2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

An Agent-Based Model of Reddit Interactions and Moderation

Isabel Murdock
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
iem@andrew.cmu.edu

Kathleen M. Carley
Software and Societal Systems
Carnegie Mellon University
Pittsburgh, USA
kathleen.carley@cs.cmu.edu

Osman Yağan
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
oyagan@ece.cmu.edu

***Abstract*—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, Reddit provides a unique environment for studying the diffusion of knowledge and beliefs over social media. While assortativity, polarization, and user behavior have been examined within empirical contexts, having the ability to model the impacts of different moderation policies and rules across communities could provide useful insights for limiting the spread of misinformation. In this work, we introduce an agent-based model of Reddit interactions and moderating actions. By simulating interactions at the user level and specifying user-specific attributes, our model allows practitioners to conduct experiments with various types of actors and moderators and study their potential impact on Reddit-facilitated discussions and information diffusion. Additionally, subreddit-specific attributes enable communities to have different standards and thresholds for user conduct. To validate this model, we rely on an empirical dataset of over 100K posts and 800K comments across three U.S. political events in addition to user surveys and studies.**

***Index Terms*—Reddit, agent-based model, moderation, misinformation**

I. 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. While many studies regarding misinformation diffusion have focused on mainstream platforms, such as Twitter and Facebook, increasing attention has been paid to alternative platforms, including Reddit. This includes investigations into how the Internet Research Agency used Reddit to heighten political tensions during the 2016 U.S. presidential election and did so in a coordinated manner with actions taken on other social media platforms [1]–[3]. Recent work has also 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 [4]. In addition to political misinformation, Reddit has also been a host of "alternative health" communities and a forum for sharing health-related misinformation [5]–[7].

Although Reddit serves a largely beneficial function by allowing users to connect with others around common interests and share useful information, 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 ranks in the top 10 most used social network websites [8]. While this highlights the relevance of Reddit, especially within the U.S., the work presented in this paper focuses more on the community-based structure of the platform and the decentralized nature of moderation rather than specific features of Reddit. Consequently, it can be generalized to other platforms with similar designs.

The recent protests in reaction to Reddit's decision to increase API fees have drawn attention to the platform's decentralized moderation system. The protests demonstrated the influence and control the unpaid volunteer moderators who manage the individual communities on Reddit have on the platform [9], [10]. This results in part from the decentralized structure of Reddit, as well as much of the platform's moderation being delegated to the community moderators rather than handled centrally by the platform.

This paper introduces an agent-based model of user interactions on Reddit. It incorporates the platform's community-based structure and decentralized moderation system. Through tuning various user, moderator, and subreddit attributes, practitioners can use the model to study how different user behaviors, moderation efforts, and community standards may impact information diffusion on the platform. Validation of the model is performed using three datasets from U.S. political events and previous user surveys and studies.

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 and funding to attend this conference was provided in part by the CMU GSA/Provost Conference Funding.

ASONAM '23, November 6–9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11.
https://doi.org/10.1145/3625007.3627489

II. Background

A. 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 [11]. 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 [11], [12].

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 [11]. 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 [13]. In another work, neural networks were used to learn relationships between users within the context of linear threshold and random walk-based models [14].

Although these predictive models can yield high-fidelity simulations of the 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 [15], the adoption of competing rumors in a social media network [16], and polarization [17]. Additionally, ABMs have been used to model social media behavior and communications during natural disasters [18], [19] and health-related events [20].

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 model uses a user-to-community network structure, and posts are shared indirectly between users through shared 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. It also sets our model apart from previous work that has examined how different community structures lead users to *self-censor* [21] and how online rejection can make users vulnerable to *radicalization* [22]. Due to this community-based design, we are contributing to the emerging field of social cybersecurity by developing a model that can be used to study how influence campaigns conducted on decentralized platforms can be mitigated [23], [24].

B. Reddit Structure and Dynamics

As discussed, 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 [25]. 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 then posting [26]. 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 [27].

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 [27]. 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.

Drawing from these known trends, we aim to present a more realistic environment of Reddit interactions than previously developed by modeling users with different activity levels and propensities towards 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. We validate our model by comparing it to our own Reddit datasets, in addition to user behavior trends identified by previous studies. We then perform experiments involving different moderation approaches and subreddit rules.

The rest of the paper is organized as follows: Section III introduces the model by first providing an overview of its main components and then describing in detail how the agents interact with the Reddit environment. Section III ends with a discussion of the approaches used to validate the model. Section IV contains a description of the simulations performed, the validation results, and the outcomes from experiments

with heterogeneous subreddits and types of users. The paper concludes with a discussion of the key takeaways in section V.

III. Reddit Model

A. Model Framework

Our Reddit model is built on top of the Construct API[1]. Construct is an agent-based dynamic network framework that models agents' knowledge, beliefs, and evolution through interactions with other users [28]. 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 [29] [30]. Since the Construct API provides baseline classes and network management functions, it is a useful framework for creating social media-based models.

The Reddit model is a discrete-time model that simulates users logging onto the platform and viewing posts in their news feeds. Users then 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 and comments 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. The Reddit model has two main types of agents: *users* and *moderators*. Each agent has its own properties that determine when the user is active and which actions they take. The full set of attributes available is described in subsection III-D.

There are three main structures that are fundamental to the model: the *subreddit membership network*, the personalized *news feeds*, 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 [31].

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 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 knowledge 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.

B. User Behavior

As discussed, the model has two main types of agents: *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. We will describe how the user agents behave in the model first.

Before each simulation, user attributes and the user subscription network are specified; see subsection III-D. 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 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 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 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.

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*. The knowledge trust network is updated as:

$$\mathbf{T}_t(i, b) = (1 - update\_rate) * \mathbf{T}_{t-1}(i, b) + update\_rate * p$$

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

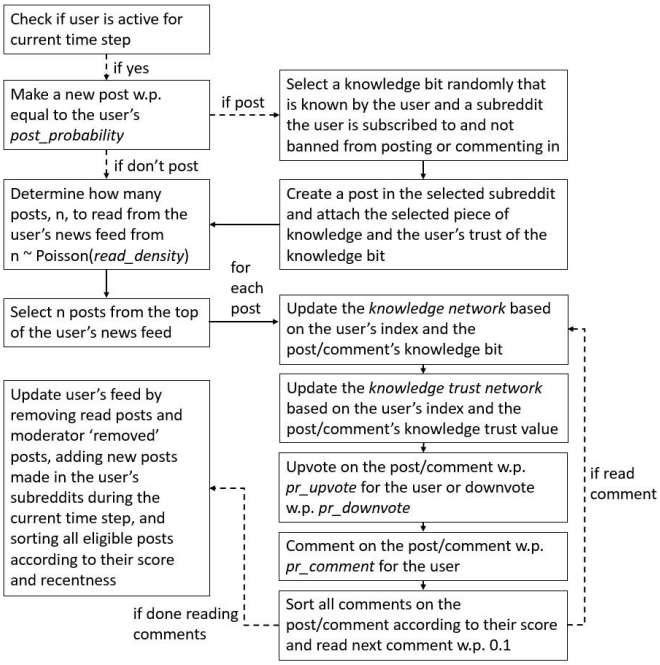

Fig. 1: User actions during each time step.

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 also introduce a *can receive trust* attribute, which can prevent certain users from updating their trust values when they view posts or comments. This feature makes it possible to simulate users who are convinced of their views and only aim to influence others within their subreddits.

C. Moderator Behavior

The second type of agent in the 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), as assigned 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

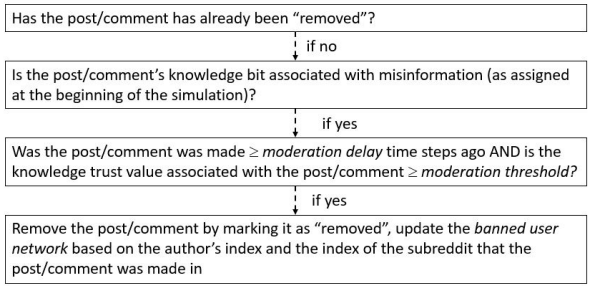

Fig. 2: Conditions that the moderators check before "removing" a post or comment from their subreddit(s).

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.

D. Model Inputs

Many of the features and attributes in our Reddit model have been covered in the previous discussion. For completeness, we present the full set of attributes in Table I. They must be initialized before the start of the simulation, along with the networks listed in Table II. They can be used to create heterogeneous agents with different activity levels, posting behaviors, and moderation styles. As discussed in the prior work, this is key for performing realistic simulations, as users on Reddit, as well as other social media platforms, vary widely in terms of activity levels and engagement types.

The networks listed in Table II 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*. Additionally, while the user attributes control the types of actions users take while they are active on the platform, the *user active time network* determines how frequently each user "logs on" to the platform.

TABLE I: Model Input Attributes

Attribute	Values	Description
User Attributes		
read_density	integer ≥ 1	average number of posts to read during an active time step
post_probability	float from 0-1	probability of making a post during an active time step
pr_upvote, pr_downvote, pr_comment	float from 0-1	probability of upvoting, downvoting, or commenting on a read post
can_receive_trust	boolean (true or false)	whether a user updates their knowledge trust when they read posts or comments
Moderator Attributes		
moderation delay	integer ≥ 0	number of time steps after a post is made that the moderator can remove the content
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 Attribute		
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 Attribute		
misinformation	boolean (true or false)	true indicates the knowledge item is fake information

TABLE II: Model Input Networks

Network	Link Type	Source → Target	Description
knowledge network	boolean	user → knowledge	true link values indicate the user is aware of the piece of knowledge
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)
user active time network	boolean	user → time step	true link values indicate that the user is active during the given time step
subreddit membership network	boolean	user → subreddit	true link values indicate that the user subscribed to the given subreddit

E. Validation Approach

Since most users on Reddit rarely post or comment on the platform, it is challenging to measure knowledge or information diffusion across the Reddit user base. Yet validation of the proposed Reddit model is critical for using it to draw meaningful conclusions. Due to this, we use a combination of input, face, and empirical validation to evaluate various aspects of our model.

First, we perform input validation of the values selected for the model attributes and networks of our experiments. This grounds our users' behaviors in previous user studies and helps ensure they follow real-world tendencies. We also draw our subreddit membership networks from empirical datasets to produce more realistic user-to-subreddit network structures.

We subsequently use 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 support 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 helps us evaluate whether the user behaviors and content diffusion produced by the model are consistent with the real world.

IV. Simulation Results

A. Validation

To evaluate the validity of our model, we first select realistic input parameters for the user attributes based on previous surveys, studies of user activity logs, and analysis of social media data. Each of these approaches has unique advantages and limitations. For example, while social media data analysis can be done at a large scale and give insights into broad populations of users, it is limited to uncovering behaviors tied to posting and commenting. Activity tracking can observe all user activity, such as the posts that users view and voting behaviors, but can have selection bias from users opting into the monitoring. Surveys can gather more information about beliefs but may have self-reporting errors. Therefore, by using these references in combination when selecting the inputs, we can create a more robust and realistic set of inputs.

Table III 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 time step in the simulation representing 5 minutes. We sample 1,000 users from one of the three Reddit datasets for each simulation run and extract their user-to-subreddit relationships. The users are then assigned attribute and network values according to Table III, with 100 knowledge items included in each simulation. After this initialization, each trial runs for 4,032 time steps, representing 2 weeks. We repeat this process 50 times for each of the three Reddit datasets for a total of 150 simulation runs.

For each run of the simulation, we gather all of the posts and comments made by the users. This allows us to collect a dataset similar to the real-world data we collected from Reddit. 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 for each simulation.

1) Temporal Patterns: The first analysis we perform relates to the temporal patterns of the users' posts and comments. Since we draw on previously observed usage patterns of Reddit users when specifying the *user active time network*, we would expect to find that the posting and commenting behaviors resulting from the active user sessions follow a similar diurnal pattern in our simulations.

To compare the pattern of discussions from the simulations to our three empirical datasets, we plot the power spectral density of the number of posts and comments made over time for each run. From this, we find that the strongest period is 24 hours across all of the simulations for each dataset. Comparing this to the posts and comments made in the actual datasets, we find that this behavior is consistent with the real-world data, see Fig. 3. In all cases, the daily frequency is much stronger than any others.

TABLE III: User-Related Input Parameters and References

Model Inputs	**Factors Considered (if applicable)**	**Values**	**References**
read_density, post_probability,	User activity levels	40% of users only read, 25% of users read and vote, 20% of users read, vote, and comment, 15% of users post, read, vote, and comment.	[25], [27], [32], [33]
pr_upvote, pr_downvote, pr_comment	Activity based on 5-minute time steps	Assuming the user can perform the given behavior: read_density=10, post_probability=0.1, pr_comment=0.1, pr_upvote=0.15, pr_downvote=0.05.	[33]
user active time network	Frequency of use	20% of users twice daily, 30% daily, 10% twice weekly, 20% weekly, 20% monthly.	[34]
	Time of use	Probability that users 'log on' between: 12am-6am: 0.1, 6am-9am: 0.1, 9am-12am: 0.8.	[35]
	Duration of session	50% of users are short-browsers (p=0.9), 25% are medium-browsers (p=0.2), 25% are long-browsers (p=0.05). For each time the user 'logs on', the number of time steps the user is active for is drawn from geometric(p).	[8], [33]
subreddit membership network		Sampled from the collected Reddit datasets.	
knowledge network		Each user has a 0.005 probability of starting with a given knowledge item (i.e., each knowledge item starts out being known by an average of 5 users).[a]	
knowledge trust network		Edge weights are drawn from uniform(0.1, 0.9) with default weights being 0.5.	

[a] Assuming simulations with 1,000 users.

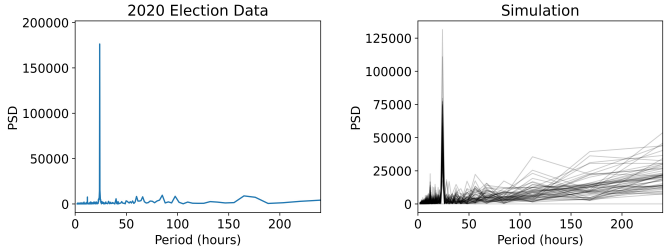

Fig. 3: Power spectral density plots of the posts made in the 2020 election dataset (left) and simulations (right). Each simulation is displayed as a line on the plot. The model reproduces the strong 24-hour period found across all three datasets in the corresponding simulations.

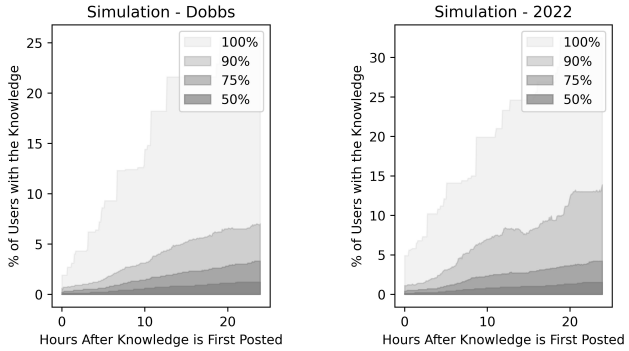

Fig. 4: Diffusion of knowledge bits in terms of the % of users who become aware of the knowledge in the first 24 hours after being posted. The diffusion of every posted knowledge bit is collected across every simulation. Temporal variations due to the *user active time network* are smoothed by taking a 6-hour moving average of each knowledge bit's diffusion. The resulting distributions for the Dobbs-related simulations (left) and 2022 election-related simulations (right) are displayed. Shading indicates the 50^{th}, 75^{th}, 90^{th}, and 100^{th} percentiles.

To further contextualize these results, we run Seasonal and Trend decomposition using Loess (STL) on the posting and commenting over time signals [36]. This allows us to extract their 'seasonal' components. We then average the seasonal signals across the weeks in the datasets or simulation runs to get a "representative" week of activity to compare across the empirical and simulation data. While the simulated agents generally have similar daily posting activity levels as the real-world users, we find that the real-world users had more gradual transitions between "night time" and "daytime" posting and commenting activities than the agents. This difference likely results from the discrete *times of use* considered when building the *user active time network*, see Table III.

Future work could explore sampling the user sessions specified by the *user active time network* directly from the posting signals found in the empirical datasets. However, our results from the power spectral density and STL analysis indicate that the 'log on' activity patterns specified by our inputs, along with the posting and commenting parameters, translated into posts and comments being created with temporal patterns similar to what has been observed in real life.

2) Knowledge Diffusion: As stated previously, models of information diffusion over networks have taken many forms. One common behavior exhibited by many of these models [37], as well as found through empirical work [38], [39], is an S-shaped 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 content or news stories go viral, many never gain enough traction to spread widely. For example, of 66K URLs linking to news articles in the 3 Reddit datasets, 74% were only posted once. Meanwhile, about 3% were posted more than ten times each. We expect our Reddit model to produce similar content diffusion patterns.

To evaluate whether this is the case, we plot the knowledge diffusion of all the knowledge bits posted in the simulations. As shown in Fig. 4, most knowledge bits experience very little diffusion, with less than 3% 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 10% of the users becoming aware of the knowledge within the first 24 hours.

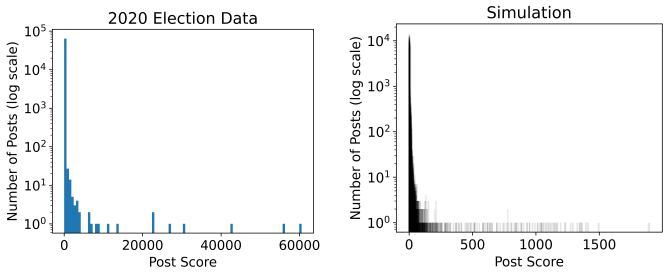

Fig. 5: Distribution of scores on posts in the 2020 election dataset (left) and simulations (right). The empirical dataset included more users than the simulations and consequently has a longer tail, with some posts receiving scores $>$ 10K.

This agrees with our expectations regarding viral content.

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 90^{th} percentiles) appears to exhibit such behavior. While the most viral knowledge spreads rapidly from the start, as might be expected with a viral news story, and the low-spreading content never spreads enough to experience the acceleration associated with the S-shape, the moderately spreading content has slow initial diffusion and then starts to spread faster around 5 hours after being posted. In combination, these findings help validate that the knowledge diffusion process implemented by the Reddit model fits with the real world.

3) Distribution of Post Scores: Turning to validation of the model's news feed and recommendation implementation, we compare the distributions of simulated post scores to our collected datasets. Since the algorithm prioritizes higher scoring and newer content, posts with larger scores are more likely to be seen and, therefore, continue to receive more attention. This positive feedback loop results in a few posts having very high scores. Meanwhile, the rest that does not gain such immediate attention drops lower in the news feeds due to both lower scores and being replaced by newer content.

We find that about 90% of the posts in the empirical datasets had a score of 1. This indicates that most of them did not receive votes from other users. However, as expected, we also find that a few posts garnered very positive scores, see Fig. 5. The simulated posts exhibit similar behavior, with most receiving scores of 1 and about 83% having a score of 3 or less. Although fewer users are simulated than were involved in the 2020 election dataset, the general trend of scores on posts is consistent with the outputs of the model.

B. Heterogeneous Users and Subreddits

We now present a short example of how the model can simulate the impacts of different types of users and moderation policies. For these simulations, we introduce two types of active spreaders. First are "good" agents who start out knowing all of the "true" knowledge bits and 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 bits and actively spread them.

TABLE IV: Knowledge Diffusion with Moderation

		Avg. % of users who know a given:	
Active spreaders?	**% of subreddits with moderation**	**"true" knowledge bit**	**"misinfo" knowledge bit**
No	0	5.0	4.9
No	100	5.4	2.8
Yes	0	9.7	9.4
Yes	25	9.8	7.2
Yes	50	8.9	5.7
Yes	100	9.1	5.8

Both types of agents have fixed trust of their knowledge bits, 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 our model's flexibility in assigning different moderation policies across subreddits, we also introduce moderators into varying percentages of the subreddits, as shown in Table IV. We then measure and compare the "normal" users' awareness of the "true" and "misinfo" knowledge bits at the end of each 14-day run. These simulations are run 50 times for each setting.

As expected, the results show that adding the active spreaders to the simulations resulted in greater diffusion of the knowledge bits among the "normal" users. Additionally, introducing moderators to the subreddits decreased the users' awareness of the "misinfo" knowledge. Interestingly, the benefit of adding moderators into the subreddits appears to plateau around 50%; see Table IV. Though further and more detailed experiments with this model are needed to investigate this relationship, this finding is promising as it suggests that targeted moderation efforts could have a significant impact across the platform.

V. 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 an agent-based model of Reddit interactions that facilitates simulations involving heterogeneous users, moderators, and subreddits. The model incorporates key features of Reddit, including personalized news feeds and a community-based structure. We show how the model produces results in alignment with real-world behaviors and can be customized to run specific experiments related to moderation and bad actors. The model is publicly available for use through the Construct API[2].

We recognize the ethical concerns of developing and publically 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. We articulate the permitted uses of the model in the model's user guide. Additionally, the Reddit datasets used in the analysis of this model were collected and analyzed with IRB approval.

[2]https://github.com/CASOS-IDeaS-CMU/Construct-API/tree/Reddit-public

ACKNOWLEDGMENT

The authors thank Dr. Stephen Dipple for his recommendations and assistance using the Construct API.

REFERENCES

[1] J. Lukito, "Coordinating a multi-platform disinformation campaign: Internet research agency activity on three u.s. social media platforms, 2015 to 2017," *Political Commun.*, vol. 37, no. 2, pp. 238–255, 2020, doi: 10.1080/10584609.2019.1661889.

[2] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn, "Who let the trolls out? towards understanding state-sponsored trolls," in *Proc. 10th ACM Conf. on Web Science*, ser. WebSci '19. New York, NY, USA: ACM, 2019, p. 353–362, doi: 10.1145/3292522.3326016.

[3] C. Sowa and K. M. Carley, "The russian strategy for interference and influence on reddit: An analysis of infiltrating new social media platforms," in *Proc. Int. Conf. SBP-BRiMS 2020*, Washington DC, USA, Oct. 2020.

[4] H. W. A. Hanley, D. Kumar, and Z. Durumeric, "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*, vol. 17, no. 1, Jun. 2023, pp. 327–338, doi: 10.1609/icwsm.v17i1.22149.

[5] M. Zimdars, M. E. Cullinan, and K. Na, "Alternative health groups on social media, misinformation, and the (de)stabilization of ontological security," *New Media Soc.*, 2023, doi: 10.1177/14614448221146171.

[6] J. Du *et al.*, "Using machine learning–based approaches for the detection and classification of human papillomavirus vaccine misinformation: Infodemiology study of reddit discussions," *J Med Internet Res*, vol. 23, no. 8, p. e26478, Aug 2021, doi: 10.2196/26478.

[7] N. Kumar *et al.*, "Covid-19 vaccine perceptions in the initial phases of us vaccine roll-out: an observational study on reddit," *BMC Public Health*, vol. 22, no. 446, 2022, doi: 10.1186/s12889-022-12824-7.

[8] S. Dixon, "Reddit - statistics & facts," statista, Sep 2022. [Online]. Available: https://www.statista.com/topics/5672/reddit/#topicOverview

[9] D. Ingram, "Reddit ceo slams protest leaders, saying he'll change rules that favor 'landed gentry'," NBC News, Jun 2023. [Online]. Available: https://www.nbcnews.com/tech/tech-news/reddit-protest-blackout-ceo-steve-huffman-moderators-rcna89544

[10] B. Collins, "Reddit strike one week on: A third of subreddits still down," Forbes, Jun 2023. [Online]. Available: https://www.forbes.com/sites/barrycollins/2023/06/21/reddit-strike-one-week-on-a-third-of-subreddits-still-down/

[11] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks: Models and methods," *Information*, vol. 8, no. 4, 2017, doi: 10.3390/info8040118.

[12] E. Serrano, C. A. Iglesias, and M. Garijo, "A survey of twitter rumor spreading simulations," *Computational Collective Intelligence*, vol. 9329, pp. 113–122, 2015, doi: 10.1007/978-3-319-24069-5_11.

[13] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," *Scientific Reports*, vol. 3, no. 2522, aug 2013, doi: 10.1038/srep02522.

[14] Z. Qiang, E. L. Pasiliao, and Q. P. Zheng, "Model-based learning of information diffusion in social media networks," *Appl. Netw. Sci.*, vol. 4, no. 1, Nov. 2019, doi: 10.1007/s41109-019-0215-3.

[15] R. Fan, K. Xu, and J. Zhao, "An agent-based model for emotion contagion and competition in online social media," *Phys. A: Stat.*, vol. 495, pp. 245–259, 2018, doi: 10.1016/j.physa.2017.12.086.

[16] C. Kaligotla, E. Yücesan, and S. E. Chick, "An agent based model of spread of competing rumors through online interactions on social media," in *Proc. 2015 Winter Sim. Conf. (WSC)*, 2015, pp. 3985–3996, doi: 10.1109/WSC.2015.7408553.

[17] M. Coscia and L. Rossi, "How minimizing conflicts could lead to polarization on social media: An agent-based model investigation," *PLOS ONE*, vol. 17, no. 1, pp. 1–23, 01 2022, doi: 10.1371/journal.pone.0263184.

[18] M. F. DiCarlo and E. Z. Berglund, "Connected communities improve hazard response: An agent-based model of social media behaviors during hurricanes," *Sustain. Cities Soc.*, vol. 69, p. 102836, 2021, doi: 10.1016/j.scs.2021.102836.

[19] E. Du, X. Cai, Z. Sun, and B. Minsker, "Exploring the role of social media and individual behaviors in flood evacuation processes: An agent-based modeling approach," *Water Resour. Res.*, vol. 53, no. 11, pp. 9164–9180, 2017, doi: 10.1002/2017WR021192.

[20] P. Sobkowicz and A. Sobkowicz, "Agent based model of anti-vaccination movements: Simulations and comparison with empirical data," *Vaccines*, vol. 9, no. 8, 2021, doi: 10.3390/vaccines9080809.

[21] B. Cabrera, B. Ross, D. Röchert, F. Brünker, and S. Stieglitz, "The influence of community structure on opinion expression: an agent-based model," *J. Bus. Econ*, vol. 91, p. 1331–1355, 11 2021, doi: 10.1007/s11573-021-01064-7.

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

[23] K. M. Carley, "Social cybersecurity: an emerging science," *Comput. Math Organ. Theory*, vol. 26, no. 4, pp. 365–381, 2020, doi: 10.1007/s10588-020-09322-9.

[24] Nat. Academies of Sci. Eng. and Med., *A Decadal Survey of the Social and Behavioral Sciences: A Research Agenda for Advancing Intelligence Analysis*. The Nat. Academies Press, 2019, ch. 6, doi: 10.17226/25335.

[25] A. Medvedev, R. Lambiotte, and J. Delvenne, "The anatomy of reddit: an overview of academic research," vol. DOOCN 2017. Springer, 2019, pp. 183–204, doi: 10.1007/978-3-030-14683-2_9.

[26] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier, "Evolution of reddit: From the front page of the internet to a self-referential community?" in *Proc. 23rd Int. Conf. on World Wide Web*, ser. WWW '14 Companion. New York, NY, USA: ACM, 2014, p. 517–522, doi: 10.1145/2567948.2576943.

[27] S. Thukral, H. Meisheri, T. Kataria, A. Agarwal, I. Verma, A. Chatterjee, and L. Dey, "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)*, 2018, pp. 662–669, doi: 10.1109/ASONAM.2018.8508687.

[28] S. Dipple, M. Kowalchuck, N. Altman, and K. M. Carley, "Construct user guide 2022," Carnegie Mellon University, School of Computer Science, Inst. for Softw. Res., Tech. Rep. CMU-ISR-22-102, 2022.

[29] C. Schreiber and K. M. Carley, "Validating agent interactions in construct against empirical communication networks using the calibrated grounding technique," *IEEE Int. Conf. Syst., Man, Cybern*, vol. 43, no. 1, pp. 208–214, 2013, doi: 10.1109/TSMCA.2012.2192104.

[30] K. Joseph, K. M. Carley, D. Filonuk, G. P. Morgan, and J. Pfeffer, "Arab spring: From newspaper data to forecasting," *Soc. Netw. Anal. Min.*, vol. 4, no. 177, 2014, doi: 10.1007/s13278-014-0177-5.

[31] A. Salihefendic, "How reddit ranking algorithms work," Dec 2015. [Online]. Available: https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9

[32] T. Bogers and R. Nordenhoff Wernersen, "How 'social' are social news sites? exploring the motivations for using reddit.com," in *Proc. iConference 2014*. iSchools, 2014, pp. 329–344, doi: 10.9776/14108.

[33] M. Glenski, C. Pennycuff, and T. Weninger, "Consumers and curators: Browsing and voting patterns on reddit," *IEEE Trans. Comput. Soc.*, vol. 4, no. 4, pp. 196–206, 2017, doi: 10.1109/TCSS.2017.2742242.

[34] S. Dixon, "Reddit usage frequency in the united states 2020," statista, Apr 2022. [Online]. Available: https://www.statista.com/statistics/815177/reddit-usage-frequency-usa/

[35] Y. Lin, "10 reddit statistics every marketer should know in 2023 [infographic]," Oct 2022. [Online]. Available: https://www.oberlo.com/blog/reddit-statistics

[36] R. B. Cleveland, W. S. Cleveland, and I. Terpenning, "Stl: A seasonal-trend decomposition procedure based on loess," *J. Off. Stat.*, vol. 6, no. 1, p. 3, 03 1990.

[37] R. Zafarani, M. A. Abbasi, and H. Liu, *Information Diffusion in Social Media*. Cambridge University Press, 2014, p. 179–214, doi: 10.1017/CBO9781139088510.008.

[38] P. English, "Twitter's diffusion in sports journalism: Role models, laggards and followers of the social media innovation," *New Media Soc.*, vol. 18, no. 3, pp. 484–501, 2016, doi: 10.1177/1461444814544886.

[39] Z. Yang, C. Yang, C. Lu, F. Wang, and W. Zhou, "Diffusion between groups: the influence of social brokers on content adoption in social networks," *Eur. J. Mark.*, vol. 57, no. 4, pp. 1039–1067, 2023, doi: 10.1108/EJM-11-2020-0811.