristics of the platform environments and general facts about user activity and information diffusion on the platforms. Section 4 describes the proposed cross-platform model, including a review of the Reddit model and an explanation of the Twitter model. Section 5 provides a discussion of the Twitter model validation, which relies on a collection of tweets related to three political events and previous user surveys and studies, and presents the results of simulations with Twitter moderation. Section 6 explores the results of the cross-platform virtual experiments, and the paper concludes with discussion of the limitations in Section 7 and key takeaways in Section 8.

2 Motivation and related work

2.1 Cross-platform misinformation diffusion

The spread of misinformation across multiple social media platforms has been studied in topics ranging from natural disasters (Hunt et al. 2020) to global pandemics (Papakyriakopoulos et al. 2020). Similarly, the role that multiple platforms can play in the execution of disinformation campaigns has been examined in various contexts. One such study analyzed how Twitter and YouTube were used in a campaign against the White Helmets during the Syrian civil war. It found that “alternative” news websites were used effectively to direct users from short-form content on Twitter to large sets of videos on YouTube (Starbird and Wilson 2020). Another study, focusing on Russian disinformation during the 2016 U.S. election, found that the Internet Research Agency (IRA) may have used Reddit to test content before spreading it on Twitter (Lukito 2020). Meanwhile, an analysis of the IRA’s use of Twitter and YouTube found that the group relied on YouTube to spread news and other information on Twitter, particularly from conservative sources (Golovchenko et al. 2020).

The use of multiple platforms by users and bad actors alike has complicated the effects of moderation taken by individual platforms. Work analyzing the spread of malicious COVID-19 content over multiple mainstream and less-moderated platforms found that blocked content was effectively funneled through less-moderated platforms to avoid detection and moderation on the mainstream platforms. It showed how cross-platform connections make it harder for platforms to get rid of fake or harmful information and can obscure the actual level of such content on the platforms (Velásquez et al. 2021). However, a different study of COVID-19 conspiracy theories concluded that removing and flagging content reduced the spread of conspiracies across multiple paltforms (Papakyriakopoulos et al. 2020). Another study found that 2020 U.S. election-related tweets that had been posted by President Trump and were subsequently blocked from user engagement were posted more often and received more attention on Facebook, Instagram, and Reddit than those that were labeled or received no intervention from Twitter (Sanderson et al. 2021). This suggests that the multi-platform information environment may limit the effectiveness of single platform moderation efforts.

Besides harmful content and misinformation spreading across the social media ecosystem, users can also migrate between various platforms. One example of this was the rise in membership on alternative and less regulated platforms following the response by mainstream platforms to the January 6th attack. The deplatforming of high-profile users involved in organizing the event and removal of content promoting the protest corresponded with increased membership on the less regulated platform Gab. Gab also experienced a sustained increase in toxicity on the platform following the intervention, unlike the more mainstream platforms of Twitter and Reddit (Buntain et al. 2023). Related research on Gab has found that while users who migrate to the platform after being banned on Twitter and Reddit may have smaller audiences, they also exhibit increased activity and toxicity (Ali et al. 2021). However, there is active debate on the effectiveness of deplatforming, as a different study concluded that the tactic was effective in reducing harmful discussions and toxicity on Twitter, especially when performed simultaneously with other mainstream platforms (Jhaver et al. 2021). Together, these findings illustrate the importance of researching moderation at the ecosystem level and the benefits of developing models that can simulate the potential cross-platform effects of moderation taken by individual platforms (Sanderson et al. 2021; Ali et al. 2021).

2.2 Social media models

Approaches to modeling the spread of information across social media have taken several forms. One straightforward approach for visualizing information diffusion over social media users involves modified versions of the SIR epidemic model (Li et al. 2017). However, as information spread over social media often involves competing ideas and interactions between users with evolving beliefs and unique behaviors, these models can lack the complexity needed to reflect the real world (Serrano et al. 2015; Li et al. 2017).

In contrast, predictive models have been used to learn diffusion patterns from real-world datasets with high accuracy. Such approaches range from variants of independent cascade and threshold models to evolutionary game theory (Li et al. 2017). For example, random forests were used to predict hashtag virality on Twitter based on the community concentration present during the initial diffusion of the hashtags (Weng et al. 2013). In another work, neural networks were used to learn relationships between users within the context of linear threshold and random walk-based models (Qiang et al. 2019).

Although these predictive models can yield high-fidelity simulations of training datasets and serve as useful mechanisms for predicting misinformation diffusion, they can be challenging to generalize to new scenarios and lack interpretability. Due to this, they are not conducive to performing "what-if" analysis and experiments to study how structural and environmental changes to the relevant social media platforms will impact information diffusion or user interactions. Additionally, they rely on access to large, high-quality, and unbiased datasets.

An alternative to the epidemic and predictive models for studying information diffusion over social media is agent-based models (ABMs). Due to their agent-focused design and bottom-up approach, these models can simulate a diverse set of behaviors for different social media users. They can also provide highly explainable results and insights into how environmental factors impact information diffusion. Some of the topics explored by prior social media ABMs include the impact of emotion on user interactions (Fan et al. 2018), the adoption of competing rumors in a social media network (Kaligotla et al. 2015), and polarization (Coscia and Rossi 2022). Additionally, ABMs have been used to model social media behavior and communications during natural disasters (DiCarlo and Berglund 2021; Du et al. 2017) and health-related events (Sobkowicz and Sobkowicz 2021).

Most of these existing models facilitate user interactions based on direct user-to-user networks. While these types of relationships are the underlying structure of many platforms, such as Twitter and Weibo, they do not reflect how connections are formed on a platform like Reddit. Therefore, our Reddit model uses a user-to-community network structure, and posts are shared between users indirectly through subscriptions to the same communities. A key benefit of this approach is that it allows different communities to have different policies and moderators that impact what can be posted in a given community. This aspect of our model sets it apart from previous work that has examined how different community structures lead users to *self-censor* (Cabrera et al. 2021) and how online rejection can make users vulnerable to *radicalization* (Haddad et al. 2021).

In contrast to the lack of models focused on community-based platforms, many models have been designed to simulate information diffusion over Twitter (Serrano et al. 2015). Relevant to our work, agent-based models with this focus have been used to simulate the impacts of algorithmic curation on misinformation diffusion and polarization (Gausen et al. 2022), evaluate the influence of bots on user beliefs (Averza et al. 2022; Beskow and Carley 2019), and explore the impact of social-cyber maneuvers in deterring disinformation campaigns (Blane et al. 2021). The Twitter model used in this paper draws on key elements included in these previous models. First, it allows for the specification of individual user behavior through user attributes, which enables the implementation of malicious agents and good actors. Second, it models changes in users' beliefs in response to the content they read, in addition to their acquisition of knowledge. Finally, while we do not explore the impact of newsfeed algorithms in this work, the Twitter and Reddit models both use a newsfeed system that can be modified in future work to simulate the effects of prioritizing different types of content.

2.3 Multi-platform models

A few previous agent-based models have explored simulating user behavior across multiple platforms. One such study used machine learning models to determine user behavior when simulating Twitter and Reddit users' engagement with tweets and posts (Murić et al. 2022). It found that combining machine learning to learn agent behaviors with an explicit modeling of bursts in activity produced results closest to real-world information propagation measures. Other research has explored using variations of SIR and Bass models to investigate diffusion across multi-layer and multi-platform networks (Kim et al. 2013; Yağan et al. 2013; Tian and Yağan 2022). In particular, they examined how information can spread faster and further when users are connected to additional, conjoining networks (Yağan et al. 2013) and explored how different categories of information have different cross-platform tendencies on social media (Kim et al. 2013). Related research on the spread of misinformation on correlated multiplex networks found that characteristics like interlayer correlation impacted the outbreak of misinformation and concluded that more research was urgently needed in this area (Xian et al. 2019).

We build on these existing studies of diffusion over multilayer networks by building a multi-platform model that not only facilitates knowledge and belief diffusion across heterogeneous platforms but also allows for the simulation of platform-specific moderation efforts. Our model also allows incorporating cross-platform user responses to moderation whereby practitioners can study how moderation on a single platform may impact the spread of misinformation on alternative platforms. Consequently, the paper contributes to the emerging field of social cybersecurity by developing a model that can be used to study how influence campaigns and misinformation diffusion on both centralized and decentralized platforms can be mitigated in a meaningful cross-platform manner (Carley 2020).

3 Background

3.1 Reddit structure and dynamics

Reddit serves a beneficial function by allowing users to connect with others around common interests and share helpful information. However, the risks posed by misinformation to both the health of individuals and of democratic political systems necessitate more effective intervention strategies. This is even more significant as Reddit is one of the most-visited websites worldwide, with over 1.5 billion monthly visits and about a billion monthly users. In the U.S., the platform has ranked in the top 10 most used social network websites (Dixon 2022).

Reddit is a community-based platform where users join communities, called *subreddits*, based on their interests. Within the subreddits, users can make posts and comment on existing posts and comments. They can also react to posts and comments by voting them up or down. This results in each post and comment having a score (i.e., the number of upvotes minus downvotes). Users can view new posts through their "news feed," which displays posts based on the subreddits they have subscribed to, with the order of the posts determined by their scores or recentness. The subreddits decide what their members can post and view within their communities by having their own rules and moderators.

In terms of the behavior of users on Reddit, prior work has found that most users prefer to browse content passively and infrequently interact with posts or comments (Medvedev et al. 2019). When users do interact with content, lower-effort activity is more popular, with voting being the most common form of engagement, followed by commenting and posting (Singer et al. 2014). Furthermore, a small percentage of users is responsible for the majority of the posts on the platform. One study found that the number of posts made by users in the dataset follows an asymptotic power-law decay (Thukral et al. 2018).

Considering the characteristics of posts made on Reddit, most receive a small number of comments and stop getting comments within a day of being posted. However, a significant yet diminishing number of posts accumulate many comments (Thukral et al. 2018). This reflects how posts that receive higher scores are more likely to be seen by users, which results in a positive feedback loop of them receiving more votes and comments and remaining active longer.

The Reddit model proposed in our prior work (Murdock et al. 2023) reflects these trends in user behavior and information diffusion. Compared to existing models, it provides a more realistic environment for Reddit interactions by modeling users with different activity levels and propensities toward posting, voting, and commenting. Our model also considers the scores and recentness of posts and the users' subreddit subscriptions when determining the content that users interact with. In our previous work, we validated the model by comparing it to Reddit datasets and information diffusion trends identified by previous studies (Murdock et al. 2023). Though we do not include this validation analysis in the current paper, we describe the model in Section 4.2 and use the input parameters outlined in our previous paper for the cross-platform virtual experiments. This allows us to draw on the validation from the prior work for the experiments performed in this paper.

3.2 Twitter structure and dynamics

As of 2023, almost a quarter of U.S. adults reported using Twitter (Gottfried 2024), and previous reports found that almost half of Twitter users visit the site daily (Auxier and Anderson 2021). While the platform (recently renamed "X") has been used to share breaking news, discuss political events, and engage in pop culture debates, it has also been used for disinformation and influence campaigns (Lukito 2020; Nimmo et al. 2020) and facilitated the spread of rumors and misinformation (Allcott et al. 2019). Consequently, understanding misinformation diffusion on the platform and simulating the impacts of moderation efforts may be useful for improving the quality of content on the platform.

Users on Twitter engage with each other based on directed, followership connections. These connections influence the content that users are served through their newsfeeds. On the platform, users can share information through short messages called tweets. They can also respond to the tweets they view through retweets, replies, quote tweets, and likes. *Retweets* are direct reshares of tweets, while *quote tweets* are reshares that include a comment on the tweet. Meanwhile, *replies* are just comments made on a tweet. Through these actions, users can spread content to their followers and increase the reach of a given tweet.

Similar to user behavior on Reddit, many previous studies have found that most Twitter activity is concentrated among a small percentage of users. For example, one study found that the most active 25% of U.S. adults on Twitter produced 97% of all tweets, while the bottom 75% of users posted a median of zero tweets per month (McClain et al. 2021). Similarly, another study found that 75% of Twitter users could be considered lurkers (Antelmi et al. 2019). Earlier work found an even more skewed distribution of tweet authors, with the top 2% of users creating 80% of tweets and the top 20% responsible for nearly all content (Liang and Kw 2015). User posting activity may also be correlated with frequency of use, with lurkers reporting to visit the platform less frequently than more active tweeters (McClain et al. 2021).

As for the characteristics of tweets, some work has investigated the properties of the retweet and reply trees that are formed when users either retweet or reply to tweets. Prior work suggests that the sizes of retweet and reply trees approximate a power-law distribution (Nishi et al. 2016; Kwak et al. 2010). This indicates that most tweets receive little reaction while few go viral, similar to the behavior on Reddit. It is consistent with other research that discovered that URL cascade sizes on Twitter also follow a power-law distribution, with the largest cascades tending to be generated by users with many followers (Bakshy et al. 2011). Using these and other established findings regarding Twitter, combined with a dataset of 40 M tweets, we validate the outputs of the Twitter model in Section 5.

Our decision to model Twitter stems from its contrast with Reddit's design in two main ways. First, Twitter uses a followership-based structure with direct user-to-user relationships, as opposed to Reddit's community-based setup. Second, the platform primarily engages in centralized moderation practices rather than the decentralized approach of Reddit. These two distinctions allow for the simulation of information diffusion and moderation in two heterogeneous environments and are reflected in our models of Reddit and Twitter user interactions, as outlined in the following section.

4 Model description

The cross-platform agent-based model is developed using the Construct API.[1] Construct is an agent-based dynamic network framework that models agents' knowledge, beliefs, and evolution through interactions with other users (Dipple et al. 2022). It has previously been used to model the spread of knowledge and beliefs related to the Arab Spring and the social and behavioral characteristics that lead to revolutions (Schreiber and Carley 2013; Joseph et al. 2014). Since the Construct API provides baseline classes and network management functions, it is a useful framework for creating social media-based models.

The cross-platform model acts as a wrapper for the Reddit and Twitter models. It handles cross-platform user reactions to moderation and determines which platform the users are active on for a given timestep. The Reddit model, introduced in our previous paper, was built using the social_media_no_followers class provided by the Construct API. It has since been integrated into the API and is available for public use. Meanwhile, the Twitter model is largely based on

[1] https://github.com/CASOS-IDeaS-CMU/Construct-API.

the twitter_with_followers class, also provided by the API. Minor modifications have been made to the existing model to increase compatibility with the Reddit model, and we have added multiple moderation mechanisms to the model. Both models can be used independently to perform single-platform simulations.

To describe the complete model, we first outline the main framework that underlies the Reddit, Twitter, and cross-platform models. We then review the user behavior and moderation mechanisms provided by the Reddit model. Next, we describe user interactions on the Twitter model and its moderation system. Finally, we explain how the cross-platform model facilitates user participation within both models and handles cross-platform moderation responses.

4.1 Main framework

The Reddit, Twitter, and cross-platform models are discrete-time models that simulate users logging onto a given platform. Users who are active during a given timestep view content from their personalized news feeds. The newsfeed algorithms are specific to the respective platform and reflect its structure. Based on user-specific attributes, the users may choose to make posts during any timestep that they are active. They may also decide to respond to the posts that they read by commenting, in the case of Reddit, or retweeting, replying, or quote tweeting, in the Twitter model.

After viewing a post, users update the information, represented as bits called *knowledge*, they are aware of. These knowledge bits represent abstract statements or news stories that could be true or false. Users also update their trust of each piece of knowledge, called *knowledge trust*, based on the posts they read at each time step. The model tracks the knowledge that each agent is aware of with an agent-to-knowledge binary network, called the *knowledge network*, where the links indicate whether the agent has seen the knowledge item before. Similarly, to track the users' knowledge trust, the model maintains another agent-to-knowledge network, the *knowledge trust network*, where the link values are floats that range from 0 to 1, with values closer to 1 indicating higher trust in the associated knowledge item and lower values representing lower trust.

Updates to the knowledge network, $\mathbf{K}$, are made when a user reads a post that contains a new knowledge item. When a user i reads a post at time t with knowledge index b, the associated link in the knowledge network becomes 1, i.e., we set $\mathbf{K}_t(i, b) = 1$. Updates to the knowledge trust network, $\mathbf{T}$, take into account both the user's prior trust of the knowledge index, b, and the trust stored with the post, p. If a user does not have a prior trust associated with the given knowledge item, b, represented by the trust value being set to -1, it adopts the knowledge trust value of the first post it sees containing the knowledge item. The knowledge trust network is updated as:

$$\mathbf{T}_t(i,b) = \begin{cases} (1 - update\_rate) * \mathbf{T}_{t-1}(i,b) + update\_rate * p & \text{, if } \mathbf{T}_{t-1}(i,b) \neq -1 \\ p & \text{, if } \mathbf{T}_{t-1}(i,b) = -1 \end{cases}$$

where the *update_rate* controls how quickly the users update their trust based on the trust values they observe in the posts and comments. For the experiments conducted in this paper, the *update_rate* is set to 0.05, but this could be changed in future work.

While users update their trust regarding the knowledge bits by default, we introduce a user-specific attribute called *can receive trust*, which can be used to prevent certain users from updating their trust values when they view posts or comments. This attribute makes it possible to simulate users who are convinced of their views and only aim to influence others.

4.2 Review of the Reddit agent-based model

The Reddit model was published and validated in earlier work (Murdock et al. 2023). We now provide a review of the main features of the model. In terms of agents, the Reddit model contains two main types: *users* and *moderators*. The users reflect individuals who join subreddits to read posts and make posts and comments. The moderators enforce the rules of the subreddit by removing posts and banning users. Each agent has its own properties that determine when the agent is active and which actions they take. The full set of attributes available is described in subsection 4.5.

There are three main structures that are fundamental to the Reddit model: the *subreddit membership network*, the personalized *newsfeeds*, and the *banned user network*. The subreddit membership network specifies which subreddits each agent is a member of and is used to control both what posts and comments the users can view and the subreddits that the moderators can act in. It is specified at the start of a given simulation. The personalized news feeds "serve" posts to the users based on their subreddit subscriptions and order the posts according to a combination of their scores and the recentness. We use the previously public version of Reddit's ranking algorithm to rank the posts in each user's feed, which prioritizes newer and higher scoring posts (Salihefendic 2015).

The banned user network, in combination with the moderator agent attributes, facilitates subreddit-specific moderation. The banned user network is a user-to-subreddit network

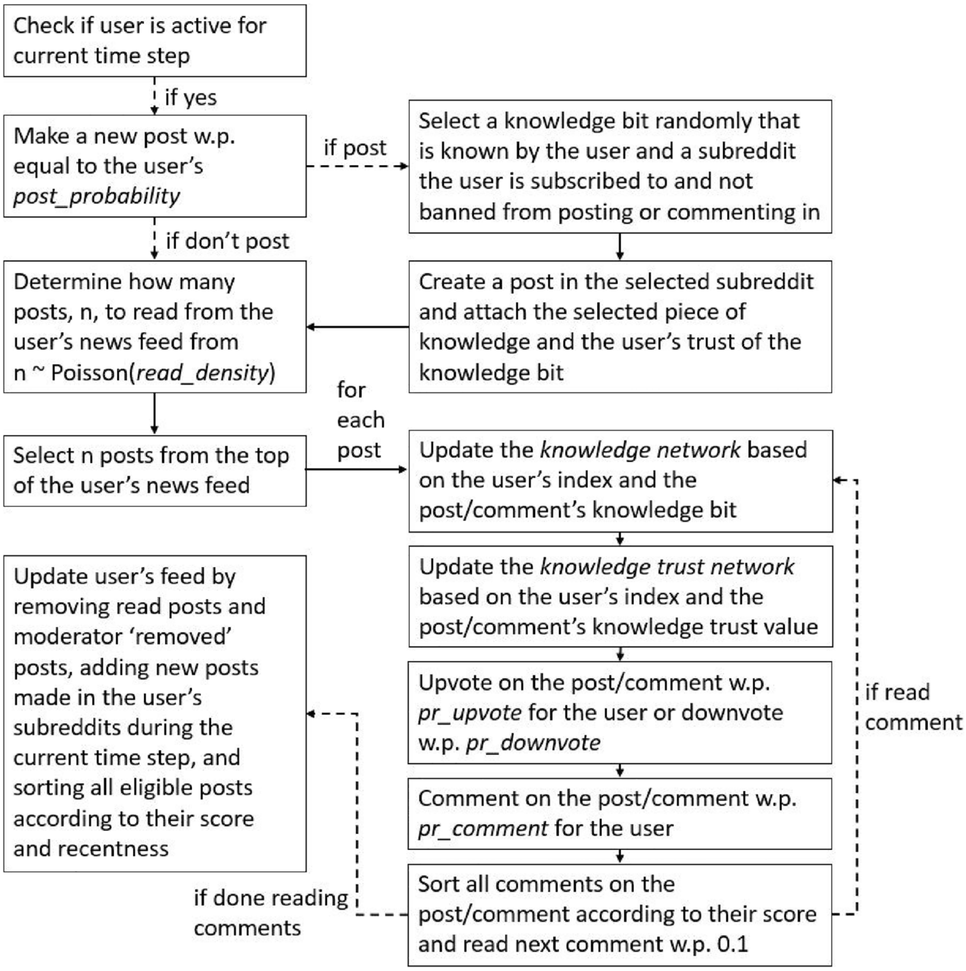
Check if user is active for current time step
if yes
Make a new post w.p. equal to the user's *post_probability*
if post
Select a knowledge bit randomly that is known by the user and a subreddit the user is subscribed to and not banned from posting or commenting in
Create a post in the selected subreddit and attach the selected piece of knowledge and the user's trust of the knowledge bit
if don't post
Determine how many posts, n, to read from the user's news feed from n ~ Poisson(*read_density*)
Select n posts from the top of the user's news feed
for each post
Update the *knowledge network* based on the user's index and the post/comment's knowledge bit
Update the *knowledge trust network* based on the user's index and the post/comment's knowledge trust value
Upvote on the post/comment w.p. *pr_upvote* for the user or downvote w.p. *pr_downvote*
Comment on the post/comment w.p. *pr_comment* for the user
Sort all comments on the post/comment according to their score and read next comment w.p. 0.1
if read comment
if done reading comments
Update user's feed by removing read posts and moderator 'removed' posts, adding new posts made in the user's subreddits during the current time step, and sorting all eligible posts according to their score and recentness

Fig. 1 User actions during each time step of the Reddit model (Murdock et al. 2023)

whose edge weights track the number of times a given user's posts or comments are removed in a specific subreddit. The moderator agents remove posts based on the moderators' attributes and the knowledge and trust associated with the given post or comment. This allows each subreddit to have moderators with specific attributes and to set a threshold for the number of times a user can "break the rules" before becoming banned from making posts and comments in the subreddit.

In summary, these structures allow the following main platform features to be implemented in the model:

- Users belong to subreddits, which impact the posts and comments they see.
- Users share information through posts and comments.
- Users vote up or down on posts and comments.
- News feeds prioritize new and higher-scoring posts.
- Subreddit moderators remove content and ban users.
- Subreddits can have heterogeneous rules and moderation thresholds.

4.2.1 User behavior

The first type of agent in the Reddit model are users. Before each simulation, user attributes and the subreddit membership network are specified; see subsection 4.5. These attributes define when each user is logged onto the platform and their likelihood to read, post, comment, and vote. The inputs also describe the knowledge bits that each user has at the start of the simulation and their associated trust in the knowledge.

Once the initialization is complete, the model runs for a specified number of time steps. During each time step, the model loops through each user to check if they are active, i.e., logged on. If the user is active, they make posts, read posts and comments, update their knowledge and knowledge trust, vote, and comment based on their assigned attribute values, as shown in Fig. 1. An important part of our implementation is that the users can view and contribute to the comment trees under posts, as they do in the real world on Reddit. This then impacts their trust in the knowledge item associated with the original post. Any comments made on a post or a comment contain the knowledge item associated

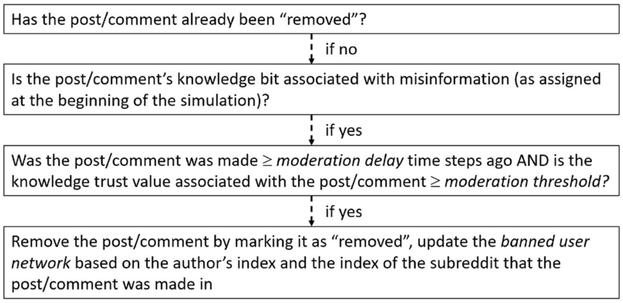
Has the post/comment already been "removed"?
if no
Is the post/comment's knowledge bit associated with misinformation (as assigned at the beginning of the simulation)?
if yes
Was the post/comment was made ≥ moderation delay time steps ago AND is the knowledge trust value associated with the post/comment ≥ moderation threshold?
if yes
Remove the post/comment by marking it as "removed", update the banned user network based on the author's index and the index of the subreddit that the post/comment was made in

Fig. 2 Conditions that the moderators check before "removing" a post or comment from their subreddit(s) in the Reddit model (Murdock et al. 2023)

with the parent post/comment. This reflects how comments on Reddit tend to discuss the topic outlined in the original post. However, the knowledge trust associated with the new comment reflects the knowledge trust value of the commenter.

4.2.2 Moderator behavior

The second type of agent in the Reddit model are moderators. These are agents who remove content and ban users within the subreddits. Their behavior is determined by the subreddit membership network and two other attributes specified for each moderator: *moderation delay* and *moderation threshold*. They also rely on the *misinformation* attributes associated with the knowledge bits that can flag knowledge as misinformation.

At each time step in the simulation, the moderators iterate over every post and comment made in their subreddit(s), specified by the subreddit membership network. As shown in Fig. 2, they check if the post's knowledge bit is associated with misinformation. If it is, they check if the post was made at least *moderation delay* time steps ago and if the knowledge trust value associated with the post is greater than or equal to the *moderation threshold*. Since the post contains a knowledge bit designated as misinformation, the high trust value of the post would increase other users' trust in the misinformation when they read it. Therefore, if the moderators aim to limit the spread of misinformation, they would want to limit the viewership of this type of post. Consequently, if all of these conditions are met, the moderator "removes" the post, preventing it from appearing in the users' news feeds at any future time step.

When the moderators remove a post or comment, they also increment the link weight in the banned user network that connects the author of the post or comment to the subreddit in which they made the post. The banned network is used to prevent users from posting or commenting in a given subreddit once their content has been removed a *ban threshold* number of times from the subreddit. The *ban threshold* is specific to each subreddit and can be used to reflect the strictness of different subreddits' rules.

The moderation delay and moderation threshold can also be used to vary moderation policies across the subreddits. Additionally, they can be used to model different types of moderators. For example, automated moderators are common on Reddit and can perform mundane or repetitive checks on content. While they can respond faster than human moderators, represented in the model by a shorter moderation delay, they are not as adept at handling borderline cases or considering the context of posts. Therefore, they may have a higher moderation threshold for removing posts in the model.

4.2.3 Validation

We used a combination of input, face, and empirical validation to validate the described Reddit model. First, we performed input validation of the values selected for the model attributes and networks of our experiments. This grounded our users' behaviors in previous user studies and helps ensure they follow real-world tendencies. We also drew our subreddit membership networks from empirical datasets to produce more realistic user-to-subreddit network structures. We subsequently used a combination of face and empirical validation to examine whether our model generates posts with characteristics that align with patterns identified in prior work. We supported this analysis with a collection of three Reddit datasets from the 2020 U.S. presidential election, the Dobbs v. Jackson Women's Health Organization U.S. Supreme Court decision, and the 2022 U.S. midterm election. This set of over 100K posts and 800K comments helped us evaluate whether the user behaviors and content diffusion produced by the model are consistent with the real world. The findings of our validation analysis can be found in our previous work (Murdock et al. 2023).

4.3 Description of the Twitter agent-based model

The Twitter model only has one main type of agent, *users*. Within this model, the users reflect individuals who can follow other users to read tweets. Consistent with the Reddit model, each user has its own properties that determine when the agent is active and which actions they take. The full set of attributes available is described in subsection 4.5. Since moderation within the Twitter model is handled at the platform level, there are no moderator agents in this model.

There are three main structures that are fundamental to the Twitter model: the *twitter follower network*, the personalized *newsfeeds*, and the *removal counter*. The twitter follower network specifies which users each agent follows and is used to control what tweets, retweets, replies, and

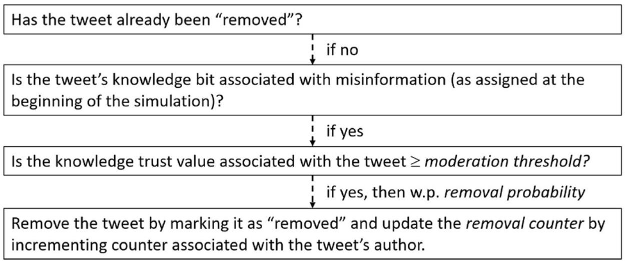
Has the tweet already been "removed"?
if no
Is the tweet's knowledge bit associated with misinformation (as assigned at the beginning of the simulation)?
if yes
Is the knowledge trust value associated with the tweet ≥ *moderation threshold?*
if yes, then w.p. *removal probability*
Remove the tweet by marking it as "removed" and update the *removal counter* by incrementing counter associated with the tweet's author.

Fig. 3 Conditions that are checked at each timestep in the Twitter model before content is removed

quote tweets the users can view. It is specified at the start of a given simulation. The personalized news feeds "serve" tweets to the users based on their followership networks and order the tweets according to a combination of the number of reactions they have received and their recentness. They also filter out content that has been moderated. The removal counter keeps track of the number of times each user's content has been removed from the platform.

Moderation is performed at the platform level by reviewing all active (i.e., not previously removed) tweets at each timestep, see Fig. 3. First, the model checks whether each active tweet's knowledge bit is associated with misinformation. If it is, the model checks if the tweet's trust value is greater than the *moderation threshold*. Like in the Reddit model, the moderation threshold helps remove content with high trust in knowledge that is associated with misinformation. If both of these conditions are met, the model removes the tweet with probability equal to the *removal probability* and increments the counter associated with the tweet's author in the removal counter. Once a user's content has been removed a *ban threshold* number of times, they are banned from making any type of tweet for the remainder of the simulation.

The moderation threshold, removal probability, and ban threshold are currently fixed values set at the beginning of the simulation. The moderation threshold can be used to change how strict the model is about misinformation. Meanwhile, the removal probability can be used to change how long it takes for undesirable content to be removed. The ban threshold determines how many "strikes" users get before they can no longer interact with others on the platform. Future versions of the Twitter model may improve these moderation mechanisms, such as by making the removal probability proportional to the number of reactions a tweet has received or the trust value associated with the tweet. However, for the current version, these parameters allow practitioners to vary both the likelihood of content removal and the strictness of user banning and explore their potential effects on information diffusion.

In summary, the key structures of the Twitter model allow the following main platform features to be implemented:

- Users follow others, which impacts the tweets they see.
- Users share information through tweets, retweets, replies, and quote tweets.
- News feeds prioritize tweets that are more recent and have garnered larger reactions.
- Moderation is performed at a platform level but can be adjusted to focus more on content removal or user banning.

4.3.1 User behavior

Similar to how users are processed in the Reddit model, user attributes and the followership network are specified before each simulation; see subsection 4.5. These attributes define when each user is logged onto the platform and their likelihood to read, tweet, retweet, reply, and quote tweet. The inputs also describe the knowledge bits that each user knows at the start of the simulation and their associated trust in the knowledge.

Once the initialization is complete, the model runs for a specified number of time steps. During each time step, the model loops through each user to check if they are active, i.e., logged on. If the user is active, they make tweets, read tweets, update their knowledge and knowledge trust, and react to tweets based on their assigned attribute values. Consistent with the Reddit model and real-world interactions on Twitter, users can view and contribute to the retweet and reply trees of tweets or start new potential trees by making tweets or quote tweets. All retweets, replies, and quote tweets contain the knowledge item associated with the parent tweet. However, the knowledge trust associated with the new post reflects the knowledge trust value of the new post's author.

4.3.2 Validation approach

Since most users on Twitter rarely tweet on the platform, it can be challenging to measure knowledge or information diffusion across the Twitter user base. Yet, validation of the Twitter model is needed in order to draw meaningful conclusions from it. Due to this, we use a combination of input, face, and empirical validation to evaluate various aspects of our model, similar to what was done with the Reddit model.

First, we perform input validation of the values selected for the user attributes and networks of our experiments. This involves using previous user surveys and studies of social media data to help ensure that the simulated users' behaviors follow real-world tendencies. We also compare our generated followership networks to metrics identified in

Table 1 Input attributes for nodesets

Platform	Attribute	Values	Description
User attributes			
Reddit & Twitter	read_density	Integer ≥ 1	Average number of posts to read during an active time step
Reddit & Twitter	post_probability	Float from 0-1	Probability of making a post or tweet during an active time step
Reddit	pr_upvote, pr_downvote, pr_comment	Float from 0-1	Probability of upvoting, downvoting, or commenting on a read post
Twitter	pr_repost, pr_reply, pr_quote	Float from 0-1	Probability of retweeting, replying, or quote tweeting a read tweet
Reddit & Twitter	can_receive_trust	Boolean (true or false)	Whether a user updates their knowledge trust when they read posts or comments
Moderator attributes			
Reddit	moderation delay	Integer ≥ 0	Number of time steps after a post is made that the moderator can remove the content
Reddit	moderation threshold	Float from 0-1	Value that the knowledge trust associated with the post must be *ge* for the content to be removed
Subreddit attributes			
Reddit	ban threshold	Integer ≥ 1	Number of times a users' content must be removed within the subreddit before they are banned from making posts or comments
Knowledge attributes			
Reddit & Twitter	misinformation	Boolean (true or false)	True indicates the knowledge item is fake information

real Twitter networks. We subsequently use a combination of face and empirical validation to examine whether the model generates tweets, retweet trees, and reply trees with characteristics that align with patterns identified in prior work. We supplement this analysis with a collection of three Twitter datasets from the 2020 U.S. presidential election, the Dobbs v. Jackson Women's Health Organization U.S. Supreme Court decision, and the 2022 U.S. midterm election. This set of more than 40 M tweets helps us evaluate whether the user behaviors and content diffusion produced by the model are consistent with the real world.

4.4 Cross-platform model

The cross-platform model builds on the Reddit and Twitter models by facilitating the participation of agents on both of the individual platform models. The cross-platform model maintains the users' knowledge and trust networks but allows the users to interact and update their knowledge and trust values based on their activities on each platform. This enables the transfer of knowledge between the platforms and reflects how interactions across multiple platforms influence users in the real world.

For each timestep in the cross-platform model, users may be active on one or neither of the individual platform models. To determine when users are active on each platform, the model takes as input an *active agent by platform network*. Similar to the *user active time networks* that are used by the Reddit and Twitter models to determine when users are logged on, this network is a user-to-timestep network. However, the edges in this network indicate which platform a user is active on for the given timestep. With this setup, the amount of cross-platform links between Twitter and Reddit is determined by the number of users who log onto both platforms. When a user is active on one of the platforms, their behavior is governed by the individual platform model they are interacting on. Based on the posts they read from their newsfeed on the given platform, they update their knowledge and trust values, which are tracked centrally by the cross-platform model.

In addition to allowing for the transfer of knowledge and trust between the platforms, this model design allows for the implementation of cross-platform responses to moderation. By storing the *active agent by platform network* as a variable of the cross-platform model, the model can dynamically alter users' log on behavior and platform preferences in response to moderation. To achieve this, the Twitter and Reddit models notify the cross-platform model whenever they remove content or block a user from making a post. When either of these situations occurs, the cross-platform model examines the active agent by platform network and switches any future instances of the user logging on to the moderating platform to the alternate platform with probability *platform switch probability*. This allows the model to simulate how users may respond to moderation by migrating to platforms with fewer regulations.

Table 2 Input networks for models

Platform	Network	Link Type	Source → Target	Description
Reddit & Twitter	Knowledge network	Boolean	user → knowledge	True link values indicate the user is aware of the piece of knowledge
Reddit & Twitter	Knowledge trust network	Float between 0-1	user → knowledge	The trust that the user has in the given piece of knowledge (higher values indicate more trust)
Reddit & Twitter	User active time network	Boolean	user → time step	True link values indicate that the user is active during the given time step
Cross-platform	Active agent by platform network	Integer	user → time step	Link values indicate which platform the user is active on during the given time step
Reddit	Subreddit membership network	Boolean	user → subreddit	True link values indicate that the user subscribed to the given subreddit
Twitter	Twitter follower network	Boolean	user → user	True link values indicate that the first user follows the second user

4.5 Summary of model inputs

Many of the features and attributes of the Reddit, Twitter, and cross-platform models have been covered in the previous discussion. We present the full set of node attributes in Table 1 for completeness. They must be initialized for each of the relevant platforms before the start of a simulation. The relevant networks listed in Table 2 must also be initialized once for all of the platform models. The combination of attributes and networks can be used to create heterogeneous agents with different activity levels and posting behaviors. They can also adjust the types of moderation implemented on each platform. As discussed in the prior work, this is key for performing realistic simulations, as Reddit and Twitter users vary widely in terms of activity levels and engagement preferences.

An important note is that, within the Construct framework, the input networks and nodesets (i.e., agents, knowledge items, timesteps) are shared across the social media models. This is useful as it allows the cross-platform model to simulate agents logging onto multiple platforms and updating their knowledge and trust values based on their interactions across platforms. In doing so, the agents can transfer knowledge between platforms, as users do in real life. However, unlike the nodes and networks, the node attributes (e.g., agent attributes, moderator attributes, subreddit attributes) must be specified separately for both the Reddit and Twitter models. This is also beneficial, as it means that users can have different levels of engagement on different platforms.

The networks listed in Table 2 are important for determining the seed users who start with the knowledge items, through the *knowledge network*, and the other users they can potentially share information with, through the *subreddit membership network* or the *twitter follower network*. Additionally, while the user attributes control the types of actions users take while they are active on a platform, the *user active time network* determines how frequently each user “logs on” to the respective platform, and the *agent active by platform network* specifies which platform a user logs onto in the case of the cross-platform model.

5 Twitter model simulation results

5.1 Validation

To evaluate the validity of the Twitter model, we first perform baseline simulations using realistic input parameters for the user activity levels and engagement preferences. The inputs are based on a combination of surveys and analysis of social media data. Using data from a variety of sources helps limit the biases that can result from self-reporting errors, in the case of surveys, and hidden user behaviors, in the case of social media data analysis. Therefore, by using a variety of references to select the inputs, we can create a more robust and realistic set of user parameter values.

Table 3 provides the user-related input values and associated references used for the experiments. The time and activity-related user attributes are derived based on one timestep in the simulation representing 5 minutes. We first generate a followership network of 1,000 users by using a combination of configuration and preferential attachment models. The modified version of the configuration model is used first to generate a bidirectional network with a power-law degree distribution and noticeable clustering (Newman 2009). Then, the preferential attachment mechanism adds

Table 3 User-related input parameters and references

Model inputs	Factors considered	Values	References
post_probability pr_repost pr_reply pr_quote	User activity levels and engagement preferences	5% are high activity: (0.08, 0.04, 0.03, 0.02), 20% are medium activity: (0.04, 0.02, 0.02, 0.01), 25% are low activity: (0.02, 0.005, 0.01, 0.001), 50% are very low: (0.01, 0.001, 0.002, 0.001)	(Antelmi et al. 2019) (McClain et al. 2021) (Odabaş 2022)
read_density	Based on 5-minute timesteps	read_density=20 (the number of tweets read during a given timestep is drawn from Poisson(read_density))	
user active time network	Frequency of use	15% 10 times/day, 15% 5 times/day, 15% daily, 20% 3 times/week, 15% weekly, 15% monthly	(Odabaş 2022) (Auxier and Anderson 2021) (Dixon 2020)

one-way followership connections that boost the presence of highly followed users (i.e., celebrities) and reproduces the discrepancies between in-degree and out-degree distributions identified in real-world Twitter networks. After this, the users are assigned attribute and network values according to Table 3, with 100 knowledge items included in each simulation. After this initialization, each trial runs for 4,032 time steps, representing two weeks. We repeat this process to collect 100 simulation runs.

For each run of the simulation, we collect all of the tweets made by the users. This allows us to collect a dataset similar to the real-world data we collected from Twitter. Unlike our limitations in the real world, however, we can also track the agents' knowledge and knowledge trust networks, which we output every 12 hours in the simulation for each run.

5.1.1 Followership network

Within the model, users can only view content posted by the users they follow. Consequently, the structure of the followership network can have a large impact on information diffusion. The followership network used in our simulations is randomly generated for each run. In this section, we compare the features of our generated networks to those that have been identified in real-world Twitter networks. In particular, we focus on three key properties: degree distribution, transitivity, and reciprocity.

Multiple prior studies have established that the distribution of followers (in-degree) and followees (out-degree) on Twitter are highly skewed with heavy tails (Kwak et al. 2010; Bakshy et al. 2011). Some have found that the degree distributions follow power-law distributions with exponents between 2 and 3. Others have concluded that the distributions have exponents less than 2 or do not fit the power-law (Liang and Kw 2015). The heavy-tailed distributions arise due to popular users being present on Twitter with many millions of followers. However, such users do not tend to follow nearly as many users in return. This results in the followers degree distribution being more skewed than the followees distribution (Bakshy et al. 2011; Myers et al. 2014). One study found that while the maximum in-degree (followership) of their collected Twitter network was much greater than the maximum out-degree, the percentiles were higher for the out-degree distributions than those of the in-degree. It concluded that the typical Twitter user follows more people than they have followers, while the reverse is true for celebrities (Myers et al. 2014).

Turning to our generated followership networks, we evaluate whether these degree distribution characteristics are present in the generated degree distributions. We find that the generated degree distributions are skewed and that the distribution of followers has a larger maximum value (908) than the followees (522), see Fig. 4. However, across all of the users in the generated networks, the median number of

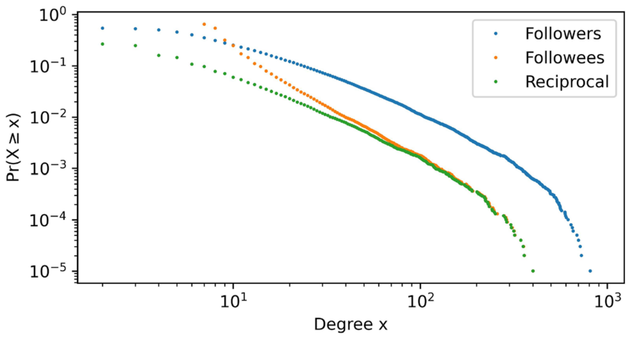

Fig. 4 The complementary cumulative distribution functions, in log-log scale, for the number of followers (in-degree), followees (out-degree), and reciprocal relationships for the users in the generated followership networks

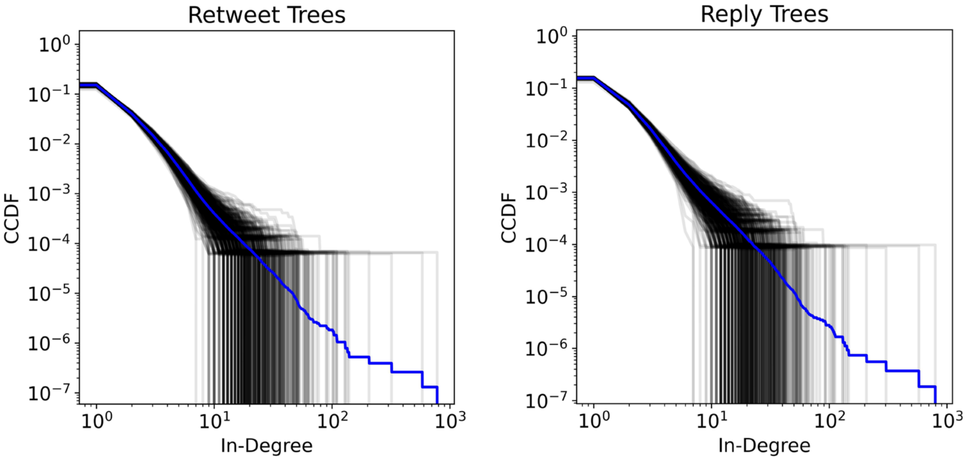

Fig. 5 The complementary cumulative distribution function (CCDF), in log-log scale, for the in-degree of all tweets involved in the retweet trees (left) and reply trees (right) from the baseline Twitter simulations. Each light grey line reflects the CCDF from one of the simulation trials. The thick blue line is the aggregated CCDF of all of the relevant trees from all of the trials

followers a user has is 5, while the median number of people a user follows is 9. This suggests that the generated networks are consistent with the prior work regarding the presence of popular users, as well as the trends of followership of typical users on the platform.

In terms of transitivity, we use the average clustering coefficient to compare the presence of connections between users who have friends in common with each other on the platform. For each generated network, we calculate the average clustering coefficient of the largest strongly connected (i.e., reciprocal) component. The average of the average clustering coefficients across the generated networks is 0.17 (*SD*: 0.04). This aligns well with prior work that measured the average clustering coefficient in a real-world Twitter network to be 0.15 for all active users and 0.12 for active users with at least two reciprocal ties (Liang and Kw 2015). Unlike other simulation studies that only use preferential attachment mechanisms to generate followership networks, our use of the modified configuration model in combination with preferential attachment allows us to generate followership networks that incorporate realistic levels of clustering.

The last metric we use to validate the generated followership networks is reciprocity. Although connections on Twitter are directed, users can decide to follow back those who follow them. This results in reciprocal relationships in the network. Such connections allow information to travel in both directions between users. Averaging across the generated followership networks, 31.1% (*SD*: 2.3) of the connections between users are reciprocal. This is consistent with the findings of previous work that reported reciprocity ranging from 22.1% to 38.3% in real-world Twitter networks (Kwak et al. 2010; Liang and Kw 2015).

5.1.2 Distributions of tweets, replies, and retweets

Shifting to focus on the outputs of the Twitter model, we now investigate whether the concentration of user activity, types of posts generated, and distributions of reactions to tweets align with previously identified trends, as well as our empirical datasets.

In terms of the concentration of users responsible for the tweets created in the simulations, we find that the results approximate statistics that prior real-world studies have reported. In the simulations, the top 10% most active users were responsible for 70% of the tweets created, while the top 25% accounted for 93% of the content. The bottom 50% of users only contributed 1% of the simulated tweets. Though slightly less skewed, this is in line with prior studies that have found that the most active 25% of U.S. adults accounted for 97% of all tweets (McClain et al. 2021) and 75% of users on Twitter could be considered lurkers (Antelmi et al. 2019). Furthermore, the bottom 75% of simulated users had a median of zero tweets made, which has also been found to be the case on Twitter (McClain et al. 2021).

Considering the different types of content generated by the users during the baseline simulations, 6.5% were original tweets, 48.7% were retweets, 39.0% were replies, and 5.8% were quote tweets. Though skewed towards retweets and replies, this breakdown is somewhat in line with a recent study of Twitter from the Pew Research Center, which found that 75% of posts from all U.S. adults on the platform were retweets and replies, with 15% being original tweets and 9% being quote tweets (Chapekis and Smith 2023). The probabilities of reposting, replying, and quote tweeting, set at the beginning of the simulations based on prior studies of user preferences, have a direct effect on this breakdown of content. Therefore, future studies interested in simulating specific types of content diffusion or categories of users could modify the probabilities to achieve the desired composition of tweets.

While informative, the distribution of different tweet types does not provide insight into the distribution of responses that the tweets received. To explore this, we

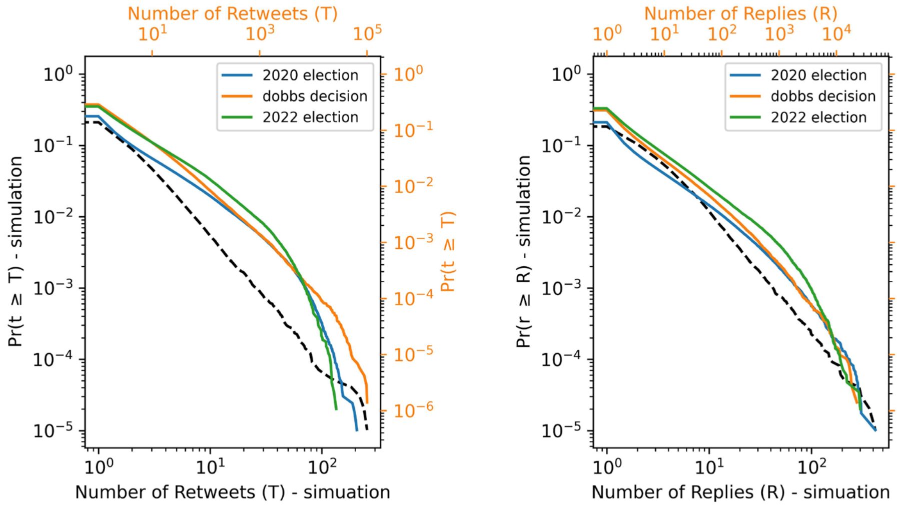

Fig. 6 The complementary cumulative distribution function, in log-log scale, for size of the retweet trees (left) and reply trees (right) that original tweets generated in the simulations (black, dashed lines) and our three Twitter datasets (solid lines). The simulation and empirical results are displayed with different axes due to the differences in scale between the number of users and tweets in the simulations as compared to real-world Twitter

collect and analyze the retweet and reply trees produced by the baseline simulations. As discussed in section 3.2, the size distribution of retweet and reply trees (i.e., the number of posts that form the respective tree) approximate a power-law distribution, with most tweets receiving few responses and a small number going viral (Nishi et al. 2016; Kwak et al. 2010). One study also found that the in-degree distribution of reply trees (i.e., the number of responses that each post in the tree received) is power-law (Nishi et al. 2016).

To compare these findings to our simulated tweets, we plot the in-degree distributions from retweets and replies in the simulation outputs, see Fig. 5. These plots of the complementary cumulative distribution functions (CCDFs) display the aggregated data from all of the trials, as well as the data from each individual trial. Although there is variation across the trials, the aggregate CCDF, in log-log scale, shows that the number of retweets and replies that any given tweet received approximates a power-law distribution. We similarly find that the size distributions of the reply and retweet trees exhibit power-law relationships.

We further investigate the distributions of the simulated retweet and reply tree sizes using our three real-world Twitter datasets. In Fig. 6, we plot the CCDFs of the tree sizes for the simulated and real-world data together. Given the smaller scale of our simulations relative to the actual number of users on Twitter and the volume of tweets made on the platform, we plot the CCDFs for the simulated and real data with separate axes. This allows us to meaningfully compare the proportional trends of the distributions using a log-log scale. Across the simulations and real-world datasets, a similar percentage of tweets belong to trees of size 1 (i.e., they did not receive any retweets or replies). This means that the Twitter model accurately simulates how much of the content on Twitter does not garner responses.

Though on a smaller scale than the real-world outputs, the simulated retweet and reply tree sizes also follow a similar heavy-tailed trend as those extracted from the empirical data, see Fig. 6. The real-world retweet trees appear to have more relatively mid-sized trees than the simulated data, while the reply trees closely follow real-world distributions. These findings help validate the Twitter model's newsfeed system. Since the newsfeed prioritizes popular and newer content, tweets with larger responses (i.e., retweets, replies, and quote tweets) are more likely to be seen and, therefore, continue to receive more attention. Similar to the findings of post growth in the Reddit model, this positive feedback loop results in a few tweets building large retweet and reply trees. Meanwhile, the rest that do not gain such immediate attention drop lower in the newsfeeds and are replaced by newer and more popular content.

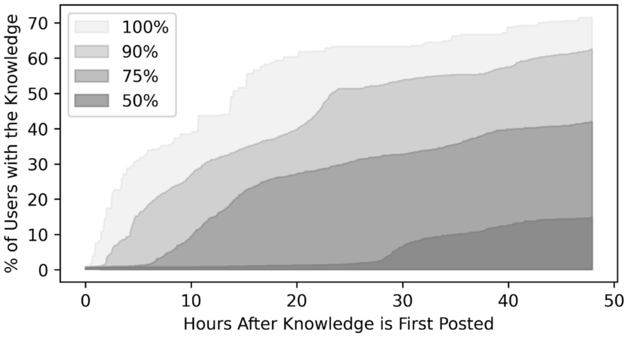

Fig. 7 Diffusion of knowledge bits in terms of the percentage of users who become aware of the knowledge in the first 48 hours after being posted for the first time. The diffusion of every posted knowledge bit is collected across every simulation. Temporal variations due to the *user active time network* and collecting the knowledge network in 12 hour increments are smoothed by taking a 12-hour moving average. Shading indicates the 50th, 75th, 90th, and 100th percentiles

5.1.3 Knowledge diffusion

A common behavior exhibited by many of the existing models of information diffusion over networks (Zafarani et al. 2014), as well as found through empirical work (English 2016; Yang et al. 2023), is an S-shaped information diffusion pattern. This occurs when information initially spreads slowly and then accelerates as more people share the information with their respective connections. Eventually, the diffusion slows as the network becomes saturated. In addition to this phenomenon, we also know that while some URLs or news stories go viral, many never gain enough traction to spread widely (Bakshy et al. 2011). We expect the Twitter model to produce similar knowledge diffusion patterns.

To study whether this is the case, we plot the knowledge diffusion of all the knowledge bits in the simulations. As shown in Fig. 7, most knowledge bits experience relatively little diffusion, with less than 5% of users learning about the given knowledge bit within the first day of it being posted. Conversely, about 10% of the knowledge bits experience widespread diffusion, with more than half of the users becoming aware of the knowledge within the first 24 hours. We see that most of the knowledge bits in the upper 50^{th} percentile experience significant diffusion within the first day of being introduced. The rest either do not diffuse at all or experience a delay in reaching any significant portion of the users in the simulation. This delayed reach is not unexpected as a user with few followers may tweet about the knowledge bit first, causing the knowledge item to have little diffusion. Then, a while later, a popular user may tweet about the knowledge item, causing a rapid increase in the diffusion.

In terms of the S-shaped diffusion curve, we find that the moderately spreading content (i.e., knowledge that falls between the 50^{th} and 75^{th} percentiles) exhibits such behavior. While the most viral knowledge spreads rapidly to many users from the start, as might be expected when a popular user makes a tweet, the moderately spreading content has a somewhat slow initial diffusion and then starts to spread faster between 8 to 28 hours after being posted. Across all of the percentiles displayed in Fig. 7, we see that the period of rapid knowledge diffusion is followed by a leveling off of the percentage of users who are aware of the knowledge bit, consistent with S-shaped diffusion.

Compared to the knowledge diffusion produced by the Reddit model, the Twitter model displays faster and more widespread diffusion. While the top 10^{th} percentile of knowledge items in the Reddit model simulations only reached 10-35% of users within the first 24 hours of being posted, the top knowledge items in the Twitter simulations reached 40-60% of users. Furthermore, a greater share of the knowledge items in the Twitter simulations had widespread diffusion, with more than 25% reaching at least 20% users within the first day of being introduced, compared to less than 10% reaching the same proportion of users in the Reddit simulations. This faster and more widespread diffusion agrees with prior work that has found that tweets of news stories have much shorter inter-arrival times and longer lifespans than Reddit posts and comments about the same topics (Priya et al. 2019). In combination with the virality and diffusion pattern results, these findings help validate that the knowledge diffusion process implemented by the Twitter model fits with the real world.

5.2 Twitter moderation virtual experiments

We now present an example of how the Twitter model can simulate the impacts of different types of users and levels of moderation. For these simulations, we introduce two types of active spreaders. First are “good” agents who start out knowing all of the “true” knowledge items and have trust values of 1.0 for the true knowledge items and 0.0 for the “misinfo” knowledge items. They create posts and comments whenever they are active on the platform to spread the “true” knowledge. Second are “bad” agents who start out knowing all of the “misinfo” knowledge items and actively spread them. Their trust values are reversed with full trust of the “misinfo” items and zero trust of the “true” items. Both types of agents have fixed trust of their knowledge items, specified by setting their *can_receive_trust* attributes to “false”. We add 20 agents of each type to the simulations.

To combat the bad agents and demonstrate the model’s ability to simulate different levels and types of moderation, we vary the *removal probability* and *ban threshold* values. The removal probability impacts how long it takes tweets to be removed, and the ban threshold changes how many times a user’s content can be removed before they are banned from tweeting. The remaining moderation setting in the Twitter model, the *moderation threshold*, was kept constant at 0.5 for this experiment. For each combination of settings, we performed 100 trials and measured the “normal” users’ awareness and trust of the “true” and “misinfo” knowledge bits at the end of each trial.

As expected, introducing stricter moderation in the experiments resulted in less diffusion of the misinformation items among the “normal” users, see Fig. 8. With no moderation, the average percentage of users who knew a given piece of knowledge was about 15% for both “true” and “misinfo” knowledge items. This gradually decreases for the “misinfo” knowledge items as the probability of removal of tweets containing misinformation increases. Once the removal probability reaches 1.0, which means that tweets are immediately

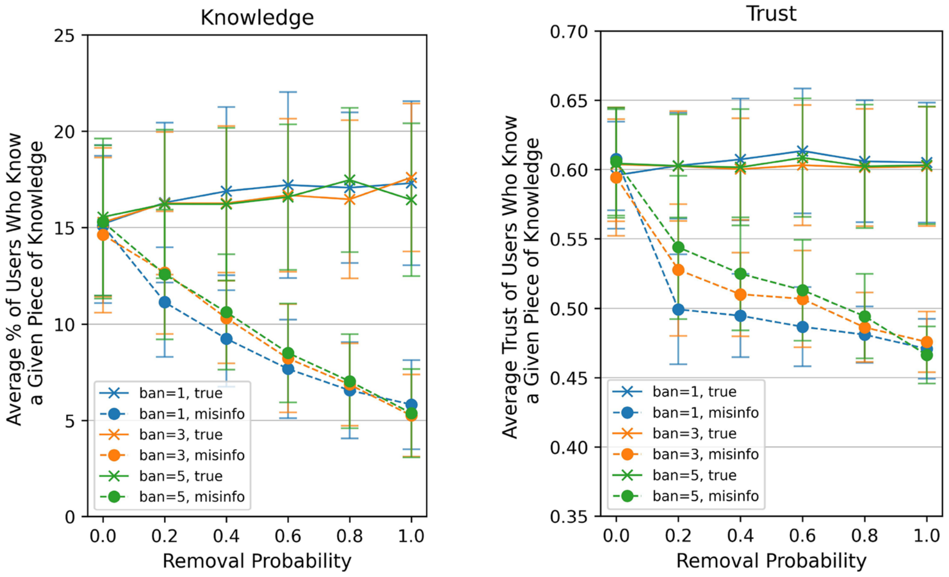

Fig. 8 Plots showing the impact of different moderation settings on the diffusion of knowledge items (left) and the trust of the knowledge items (right) recorded at the end of the simulations. In the figures, solid lines indicate awareness or trust of the "true" knowledge items and dashed lines represent "misinfo" items. The values are averaged across all of the trials for each setting with the standard deviation represented by the error bars. Increasing the removal probability decreased awareness of and trust in the misinformation. Meanwhile, increasing the ban threshold mostly affected only trust of the misinformation

removed if they contain misinformation and have trust values greater than 0.5, only about 6% of the users are aware of the misinformation at the end of the trials. This makes sense as it is only slightly greater than the 5% of users, on average, who begin the experiments with awareness of a given knowledge item.

In addition to impacting the number of users who learn about the misinformation, the removal probability also affects the users' trust of the misinformation, in the event that they become aware of it. Figure 8 shows that the average trust of the misinformation decreases as the removal probability increases. Since users are only active during certain timesteps, increasing the removal probability increases the chances of the tweets with high trust in misinformation being removed before the users see them. It follows then that users who find out about the misinformation knowledge items will be more likely to learn about them through tweets with trust values less than the moderation threshold, which is 0.5 for these experiments.

The ban threshold also affects the users' trust of the misinformation, especially when the removal probability is smaller. We see that when the removal probability is 0.2, the average trust of the misinformation is 0.54 when the ban threshold is 5 and 0.50 when the threshold is 1. When users are given more "strikes" before being banned from tweeting, they can make more tweets about the misinformation. With smaller removal probabilities, these posts survive longer, exacerbating their negative impact on impressionable users. A potential consequence of this is that platforms may be able to compensate for weaker content removal policies by enforcing stricter user bans or suspension thresholds. The reverse also holds and may be a source for future studies.

6 Cross-platform model results

With an understanding of how the moderation mechanisms impact misinformation diffusion on each platform individually, we now turn to the cross-platform model results. For the experiments with this model, we are interested in measuring the cross-platform effects of moderation. We use four moderation settings and three starting scenarios to conduct the experiments. The four moderation settings are: i) no moderation, ii) moderation on Twitter only, iii) moderation on Reddit only, and iv) moderation on both platforms. The three starting scenarios are: a) no active spreaders, b) 25 good and 25 bad agents on Reddit, and c) 25 good and 25 bad agents on Twitter. For each moderation and starting scenario combination, we perform 75 trials with 1,000 agents for 4,032 time steps.

6.1 Results without multi-platform normal users

We first conduct the experiments with disjoint sets of normal users for each platform. This acts as a control setting where the only source of cross-platform information diffusion is users who experience moderation on one of the platforms. When this happens, the *platform switch probability* parameter of the cross-platform model causes the users to gradually migrate to the alternative platform. Over time, we would expect these users to increase the amount of misinformation on the non-moderating platform, especially in cases where there are no good actors to counter the misinformation on the non-moderating platform.

The results of the cross-platform simulations demonstrate this pattern. In the trials where moderation is enacted on Reddit, the share of tweets containing "misinfo" knowledge

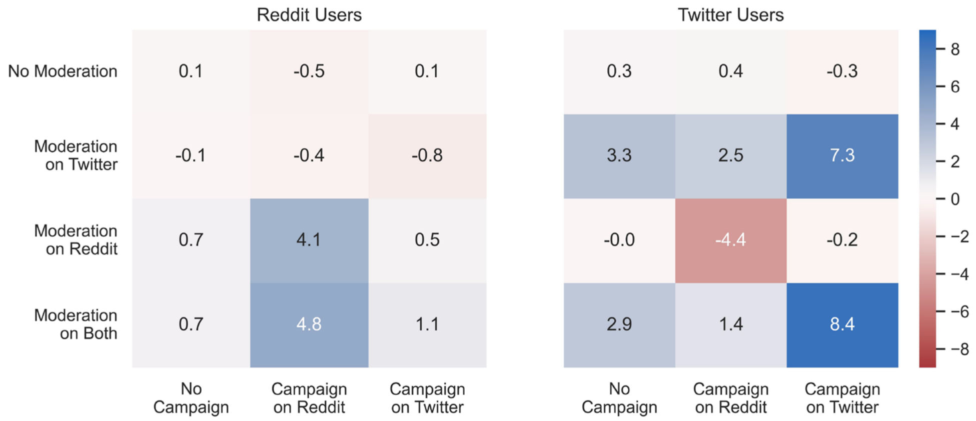

Fig. 9 Heatmaps reflecting the difference between the average percentage of normal users on Reddit (left) and Twitter (right) who know about a random "true" knowledge item versus a random "misinfo" knowledge item. The x-axis specifies the starting platform of the active spreaders and the y-axis indicates the moderation status of the platforms. Blue reflects a greater awareness of true information and red shows greater awareness of misinformation

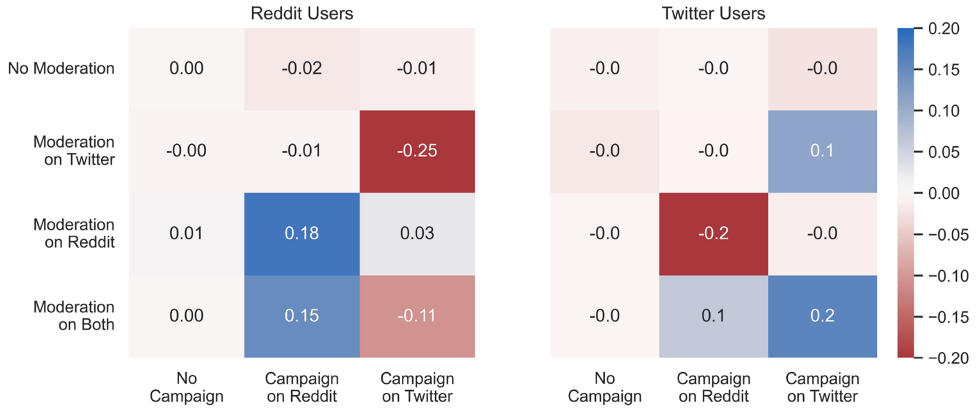

Fig. 10 Heatmaps displaying the difference between the normal users' average trust of true and misinformation knowledge items on Reddit (left) and Twitter (right). The x-axis specifies the starting platform of the active spreaders and the y-axis indicates the moderation status of the platforms. Blue represents a greater trust of true knowledge items than the misinformation knowledge items and red indicates higher relative trust in the misinformation than the true information

items is 55.9% when the bad actors start on Twitter and 55.8% when the bad actors start on Reddit. This is far more than the 28.9% and 33.6% of tweets that contain misinformation, respectively, when moderation is implemented on Twitter. The percentages of misinformation tweets in the trials when moderation is implemented on Reddit are also greater than the baseline of no moderation on either platform. The Reddit posts and comments followed similar trends to that of Twitter content.

The impact of moderation on misinformation diffusion can also be measured by the users' awareness of the true information and misinformation at the end of the simulations. As anticipated, we find that the combination of moderation on both platforms results in users being more aware of the true information than the misinformation, as shown in Fig. 9. However, we also find that the experiments with single-platform moderation resulted in the non-moderating platform mainly having worse outcomes than when neither platform implemented moderation. This effect was particularly strong when the bad actors started on Reddit and faced moderation on the platform. Their ensuing migration to Twitter ultimately resulted in a global increase in the awareness of the misinformation compared to no moderation on either platform.

The knowledge diffusion patterns from the individual platform models provide an explanation for these results. From the individual platform experiments, we observed that knowledge items tended to diffuse more quickly and to a larger share of the users on Twitter than on Reddit. This means that moderation performed on Twitter that encouraged the "bad" actors to migrate to Reddit allowed the "true" information to spread more effectively among the Twitter users and resulted in the Reddit users being less affected by the bad actors than the Twitter users would have been. An important note, however, is that while the Reddit users were less affected by the bad actors migrating to the platform than the Twitter users were in the opposite scenario, they still experienced an increase in awareness of the misinformation compared to when moderation was not implemented on Twitter. Therefore, recommendations regarding moderation and its impact on users depend on the users we are concerned about protecting.

The different combinations of moderation and active spreaders also impacted the users' average trust in the misinformation. Figure 10 shows that when the active spreaders started on Reddit, moderation on the platform increased the relative trust of the true information. Yet,

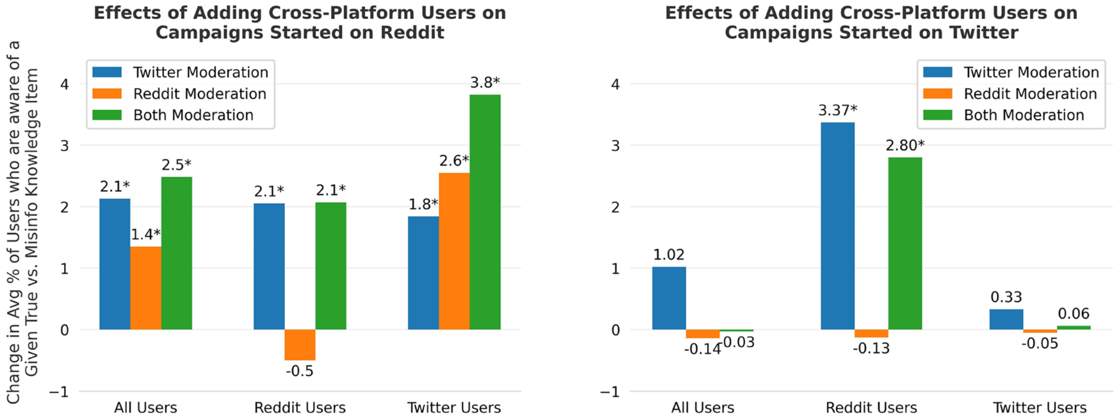

Fig. 11 Plots showing the impact of adding multi-platform “normal” users into the simulations. The bars show the change in the average percentage of users who are aware of a random true knowledge item versus a misinformation one compared to the baseline when there are no multi-platform users. The left plot shows the changes when the active users started on Reddit and the right one shows when they started on Twitter. (*) indicates changes that are statistically significant, using a Welch’s t-test and $\alpha = 0.05$

when the active spreaders began on Twitter and experienced moderation on the platform, there was an even larger decrease in the relative trust of the true information among the Reddit users. Similarly, on Twitter, there was a larger decrease in the relative trust of true information when the active spreaders started on Reddit and migrated to Twitter than the increase in trust that was gained when the spreaders started on Twitter and faced moderation. Although these trends relate to the trust of information, they are reflective of the real-world findings that users can increase the toxicity of alternative platforms when they migrate due to censorship and deplatforming (Buntain et al. 2023; Ali et al. 2021). In such cases, similar to the findings regarding the awareness of the misinformation, the relative sizes of the audiences on the moderating and alternative platforms should be considered to determine the net benefit of the overall system.

6.2 Impact of adding multi-platform normal users

In addition to simulating the impact of bad actors migrating across platforms, we are interested in the effects of normal users participating on both platforms. By being active on multiple platforms, the users can update their knowledge and beliefs based on viewing posts from either platform. To investigate this, we select 50% of the “normal” users (i.e., not active spreaders) to log on to both platforms. They serve as pathways for information and trust to spread between Twitter and Reddit. For these experiments, we set the users’ activity preferences on the platforms independently. This means it is possible to have users with different engagement styles on each platform, such as users who actively post on Reddit but only read on Twitter. We measure the effects of these users on the diffusion of knowledge and the users’ average trust values by adding them to the simulations where the active spreaders are present.

In general, we observe that adding the multi-platform users improves the users’ relative awareness of the true knowledge items, see Fig 11. When the active spreaders start out on Reddit, there is a statistically significant increase in the difference between the average share of users who know about a given true knowledge item versus a misinformation one, regardless of the moderation setting. This is especially true for the users of Twitter, a subset of whom can now receive messages from the “good actors” on Reddit and spread that information and trust to their peers. The only setting in which the active spreaders start on Reddit and the presence of the multi-platform users increases the relative awareness of the misinformation is among Reddit users when Reddit implements moderation. In this case, the fact that the “bad actors” can migrate to Twitter means that they can now continue to influence the users who are active on both platforms.

When the active spreaders start on Twitter, we similarly find that adding multi-platform users helps the Reddit users in the instances when Twitter performs moderation. Rather than the Reddit users being isolated and primarily influenced by the “bad actors” who migrate to Reddit, some can view the true information on Twitter and benefit from Twitter’s moderating activity, where the “good actors” are likely to experience more successful information diffusion. These results illustrate how having overlapping audiences between the platforms can potentially make it harder to cut off the influence of bad actors within a given platform while also making it possible to facilitate the reach of positive actors and diverse viewpoints. Thus, it may amplify the impact of moderation of platforms with greater reach and faster diffusion properties. In additional experiments in which we vary the level of multi-platform users, we find that the cross-platform trends are observed with even just 20% of the users being active on both platforms. This further highlights the importance of studying the impacts of moderation in a cross-platform context, as it can significantly impact the outcomes of different moderation strategies.

As for the multi-platform users' impact on trust, we find trends similar to their impact on knowledge diffusion. They tend to have a positive impact on the trust of the non-moderating platform's users, especially when the bad actors migrate from Reddit to Twitter. In that case, the presence of the multi-platform users results in the difference between the average trust of true knowledge items and "misinfo" items increasing by 0.15. By being able to use Reddit, they are exposed to the good agents' trust of the true items and distrust of the misinformation. Conversely, in the same situation, the difference in average trust between true items and misinformation items among Reddit users decreases by 0.05. Same as the knowledge diffusion case, they are better off when none of them have connections to Twitter, where the bad agents migrated. Interestingly, when the bad actors start on Reddit, and moderation is enforced only on Twitter, the average trust of both the Reddit users and Twitter users worsens by a statistically significant amount compared to when there were no multi-platform users. This is not the case when the bad actors start on Twitter, and moderation is implemented only on Reddit. There is no statistically significant change in the users' trust levels in this situation. Further and more detailed experiments with this model could investigate relationships like this one, aiming to improve trust in true information and distrust of misinformation for the entire ecosystem.

7 Limitations

By combining the independent models into a cross-platform model, this work provides a tool for simulating the cross-platform flow of information on social media and the complex impacts of moderation efforts. While the individual platform models incorporate the key structural features of each platform, their main limitations relate to their abstraction of user behavior and post content. First, they do not model within-session behavior changes or potential responses to information overload. Prior work has found that users' behavior can evolve over the duration of a single session such that users prefer simpler and easier actions as sessions become longer (Kooti et al. 2016). Other studies have indicated that when users receive information faster than they can process, their behavior regarding how much information they process and the sources they prioritize receiving information from can change (Rodriguez et al. 2014). Furthermore, information overload can reduce users' susceptibility to social contagions (Rodriguez et al. 2014). While the current versions of the models do not include these user behaviors, they provide a framework that can be modified to update user engagement attributes based on session length. The models can also be extended so that the updates to users' trust values account for the level of information they are exposed to during each session.

The individual platform models also contain some abstractions and simplification of their features. Although the Reddit model performs moderation in a distributed and heterogeneous manner, it performs threshold-based moderation. This simplification limits the model's ability to reflect the chance that misinformation goes undetected by the moderators, even when it meets their criteria for removal. Additionally, it results in the communities having their own general standards of user behavior but without the ability to have multiple rules, each with different potential consequences. Another limitation is that the Reddit and Twitter models only implement subscription and network-based recommendation systems to prioritize content in users' newsfeeds. Over the past decade, many platforms have shifted to algorithmic recommendation systems (Narayanan 2023). These systems learn user preferences and consider post attributes when ranking posts to show users. While such systems are not the focus of our work, the presented models allow for the implementation of alternative approaches to prioritizing content in the users' feeds through the Construct API. Consequently, they could be used to study the effects of different newsfeed algorithms on the spread of misinformation.

Since the cross-platform model is based on the two individual models, it inherits the limitations of those models. Additionally, the cross-platform model makes certain assumptions regarding user reactions to moderation and simplifies the full range of ways users could react to moderation. More specifically, we assume that users switch their future platform preferences with a fixed probability each time their content is removed or they are blocked from making a post. However, prior studies have found that user responses to moderation can differ based on how the moderation is performed and whether explanations are provided (Jhaver et al. 2019). Additionally, the position of the bad actors in the followership network could impact the effectiveness of user bans. If well-followed users are banned from the platform, it might have a greater impact on the within-platform diffusion but have negative cross-platform effects as the followers of the banned user migrate with them to an alternate platform. These behaviors are not implemented in the model or reflected in the simulations.

As for the experiments we performed, the knowledge items and trust values were randomly assigned to the users, excluding the good and bad actors. Consequently, our experiments did not reflect homophily between the users in the social networks. For example, on Reddit, we would expect to find subreddits where users share similar political beliefs or levels of awareness of particular knowledge items.

Similarly, users on Twitter tend to follow those with beliefs similar to their own. To study the effects of these properties on the spread of misinformation, future experiments could specify the input knowledge and knowledge trust networks in coordination with the subreddit membership or followership networks to connect users with similar beliefs.

8 Conclusion

Recent moves to restrict data access on platforms such as Twitter and Reddit heighten the benefits of having realistic agent-based models of social media environments. We present a cross-platform agent-based model of Twitter and Reddit interactions that facilitates simulations involving heterogeneous platforms, users, and moderation measures. The model incorporates two independent models of Twitter and Reddit interactions. The Reddit model includes the platform's key features, such as personalized news feeds, a community-based structure, and a decentralized moderation system. Meanwhile, the Twitter model facilitates followership-based content diffusion with realistic tweeting and newsfeed mechanisms. We show that the models independently produce results in alignment with real-world behaviors and can be customized to run specific experiments related to moderation and bad actors.

Although the presented experiments are limited to a few specific scenarios and moderation settings, we show that the model can replicate user migration to alternative platforms in response to moderation and reflect their subsequent impact on the alternative community's users. These initial experiments indicate that interventions taken on platforms that facilitate faster and more widespread diffusion may be more effective and have positive effects beyond the moderating platform if there is sufficient overlap in the user bases. Meanwhile, interventions taken on platforms that are more segmented or slower spreading that result in the bad actors migrating to platforms with faster diffusion can have worse outcomes for the whole system. The differences in the simulated misinformation diffusion based on the level of multi-platform usage highlight the importance of studying and modeling the cross-platform relationships between platforms when determining effective countermeasures for limiting the spread of misinformation.

By developing a general model that simulates cross-platform information diffusion, we provide a framework that can be used to study the cross-platform effects of specific user behaviors and platform features. Ample areas for further investigation include varying the levels of moderation on the platforms and introducing users with additional behaviors and interests to measure their impact on misinformation diffusion. The presented models can also be used to simulate diffusion across more platforms. Due to the design of the individual platform models, their properties can be set to simulate other types of community-based and followership-based platforms. The cross-platform model can also simulate users logging on to additional platforms, thus allowing for more complex diffusion patterns. Beyond the opportunity for further experimentation with the current models, future versions of the models will incorporate additional moderation measures and more user responses to improve their relevance to real-world interventions.

Acknowledgements The authors thank Dr. Stephen Dipple for his work developing the starting Twitter model available in the Construct API, as well as for his recommendations and assistance in working with the API and source code. This work was supported in part by the National Science Foundation through Grants #1813637 and #2225513, the Army Research Office through Grant #W911NF-22-1-0181, and the Defence Science and Technology Agency under the 'Development of models for information diffusion and combating disinformation' grant. Additional support was provided by the Knight Foundation through the CMU IDeaS Center.

Funding Open Access funding provided by Carnegie Mellon University.

Data availability Data collected for validation of the Twitter model can be made available through the corresponding author, in accordance to Twitter's terms and conditions and in compliance with the Carnegie Mellon University IRB. Access to the Reddit, Twitter, and cross-platform models will be available through the Construct API: https://github.com/CASOS-IDeaS-CMU/Construct-API. We recognize the ethical concerns of developing and publicly releasing a model that simulates misinformation diffusion over social media. However, the benefits of allowing other researchers to perform experiments and investigate countermeasures to limit such phenomena are substantial. The permitted uses of the models will be articulated in the models' user guide.

Declarations

Conflict of interest The authors declare that there are no conflicting interests.

References

Ali S, Saeed MH, Aldreabi E, et al (2021) Understanding the effect of deplatforming on social networks. In: Proceedings of the 13th ACM Web Science Conference 2021. Association for Computing Machinery, New York, NY, USA, WebSci '21, p 187–195, https://doi.org/10.1145/3447535.3462637

Allcott H, Gentzkow M, Yu C (2019) Trends in the diffusion of misinformation on social media. Res Polit 6(2):2053168019848554. https://doi.org/10.1177/2053168019848554

Antelmi A, Malandrino D, Scarano V (2019) Characterizing the behavioral evolution of twitter users and the truth behind the 90-9-1 rule. In: Companion Proceedings of The 2019 World Wide Web Conference. Association for Computing Machinery, New York, NY, USA, WWW '19, p 1035–1038, https://doi.org/10.1145/3308560.3316705

Auxier B, Anderson M (2021) Social media use in 2021. Tech rep Pew Res Cent 1(1):1–4

Averza A, Slhoub K, Bhattacharyya S (2022) Evaluating the influence of twitter bots via agent-based social simulation. IEEE Access 10:129394–129407. https://doi.org/10.1109/ACCESS.2022.3228258

Bakshy E, Hofman JM, Mason WA, et al (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, USA, WSDM '11, p 65–74, https://doi.org/10.1145/1935826.1935845

Beskow DM, Carley KM (2019) Agent based simulation of bot disinformation maneuvers in twitter. In: 2019 Winter Simulation Conference (WSC), pp 750–761, doi: https://doi.org/10.1109/WSC40007.2019.9004942

Blane JT, Moffitt JD, Carley KM (2021) Simulating social-cyber maneuvers to deter disinformation campaigns. In: Thomson R, Hussain MN, Dancy C, et al (eds) Social, Cultural, and Behavioral Modeling. Springer International Publishing, Cham, pp 153–163, https://doi.org/10.1007/978-3-030-80387-2_15

Buntain C, Innes M, Mitts T et al (2023) Cross-platform reactions to the post-january 6 deplatforming. J Quant Descr Digit Media. https://doi.org/10.51685/jqd.2023.004

Cabrera B, Ross B, Röchert D et al (2021) The influence of community structure on opinion expression: an agent-based model. J Bus Econ 91:1331–1355. https://doi.org/10.1007/s11573-021-01064-7

Carley KM (2020) Social cybersecurity: an emerging science. Comput Math Organ Theory 26(4):365–381. https://doi.org/10.1007/s10588-020-09322-9

Chapekis A, Smith A (2023) How u.s. adults on twitter use the site in the elon musk era. Tech. rep., Pew Research Center, https://www.pewresearch.org/short-reads/2023/05/17/how-us-adults-on-twitter-use-the-site-in-the-elon-musk-era/

Coscia M, Rossi L (2022) How minimizing conflicts could lead to polarization on social media: an agent-based model investigation. PLoS ONE 17(1):1–23. https://doi.org/10.1371/journal.pone.0263184

DiCarlo MF, Berglund EZ (2021) Connected communities improve hazard response: an agent-based model of social media behaviors during hurricanes. Sustain Cities Soc 69:102836. https://doi.org/10.1016/j.scs.2021.102836

Dipple S, Kowalchuck M, Altman N, et al (2022) Construct user guide 2023. Tech. Rep. CMU-ISR-22-102, Carnegie Mellon University, School of Computer Science, Inst. for Softw. Res. https://www.cmu.edu/casos-center/research/tools/cmu-s3d-23-104.pdf

Dixon S (2020) Frequency of twitter use in the united states as of 3rd quarter 2020. AudienceProject, https://www.statista.com/statistics/234245/twitter-usage-frequency-in-the-united-states/

Dixon S (2022) Reddit - statistics & facts. statista, https://www.statista.com/topics/5672/reddit/#topicOverview

Du E, Cai X, Sun Z et al (2017) Exploring the role of social media and individual behaviors in flood evacuation processes: an agent-based modeling approach. Water Resour Res 53(11):9164–9180. https://doi.org/10.1002/2017WR021192

Du J et al (2021) Using machine learning-based approaches for the detection and classification of human papillomavirus vaccine misinformation: infodemiology study of reddit discussions. J Med Internet Res 23(8):e26478. https://doi.org/10.2196/26478

English P (2016) Twitter's diffusion in sports journalism: role models, laggards and followers of the social media innovation. New Media Soc 18(3):484–501. https://doi.org/10.1177/1461444814544886

Fan R, Xu K, Zhao J (2018) An agent-based model for emotion contagion and competition in online social media. Phys A: Stat 495:245–259. https://doi.org/10.1016/j.physa.2017.12.086

Gatta VL, Luceri L, Fabbri F, et al (2023) The interconnected nature of online harm and moderation: Investigating the cross-platform spread of harmful content between youtube and twitter. In: Proceedings of the 34th ACM Conference on Hypertext and Social Media. Association for Computing Machinery, New York, NY, USA, HT '23, https://doi.org/10.1145/3603163.3609058

Gausen A, Luk W, Guo C (2022) Using agent-based modelling to evaluate the impact of algorithmic curation on social media. J Data Inf Qual 15(1):1–24. https://doi.org/10.1145/3546915

Golovchenko Y, Buntain C, Eady G et al (2020) Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 u.s. presidential election. Int J Press/Polit 25(3):357–389. https://doi.org/10.1177/1940161220912682

Gottfried J (2024) Americans' social media use. Tech. rep., Pew Research Center, https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/

Haddad H, Baral N, Garibay I (2021) Online rejection influence on behavior deviancy and radicalization: An agent-based model approach. In: Yang Z, von Briesen E (eds) Proc. 2020 Conf. of the Comput. Soc. Sci. Soc. of the Americas. Springer International Publishing, Cham, pp 15–29, https://doi.org/10.1007/978-3-030-83418-0_2

Hanley HWA, Kumar D, Durumeric Z (2023) Happenstance: Utilizing semantic search to track russian state media narratives about the russo-ukrainian war on reddit. In: Proc. Int. AAAI Conf. on Web and Social Media, pp 327–338, https://doi.org/10.1609/icwsm.v17i1.22149

Hunt K, Wang B, Zhuang J (2020) Misinformation debunking and cross-platform information sharing through twitter during hurricanes harvey and irma: a case study on shelters and id checks. Nat Hazards 103(1):861–883. https://doi.org/10.1007/s11069-020-04016-6

Jhaver S, Appling DS, Gilbert E, et al (2019) "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. In: Proceedings of the ACM on Human-Computer Interaction 3(CSCW):1–33. https://doi.org/10.1145/3359294,number: CSCW

Jhaver S, Boylston C, Yang D et al (2021) Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. Association for computing machinery, New York. https://doi.org/10.1145/3479525

Joseph K, Carley KM, Filonuk D et al (2014) Arab spring: from newspaper data to forecasting. Soc Netw Anal Min 4:1–7. https://doi.org/10.1007/s13278-014-0177-5

Kaligotla C, Yücesan E, Chick SE (2015) An agent based model of spread of competing rumors through online interactions on social media. In: Proc. 2015 Winter Sim. Conf. (WSC), pp 3985–3996, https://doi.org/10.1109/WSC.2015.7408553

Kim M, Newth D, Christen P (2013) Modeling dynamics of diffusion across heterogeneous social networks: news diffusion in social media. Entropy 15(10):4215–4242. https://doi.org/10.3390/e15104215

Kooti F, Moro E, Lerman K (2016) Twitter session analytics: profiling users' short-term behavioral changes. In: Spiro E, Ahn YY (eds) Social informatics. Springer, Cham, pp 71–86

Kumar N et al (2022) Covid-19 vaccine perceptions in the initial phases of us vaccine roll-out: an observational study on reddit. BMC Public Health 22:446. https://doi.org/10.1186/s12889-022-12824-7

Kwak H, Lee C, Park H, et al (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. ACM, Raleigh North Carolina USA, pp 591–600, https://doi.org/10.1145/1772690.1772751

Li M, Wang X, Gao K et al (2017) A survey on information diffusion in online social networks: models and methods. Information 8(4):118. https://doi.org/10.3390/info8040118

Liang H, Kw Fu (2015) Testing propositions derived from twitter studies: generalization and replication in computational social science. PLoS ONE 10(8):1–14. https://doi.org/10.1371/journal.pone.0134270

Lukito J (2020) Coordinating a multi-platform disinformation campaign: internet research agency activity on three us social media platforms, (2015) to 2017. Polit Commun 37(2):238–255. https://doi.org/10.1080/10584609.2019.1661889

McClain C, Widjaya R, Rivero G, et al. (2021) The behaviors and attitudes of us adults on twitter. Tech rep. Pew Research Center, https://www.pewresearch.org/internet/2021/11/15/the-behaviors-and-attitudes-of-u-s-adults-on-twitter/

Medvedev A, Lambiotte R, Delvenne J (2019) The anatomy of reddit: an overview of academic research. Springer, New York, pp 183–204. https://doi.org/10.1007/978-3-030-14683-2_9

Murdock I, Carley KM, Yağan O (2023) An agent-based model of reddit interactions and moderation. In: Proc. Int Conf. on Advances in Social Networks Analysis and Mining, https://doi.org/10.1145/3625007.3627489

Murić G, Tregubov A, Blythe J et al (2022) Large-scale agent-based simulations of online social networks. Auton Agents Multi-Agent Syst 36(2):38. https://doi.org/10.1007/s10458-022-09565-7

Myers SA, Sharma A, Gupta P, et al (2014) Information network or social network? the structure of the twitter follow graph. In: Proceedings of the 23rd International Conference on World Wide Web. Association for Computing Machinery, New York, NY, USA, WWW '14 Companion, p 493–498, https://doi.org/10.1145/2567948.2576939

Narayanan A (2023) Understanding Social Media Recommendation Algorithms. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

Newman MEJ (2009) Random graphs with clustering. Phys Rev Lett 103:058701. https://doi.org/10.1103/PhysRevLett.103.058701

Nimmo B, François C, Eib C et al (2020) Ira in ghana: Double deceit. Tech. rep, Graphika. https://graphika.com/reports/ira-in-ghana-double-deceit

Nishi R, Takaguchi T, Oka K et al (2016) Reply trees in Twitter: data analysis and branching process models. Soc Netw Anal Min 6(1):26. https://doi.org/10.1007/s13278-016-0334-0

Odabaş M (2022) 5 facts about twitter 'lurkers'. Tech. rep., Pew Research Center, https://www.pewresearch.org/short-reads/2022/03/16/5-facts-about-twitter-lurkers/

Papakyriakopoulos O, Medina Serrano J, Hegelich S (2020) The spread of covid-19 conspiracy theories on social media and the effect of content moderation. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-034

Priya S, Sequeira R, Chandra J et al (2019) Where should one get news updates: Twitter or reddit. Online Soc Netw Media 9:17–29. https://doi.org/10.1016/j.osnem.2018.11.001

Qiang Z, Pasiliao EL, Zheng QP (2019) Model-based learning of information diffusion in social media networks. Appl Netw Sci 4(1):111. https://doi.org/10.1007/s41109-019-0215-3

Rodriguez MG, Gummadi K, Schoelkopf B (2014) Quantifying information overload in social media and its impact on social contagions. In: Proceedings of the international AAAI conference on web and social media, pp 170–179

Salihefendic A (2015) How reddit ranking algorithms work. https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9

Sanderson Z, Brown MA, Bonneau R, et al (2021) Twitter flagged donald trump's tweets with election misinformation: They continued to spread both on and off the platform. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-77

Schreiber C, Carley KM (2013) Validating agent interactions in construct against empirical communication networks using the calibrated grounding technique. IEEE Int Conf Syst Man Cybern 43(1):208–214. https://doi.org/10.1109/TSMCA.2012.2192104

Serrano E, Iglesias CA, Garijo M (2015) A survey of twitter rumor spreading simulations. Comput Collectiv Intell 9329:113–122. https://doi.org/10.1007/978-3-319-24069-5_11

Singer P, Flöck F, Meinhart C, et al (2014) Evolution of reddit: From the front page of the internet to a self-referential community? In: Proc. 23rd Int. Conf. on World Wide Web. ACM, New York, NY, USA, WWW '14 Companion, p 517–522, https://doi.org/10.1145/2567948.2576943

Sobkowicz P, Sobkowicz A (2021) Agent based model of anti-vaccination movements: simulations and comparison with empirical data. Vaccines 9(8):809. https://doi.org/10.3390/vaccines9080809

Starbird K, Wilson T (2020) Cross-platform disinformation campaigns: lessons learned and next steps. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-002

Thukral S, Meisheri H, Kataria T, et al (2018) Analyzing behavioral trends in community driven discussion platforms like reddit. In: Proc. 2018 IEEE/ACM Int. Conf. on Adv. in Soc. Netw. Analysis and Mining (ASONAM), pp 662–669, https://doi.org/10.1109/ASONAM.2018.8508687

Tian Y, Yağan O (2022) Spreading processes with population heterogeneity over multi-layer networks. http://arxiv.org/abs/2211.07479

Velásquez N, Leahy R, Restrepo NJ et al (2021) Online hate network spreads malicious covid-19 content outside the control of individual social media platforms. Sci Rep 11:11549. https://doi.org/10.1038/s41598-021-89467-y

Weng L, Menczer F, Ahn YY (2013) Virality prediction and community structure in social networks. Sci Rep 3:1–6. https://doi.org/10.1038/srep02522

Xian J, Yang D, Pan L et al (2019) Misinformation spreading on correlated multiplex networks. Chaos: Interdiscip J Nonlinear Sci 29(11):113123. https://doi.org/10.1063/1.5121394

Yang Z, Yang C, Lu C et al (2023) Diffusion between groups: the influence of social brokers on content adoption in social networks. Eur J Mark 57(4):1039–1067. https://doi.org/10.1108/EJM-11-2020-0811

Yağan O, Qian D, Zhang J et al (2013) Conjoining speeds up information diffusion in overlaying social-physical networks. IEEE J Sel Areas Commun 31(6):1038–1048. https://doi.org/10.1109/JSAC.2013.130606

Zafarani R, Abbasi MA, Liu H (2014) Information diffusion in social media. Cambridge University Press, Cambridge, pp 179–214. https://doi.org/10.1017/CBO9781139088510.008

Zannettou S, Caulfield T, Setzer W, et al (2019) Who let the trolls out? towards understanding state-sponsored trolls. In: Proc. 10th ACM Conf. on Web Science. ACM, New York, NY, USA, WebSci '19, p 353–362, https://doi.org/10.1145/3292522.3326016

Zimdars M, Cullinan ME, Na K (2023) Alternative health groups on social media, misinformation, and the (de)stabilization of ontological security. New Media Soc. https://doi.org/10.1177/14614448221146171

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.