Simulating Toxicity Diffusion with Moderation-Aware SEIZ Models

Laura Pollacci[1] (✉), Letizia Milli[1], Giulio Rossetti[2], and Riccardo Guidotti[1]

[1] Department of Computer Science, University of Pisa, Italy, {name.surname}@unipi.it
[2] ISTI-CNR Pisa, Italy, giulio.rossetti@isti.cnr.it

Abstract. The spread of toxic content on online platforms motivates simulation tools for analyzing hypothetical intervention mechanisms under controlled assumptions. We present a simulation-based framework that extends the Susceptible–Exposed–Infected–Skeptic (SEIZ) epidemic model for toxic behavior diffusion in Online Social Networks. The framework combines a baseline SEIZ model for unmoderated diffusion with two moderation-aware variants: SEIZ-BasicModerator for uniform moderation and SEIZ-SmartModerator for message-driven and profile-aware moderation. In the baseline run, infection and exposure persist throughout the simulation. In the SEIZ-BM settings, stronger uniform moderation lowers the infection peak and shortens the time to elimination, but both moderated runs converge to an all-skeptic population. All SEIZ-SM runs end with non-zero infection, and the toxicity-score threshold shows the clearest association with late-stage infected counts. These results show that different diffusion and moderation rules can produce distinct state evolutions within the same SEIZ state space. The framework supports controlled what-if analysis of hypothetical moderation rules.

Keywords: Online Moderation · Online Toxicity · Epidemic Models.

1 Introduction

Online Social Networks (OSNs) are central spaces for public communication, political discussion, and civic participation. Their quality depends in part on how platforms handle toxic content, including harassment, aggressive language, hate speech, and other harmful interactions [15]. Toxicity can reduce the quality of discussion, discourage participation, and alter the structure of online communities [18]. For this reason, platform moderation affects the governance of online spaces. Moderation decisions influence what users see, how they interact, and whether online environments remain usable for public debate.

Most computational approaches to online toxicity focus on detection. They classify messages, users, or interaction patterns as toxic or non-toxic by using textual, behavioral, or network-based signals [2]. These methods are useful for identifying harmful content at scale, but they provide limited support for studying how toxicity evolves after exposure and how moderation changes this evolution. For moderation design, the relevant questions concern whether a message is toxic and how different intervention strategies affect the spread and persistence of toxic behavior over time.

Diffusion models provide a complementary perspective because they represent online behavior as a dynamic process over a networked population. The Susceptible–Exposed–Infected–Skeptic (SEIZ) model is useful because it distinguishes between susceptible, exposed, infected, and skeptic users [1]. This distinction captures three key phases of online toxicity, i.e., exposure to harmful content, adoption of toxic behavior, and resistance to its propagation. However, standard SEIZ formulations do not explicitly model platform moderation, and they often treat users as homogeneous agents. These assumptions limit their use for comparing moderation strategies.

Here we propose a moderation-aware extension of the SEIZ framework for simulating the diffusion of toxic behavior in OSNs. The framework includes three SEIZ-based models that share the same state space but implement different update mechanisms. The first is a baseline SEIZ model, which captures unmoderated diffusion through susceptible, exposed, infected, and skeptic states. The second is SEIZ-BasicModerator (SEIZ-BM), which adds a profile-agnostic moderation mechanism that can move infected users back to the susceptible state. The third is SEIZ-SmartModerator (SEIZ-SM), which introduces message-driven and profile-aware moderation based on user activity, message toxicity, and behavioral profiles. In SEIZ-SM, moderation is applied when an infected user's toxic-message count reaches a threshold, and the transition probabilities depend on the user's profile. The framework[3] supports controlled what-if analysis of unmoderated diffusion, uniform moderation, and message-driven, profile-aware moderation by comparing changes across SEIZ states over time. These changes reflect the specified rules and parameters and do not estimate real-platform policy effects. Building on this motivation, the paper contributes a SEIZ-based simulation framework for toxic behavior diffusion and moderation, two moderation-aware extensions implementing uniform and message-driven intervention rules, and a profile-aware formulation in SEIZ-SM that connects message toxicity, user activity, and moderated transitions.

The remainder of the paper is organized as follows. Section 2 reviews diffusion models, online toxicity, and moderation strategies. Section 3 presents the baseline SEIZ, SEIZ-BasicModerator, and SEIZ-SmartModerator models. Section 4 reports their agent-based evaluation on synthetic OSNs. Section 5 discusses the findings and limitations, while Section 6 concludes the paper.

2 Related Work

Toxic content in OSNs affects both users and communities. It can reduce the quality of discussion, discourage participation, and alter the structure of on-

[3] Code: `https://github.com/GiulioRossetti/Eval_Moderation_OSN.git`.

line interactions [18]. For this reason, toxicity has been studied from several perspectives, including content analysis, user modeling, and platform governance. A large part of the literature focuses on automatic detection. Existing approaches [2] classify toxic comments, cyberbullying, cyberviolence, or troll-like behavior from textual, behavioral, and user-level signals. Other work [15] stresses that toxicity detection cannot rely only on surface text, because context and background knowledge are often needed to interpret harmful content correctly. These studies provide useful tools for identifying toxic content, but they usually treat toxicity as a static classification target. They say whether a message or a user is toxic, but they do not model how toxic behavior spreads over time or how moderation changes this process.

This distinction is important for the study of online moderation. Moderation shapes how online conversations evolve after toxic content appears. It is also a governance mechanism that can change the evolution of a conversation, the behavior of users, and the stability of an online community. Recent work [4] has therefore framed moderation as an intervention problem, where the timing, form, and target of the intervention matter. Personalized interventions are especially relevant in this context, because users may react differently to the same moderation message. A generic warning may reduce toxic behavior for some users, but it may also be ineffective or counterproductive for others. This motivates simulation frameworks that can test moderation strategies before they are deployed on real platforms.

2.1 Diffusion Models for Online Behavior

Epidemiological models offer a compact and interpretable way to study diffusion processes. Classical compartmental models divide a population into discrete states and define transition rules among them. The SIR model [8], for example, represents individuals as Susceptible, Infected, or Recovered and models the spread of infection through state transitions. This modeling family was originally developed for infectious diseases, but it has also been used to study information diffusion, rumor spreading, and behavioral contagion in online environments [7].

Online behavior, however, is not always well represented by a simple binary distinction between adoption and non-adoption. Users can be exposed to a message without immediately sharing it, and they can also reject it after exposure. The SEIZ model [1] addresses this issue by adding two further states to the classical contagion structure, i.e., Exposed and Skeptic. The Exposed state captures users who have encountered a message but have not yet adopted the behavior. The Skeptic state captures users who resist or reject the message. This structure makes SEIZ particularly useful for modeling phenomena in which hesitation, resistance, and delayed adoption are central.

SEIZ and related models have been widely used to study rumor and misinformation diffusion in OSNs. The authors in [6] model the spread of news and rumors on Twitter through an epidemiological framework, while those in [16] study the effect of fact-checking on viral hoaxes and show how correction mechanisms can alter diffusion dynamics. Later work [17] extends this line by considering network segregation and the interaction between misinformation and fact-checking. Other studies [12] introduce dynamic SEIZ formulations for OSNs, where the diffusion process can change over time. Recent models [11] also study optimal control strategies for false information, further showing that intervention mechanisms can be integrated into diffusion models. These works show that epidemic-style models are useful for representing online diffusion in a transparent way. They make assumptions explicit, expose the role of model parameters, and allow researchers to compare alternative scenarios. However, most of this literature focuses on rumors, misinformation, or false information. Toxicity has received less attention as a dynamic behavioral process. Moderation is often represented indirectly, for example through fact-checking or generic control parameters, rather than as an explicit platform intervention applied to toxic users.

2.2 Toxicity as a Dynamic Contagion Process

A smaller body of work studies toxicity itself as a contagion process. In this line, epidemiological models have been used to study the spread of toxicity on YouTube, shifting the focus from misinformation to toxic behavior as a process that can propagate across users [13]. Related studies [9] apply epidemic models to online discussions during periods of high polarization, including the propagation of COVID-19-related toxicity on Twitter. Other work [3] uses epidemiological models to predict information spreading on Twitter, further supporting their use for online social dynamics. A comparison between SIR and SEIZ models for polarizing COVID-19 viewpoints also shows the relevance of models that can represent resistance and re-engagement dynamics [10].

The literature provides two relevant insights. First, toxicity can be studied as a population-level process rather than only as a property of isolated messages. Second, the structure of the online environment matters. Network homophily, social approval, and interaction patterns can affect the amplification of toxic behavior [5]. This suggests that moderation strategies should be evaluated through their system-level effects on toxicity diffusion, rather than only through their immediate effect on individual messages. At the same time, existing toxicity-diffusion models still leave an important gap. They usually model the spread of toxicity, but they do not explicitly compare different moderation strategies. In particular, they rarely distinguish between uniform moderation, where all users receive the same intervention, and profile-aware moderation, where the intervention depends on user behavior or user characteristics. This limits their usefulness for studying platform governance, because real moderation systems must decide when to intervene, whom to target, and how strongly to act.

2.3 User Heterogeneity and Moderation Strategies

A further limitation of many diffusion models is the assumption that users are homogeneous. In classical epidemic models, individuals usually share the same transition rules or differ only through network position. This abstraction is useful, but it is restrictive when the target phenomenon is toxic behavior. Users may differ in their tendency to produce toxic content, in their susceptibility to toxic exposure, and in their response to moderation. Ignoring these differences can make moderation models less realistic.

Psychological and behavioral parameters offer one way to represent user heterogeneity. We use the Dark Triad framework [14], which includes narcissism, Machiavellianism, and psychopathy, as one synthetic parameterization of socially aversive tendencies. The three components are not treated as clinical measurements or as validated predictors of online toxicity. Their role is to provide a structured source of variation in message classification and moderation responses. A generic latent behavioral propensity would be a possible alternative representation. This use of psychological profiling must be treated carefully because profile-based moderation raises issues of transparency, fairness, and potential misuse. Nevertheless, as a modeling assumption, it allows researchers to test how different user responses may affect moderation outcomes.

This work follows from this gap. The SEIZ structure provides a suitable starting point because it represents exposure, adoption, and resistance within a single interpretable framework [1]. We extend this structure by modeling moderation as an explicit intervention in the diffusion process and by introducing user-level heterogeneity to represent different responses to toxic exposure and corrective actions. The resulting framework connects epidemic-style diffusion, toxicity propagation, and moderation design in a common simulation setting.

3 Moderation-Aware SEIZ Models

We model the diffusion of toxic behavior in OSNs through three SEIZ-based agent models. All models use the same state space: Susceptible, Exposed, Infected, and Skeptic [1]. They differ, however, in their update mechanisms. The baseline SEIZ model represents unmoderated diffusion. SEIZ-SM introduces a message-driven moderation process in which diffusion and intervention depend on messages produced by infected users and treated as toxic.

In our setting, each node represents a user and edges represent potential exposure relations. At each simulation step, users can change state according to the transition rules of the selected model. The four states are defined as follows:

- S denotes *susceptible* users who have not adopted toxic behavior but can be exposed to it through their contacts;
- E denotes *exposed* users who have encountered toxic content but are not yet actively propagating it;
- I denotes *infected* users who can produce and propagate toxic content;
- Z denotes *skeptic* users who resist toxic content and do not propagate it.

These states describe behavioral conditions within the modeled diffusion process. In SEIZ-SM, the behavioral profile remains fixed, whereas the user's SEIZ state can change over time. The profile affects message classification and moderated transition probabilities, but it does not by itself determine whether a user

Table 1. Parameters of the baseline SEIZ model.

Param.	**Definition**
β	Contact rate between susceptible and infected users (S-I)
b	Contact rate between susceptible and skeptic users (S-Z)
ρ	Transition rate from exposed to infected ($E \rightarrow I$)
ε	Transition rate from infected to exposed ($I \rightarrow E$)
p	Probability that a susceptible user becomes infected after contact with I
$1-p$	Probability that a susceptible user becomes exposed after contact with I
l	Probability that a susceptible user becomes skeptic after contact with Z
$1-l$	Probability that a susceptible user becomes exposed after contact with Z
Δt	Simulation time step used for rate-to-probability conversion

is infected. The epidemiological analogy represents exposure, activation, persistence, and resistance in a networked process, rather than an equivalence between toxic behavior and biological contagion.

Table 1 shows the parameters used by the baseline SEIZ model. In the baseline implementation, contact and transition rates are converted into per-step probabilities through the simulation time step Δt. The parameter β controls contact between susceptible and infected users. After contact with an infected user, a susceptible user becomes infected with probability p or exposed with probability $1 - p$. The parameter b controls contact between susceptible and skeptic users. After contact with a skeptic user, a susceptible user becomes skeptic with probability l or exposed with probability $1-l$. The parameter ρ controls the transition from exposed to infected, while ε controls the transition from infected to exposed. Together, these transitions capture exposure, adoption, and resistance in the baseline dynamics.

3.1 Baseline SEIZ Model

The baseline model follows the SEIZ structure without platform intervention. A susceptible user who encounters toxic content can either adopt the toxic behavior directly or enter an exposed state. The exposed state represents a temporary condition in which the user has been influenced by toxic content but has not yet started propagating it. An exposed user can later become infected, while an infected user can also move back to the exposed state according to the internal transition rate ε. A susceptible user can become skeptic after contact with skeptic users. Skeptic users represent resistance to the propagation of toxic behavior.

Figure 1 summarizes the baseline SEIZ model. The model captures diffusion, exposure, and resistance in the absence of moderation. In the experimental section, this baseline is used as a reference for unmoderated diffusion, not as a direct competitor of the moderated variants.

3.2 SEIZ-BM: Basic Moderator

SEIZ-BasicModerator maintains the four SEIZ states and adds uniform moderation of infected users. Susceptible users follow the corresponding S transitions.

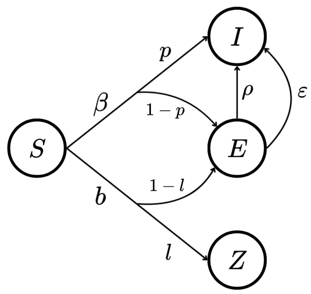


Fig. 1. Baseline SEIZ transitions among susceptible (S), exposed (E), infected (I), and skeptic (Z) users.

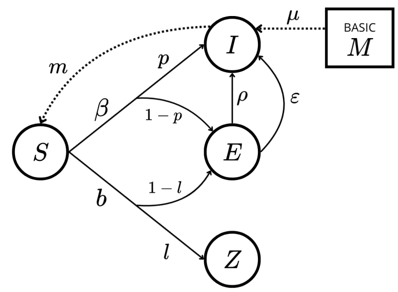


Fig. 2. SEIZ-BM introduces a profile-agnostic moderation transition from infected to susceptible users ($I \to S$), controlled by μ and m.

Table 2. Additional parameters introduced by SEIZ-BM.

Param.	**Definition**
μ	Probability of moderator intervention on an infected user
m	Probability of a moderated $I \to S$ transition

An exposed user becomes infected either through the internal transition with probability ε or, if it does not occur, after contact with an infected neighbor with probability ρ. Infected users may instead be moderated. Table 2 shows the additional parameters. In SEIZ-BM, moderation is profile-agnostic. At each simulation step, each infected user is selected for intervention with probability μ. If selected, the user returns from the infected state to the susceptible state with probability m. Otherwise, the user remains infected. The combined per-step probability of this transition is μ, m, separating intervention frequency from transition probability: $I \xrightarrow{\mu,m} S$. Figure 2 shows the SEIZ-BM extension. All infected users follow the same intervention rule, regardless of their behavioral profile or previous activity. The model therefore isolates how uniform moderation changes the infected population over time.

3.3 SEIZ-SM: Smart Moderator

SEIZ-SmartModerator introduces message-driven and profile-aware moderation. The model uses the same SEIZ state space, but diffusion is activated by messages treated as toxic rather than by every contact with an infected user. At each simulation step, n distinct users are randomly selected to generate messages. Users in any state may be selected as senders, but only infected users can produce messages treated as toxic. Messages produced by susceptible, exposed, or skeptic users are treated as non-toxic and do not contribute to diffusion.

Each user u is associated with a behavioral profile $\mathbf{p}_u = [p_1, p_2, p_3]$, where p_1, p_2, and p_3 represent narcissism, Machiavellianism, and psychopathy, respectively. The three components are independently sampled from a uniform distribution

Table 3. Additional parameters introduced by SEIZ-SM.

Param.	Definition
n	Number of users selected to generate a message at each step
θ	Moderation threshold for the toxic-message count
T	Toxicity-score threshold; higher values make classification more selective
η	Base probability of a moderated $I \to E$ transition
λ	Conditional probability of transition to Z from E or I

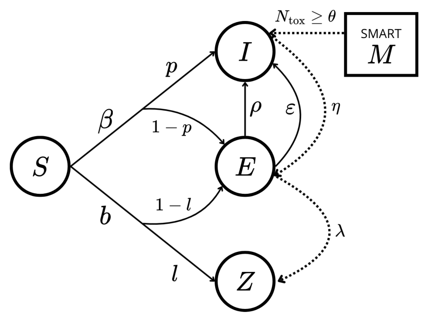


Fig. 3. SEIZ-SM model. Moderation is applied when an infected user's toxic-message count N_{tox} reaches θ; T defines which messages are treated as toxic.

over $[0, 1]$ at initialization and remain fixed throughout the simulation. They are used as synthetic behavioral parameters to introduce user heterogeneity, rather than as clinical labels or empirically estimated personality traits.

For each message produced by an infected user, the model assigns a profile-based toxicity score. Given a fixed user profile, this score is deterministic: messages produced by the same user receive the same score, while messages produced by users with different profiles may receive different scores. This choice ties message toxicity to fixed user-level assumptions, while moderation timing depends on how often the user produces messages treated as toxic and whether the toxic-message count reaches θ. The toxicity score is computed as the average of the three profile components. A message is treated as toxic if its score reaches T:

$$T_{message} = \frac{p_1 + p_2 + p_3}{3} \qquad T_{message} \geq T.$$

The arithmetic mean is used as a simple and transparent aggregation rule. This choice allows low values on two components to offset a high value on the third; maximum-based, weighted, or non-linear alternatives are left for future sensitivity analysis. The threshold T controls the selectivity of message classification. With lower values of T, more messages produced by infected users are treated as toxic; these messages activate diffusion and increase the toxic-message count. Higher values of T cause fewer messages to be treated as toxic. This may delay or prevent moderation because fewer messages contribute to N_{tox}. Only messages produced by infected users and treated as toxic trigger diffusion. When

Table 4. Fixed parameter values used in the experiments.

Model	β	b	ρ	p	ε	l	η	λ
Baseline SEIZ and SEIZ-BM	0.3	0.3	0.3	0.3	0.3	0.9	–	–
SEIZ-SM	0.3	–	0.3	0.3	0.3	–	0.6	0.2

an infected sender produces such a message, susceptible neighbors can become infected with probability p after contact, or exposed with probability $1-p$. Exposed neighbors can become infected with probability ρ. Exposed users become infected with probability ε; if this transition fails, they move to skeptic with probability λ or remain exposed.

Moderation is applied when an infected user's toxic-message count reaches the moderation threshold θ, i.e., when $N_{\text{tox}} \geq \theta$. The moderated transition probability depends on the user's behavioral profile. The effective probability of moving from infected to exposed is reduced for users with higher profile-based toxicity: $\eta_u = \eta\left(1 - T_{message}\right)$. After moderation, an infected user moves to exposed with probability η_u. If this transition fails, the user moves to skeptic with probability λ; if both transitions fail, the user remains infected:

$$I \xrightarrow{\eta_u} E, \qquad I \xrightarrow{(1-\eta_u)\lambda} Z, \qquad I \xrightarrow{(1-\eta_u)(1-\lambda)} I.$$

The SEIZ-SM rules represent moderation as a gradual transition away from infection rather than an immediate reset to susceptibility. Moderation may move a user to the exposed or skeptic state, while failed transitions leave the user infected. Exposed users may later return to infection or move to the skeptic state. Compared with SEIZ-BM, SEIZ-SM introduces three main changes. First, diffusion is activated by messages treated as toxic rather than by contact with infected users alone. Second, moderation is applied only after repeated production of such messages, with θ defining the required toxic-message count. Third, moderated transition probabilities depend on user-level behavioral profiles. Together, these rules represent message-driven and profile-aware moderation based on user activity and message toxicity.

3.4 Simulation Framework

The baseline model and the two moderation-aware variants are implemented as agent-based simulations over a networked population. They use the same SEIZ state space but different update mechanisms. The experiments therefore do not rank them through a single aggregate score. Instead, each model is evaluated through its state evolution under controlled parameter settings. The simulator records the number of users in each state at every time step, producing time series for S, E, I, and Z. Baseline SEIZ represents unmoderated diffusion, SEIZ-BM represents uniform moderation of infected users, and SEIZ-SM represents message-driven and profile-aware moderation.

4 Experiments

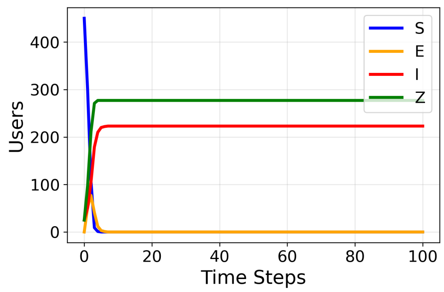
We evaluate a baseline SEIZ model and two moderation-aware variants through agent-based simulations on synthetic OSNs. They share the same state space but differ in their update rules (Section 3). Baseline SEIZ defines the unmoderated reference, while SEIZ-BM and SEIZ-SM are analyzed across moderation settings. All experiments use the same Erdős–Rényi network with 500 users and $p_{\mathrm{ER}} = 0.03$. The expected average degree is 14.97, while the generated network has 3,669 edges and an average degree of 14.68. This synthetic topology isolates the effects of model rules and parameters rather than reproducing an empirical OSN. Initially, the network contains 450 susceptible, 25 infected, and 25 skeptic users, with no exposed users. Each run spans 100 time steps and reports the number of users in each state over time. Fixed graph and simulation seeds preserve the network and initial states across configurations and make each run reproducible. Table 4 reports the fixed parameter values used in the experiments. A dash indicates that the parameter is not used. The value 0.3 is used consistently for the main contact and transition parameters to limit the number of varying factors. With $p = 0.3$, contact with an infected user leads to infection with probability 0.3 and exposure with probability 0.7. With $l = 0.9$, contact with a skeptic user leads to the skeptic state with probability 0.9 and to exposure with probability 0.1. In SEIZ-SM, $\eta = 0.6$ defines the base probability of a moderated $I \rightarrow E$ transition before profile adjustment, while $\lambda = 0.2$ defines the conditional probability of moving to Z. These values define a controlled synthetic setting and are not empirical estimates. Baseline SEIZ converts rates into per-step probabilities using $\Delta t = 1$, while SEIZ-BM and SEIZ-SM use the listed values directly in their update rules. In all plots, susceptible users are shown in blue, exposed users in orange, infected users in red, and skeptic users in green.

4.1 Baseline SEIZ Dynamics

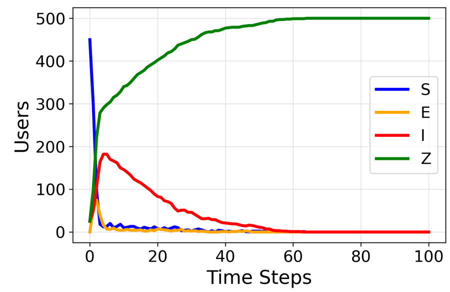
We first analyze baseline SEIZ as the reference for unmoderated diffusion. The susceptible population falls from 450 to zero within five steps, while the skeptic population rises from 25 to 294 and then remains constant. After this initial phase, infected and exposed users continue to move between the two states. From step 5 onward, the infected population ranges from 88 to 120 users and the exposed population from 86 to 118. The simulation ends with 102 infected and 104 exposed users. The continued exchange between infected and exposed users indicates persistent unmoderated diffusion, with approximately two fifths of the population remaining in these states. Figure 4(a) shows the within-model no-moderation reference for SEIZ-BM, obtained with $\mu = 0$ and $m = 0$.

4.2 SEIZ-BM Moderation Dynamics

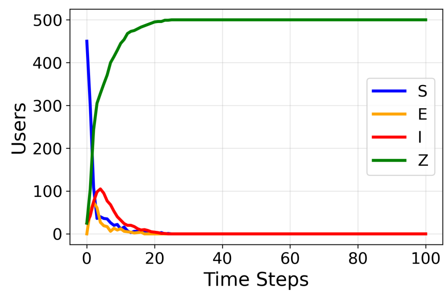
We vary the SEIZ-BM parameters μ and m, which respectively control intervention and the transition from infected to susceptible. Figure 4 compares disabled, moderate, and stronger uniform moderation. With $\mu = 0$ and $m = 0$ (panel



(a) $\mu = 0$, $m = 0$



(b) $\mu = 0.3$, $m = 0.3$



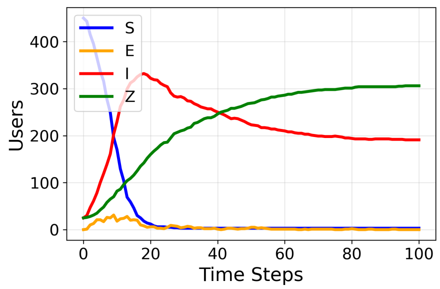
(c) $\mu = 0.6$, $m = 0.6$

Fig. 4. SEIZ-BM state evolution under no, moderate, and stronger uniform moderation. Panel (a) is the within-model reference.

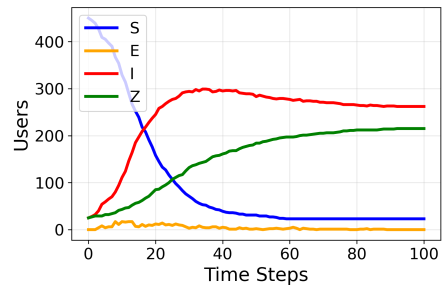
(a)), susceptible and exposed users disappear by step 7, after which the system remains at 223 infected and 277 skeptic users. The no-moderation run thus reaches a stable split, with 44.6% of the population remaining infected. With $\mu = 0.3$ and $m = 0.3$ (panel (b)), infection peaks at 182 users during steps 4–5 and declines to zero at step 63. Moderation removes infection, but only after a long decline. With $\mu = 0.6$ and $m = 0.6$ (panel (c)), the peak falls to 105 users and infection reaches zero by step 24. In the stronger setting, both early growth and infection duration are lower. Across these settings, increasing μ and m changes the system from persistent infection to elimination, but not to stable susceptible recovery. Moderated users return to susceptibility and re-enter the SEIZ dynamics; as the skeptic population grows, transitions toward skepticism become more likely. Both moderated runs therefore end with the full population in the skeptic state rather than an intermediate balance among states.

4.3 SEIZ-SM Message-Driven Moderation Dynamics

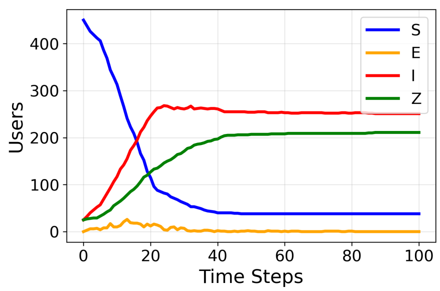
We evaluate SEIZ-SM by varying the number of users selected to generate messages n, the moderation threshold θ, and the toxicity-score threshold T. Unlike SEIZ-BM, SEIZ-SM is message-driven, and both diffusion and moderation depend on messages produced by infected users and treated as toxic. The parame-



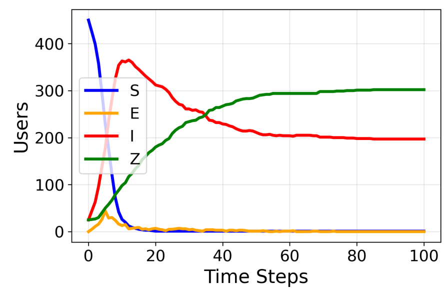
(a) $n = 150$, $\theta = 2$, $T = 0.5$



(b) $n = 150$, $\theta = 3$, $T = 0.6$



(c) $n = 250$, $\theta = 2$, $T = 0.6$



(d) $n = 250$, $\theta = 3$, $T = 0.5$

Fig. 5. SEIZ-SM state evolution for four representative settings of n, θ, and T.

ter n controls how many users are selected as message senders at each step. The parameter θ defines how many messages treated as toxic an infected user must produce before moderation is applied. The parameter T controls which messages are treated as toxic by the model. Lower values of T classify a larger share of messages as toxic.

Figure 5 shows four representative configurations. With $n = 150$, $\theta = 2$, and $T = 0.5$ (Figure 5(a)), the infected population reaches a maximum of 332 users at step 18. It then decreases throughout the remaining simulation and ends at 191 users. The skeptic population grows to 306 users, while the final susceptible and exposed populations contain 3 and 0 users, respectively. Infection decreases after the initial peak, but a large infected population remains active. With $n = 150$, $\theta = 3$, and $T = 0.6$ (Figure 5(b)), infection grows more gradually and reaches a maximum of 299 users at steps 34 and 35. It then decreases to 262 users at the end of the simulation. The final state contains 215 skeptic, 23 susceptible, and no exposed users. Compared with panel (a), the joint increase in θ and T is associated with a lower peak, a larger final infected population, and a smaller final skeptic population. With $n = 250$, $\theta = 2$, and $T = 0.6$ (Figure 5(c)), the infected population reaches a maximum of 268 users at step 24. It then decreases slightly and ends at 251 users. The final state also contains 211 skeptic and 38 susceptible users, while the exposed population reaches zero. This configuration produces the lowest peak among the four displayed runs, but infection remains

Table 5. Late-stage infected ranges and main outcomes for all SEIZ-SM configurations.

n	θ	T	I_{late}	**Main outcome**
150	2	0.5	191–193	Gradual decline toward a skeptic-dominant regime
150	2	0.6	245	Persistent infection under the stricter toxicity threshold
150	3	0.5	188–194	Lowest late infected range and skeptic-dominant regime
150	3	0.6	262–263	High persistence with delayed moderation
250	2	0.5	203	Persistent infection after a sharp early peak
250	2	0.6	251	Persistent infection with the largest susceptible remainder
250	3	0.5	197	Highest early peak followed by a marked decline
250	3	0.6	262	High persistence under the stricter toxicity threshold

active in approximately half of the population. Finally, with $n = 250$, $\theta = 3$, and $T = 0.5$ (Figure 5(d)), the infected population reaches 365 users at step 12, the highest peak among the displayed configurations. It subsequently decreases to 197 users. At the end of the simulation, 302 users are skeptic, one is susceptible, and none are exposed. This run combines a larger message volume and a lower toxicity-score threshold, which increase the number of messages that can activate diffusion, with a higher moderation threshold, which delays intervention. The resulting early infection peak is the highest among the four displayed runs.

Table 5 summarizes the late-stage infected ranges and the main pattern observed for all eight SEIZ-SM configurations. The reported range is computed over the last ten simulation steps, rather than from a single final value, to account for the small residual fluctuations that remain in some configurations. The clearest pattern concerns the toxicity threshold T. For every matched combination of n and θ, configurations with $T = 0.6$ contain more infected users in the late stage than those with $T = 0.5$. The difference ranges from about 48 to 75 users. At the same time, configurations with $T = 0.5$ end with a larger skeptic population. A higher T causes fewer messages to be treated as toxic, reducing both diffusion events and moderation triggers. In the reported runs, fewer moderation triggers are associated with greater late-stage infection. The moderation threshold θ mainly changes the transient phase. At fixed n and T, increasing θ is generally associated with a larger or later infection peak because moderation is applied only after more messages are treated as toxic. Its effect on the late-stage infected population is less regular: depending on n and T, the final difference is small or changes direction. The effect of n is also conditional on the other parameters. Increasing the number of selected senders changes the speed and height of the initial diffusion phase, but it does not produce a uniform increase in late-stage infection. Message volume, toxicity classification, and intervention timing must therefore be interpreted jointly. Each configuration uses one run, so no uncertainty is estimated. These results show that the shared SEIZ state space produces distinct regimes under different diffusion and moderation rules. In the tested SEIZ-BM configurations, infection elimination coincides with convergence of the full population to the skeptic state. SEIZ-SM instead ends with non-zero infection in every tested run, with late-stage levels most clearly differentiated by T. The next section discusses these regimes and their implications.

5 Discussion

The simulations indicate that the final infected count only partially describes the effects of the tested moderation rules. The infection peak, the timing of its decline, and the distribution across states provide additional information. In the baseline SEIZ run, susceptible users disappear rapidly, while infected and exposed users continue to move between the two states. The system therefore maintains active unmoderated diffusion throughout the simulation. The exposed population also remains substantial, so many users can still move to infection. These dynamics provide the reference for assessing how moderation changes infection and the overall state distribution.

Under SEIZ-BM, larger values of μ and m are associated with lower infection peaks and earlier elimination. However, infection removal does not produce a stable susceptible population. Moderated users return to susceptibility and re-enter the SEIZ dynamics. As the skeptic population grows, transitions toward skepticism become more likely under the model rules, and both moderated runs end with all users in the skeptic state. The observed change is therefore from persistent infection to complete convergence toward skepticism. This result depends on the specified transition rules and should not be interpreted as a general effect of uniform moderation. Moreover, because μ and m increase together, the experiments cannot attribute the differences to either parameter alone.

All tested SEIZ-SM configurations end with non-zero infection. The parameters affect the infection peak, the subsequent decline, and the late-stage infected population through their interaction with the message-based rules. Diffusion depends on messages treated as toxic, moderation is applied only after their repeated production, and moderated transitions depend on user profiles. These results describe how toxicity classification and profile-dependent intervention change infection dynamics. They do not establish that SEIZ-SM is more effective than SEIZ-BM, since the two models use different diffusion and transition rules. The toxicity-score threshold T shows the most consistent association with late-stage infection. For every matched combination of n and θ, runs with $T = 0.6$ contain more infected and fewer skeptic users near the end than runs with $T = 0.5$. A higher T reduces both the messages that activate diffusion and those that increase the toxic-message count. In the reported runs, fewer moderation triggers are associated with greater late-stage infection. The infected count must therefore be interpreted together with the model's toxicity rule.

The effects of the moderation threshold θ and message volume n are more evident during the transient phase. A higher θ delays intervention and often corresponds to a larger or later infection peak, although its association with late-stage infection is not consistent across configurations. Increasing n also changes the speed and size of early diffusion without producing a uniform late-stage effect. Message volume, toxicity classification, and intervention timing should therefore be considered jointly. Similar final counts may otherwise hide different infection peaks and rates of decline.

The findings remain conditional on the selected rules, parameters, network, and initial conditions. The experiments use one Erdős–Rényi network, one size and density, and one reproducible run per configuration, without uncertainty estimates. The models are not fitted to or validated on empirical OSN data. The infected state represents a modeled propensity to produce toxic messages, while the profile scores and their aggregation are simplified modeling choices. The results support controlled comparisons of hypothetical moderation mechanisms under explicit assumptions, but do not estimate their effects on real platforms.

6 Conclusion

This paper presents a SEIZ-based simulation framework for toxic behavior diffusion and moderation in OSNs. The framework combines a baseline model for unmoderated diffusion with two moderation-aware variants: SEIZ-BM for uniform intervention and SEIZ-SM for message-driven and profile-aware moderation.

The tested configurations show that different diffusion and intervention rules produce distinct system-level outcomes. Baseline SEIZ maintains non-zero infected and exposed populations. In SEIZ-BM, stronger uniform moderation lowers infection peaks and shortens the time to elimination, but the tested moderated runs move the full population to the skeptic state. In SEIZ-SM, infection remains present in all runs, while late-stage levels vary most consistently with the toxicity-score threshold. This result shows that message classification affects both toxic diffusion and the opportunities for moderation.

The central result is that the same SEIZ state space produces different outcomes depending on how diffusion and moderation are implemented. In the tested runs, infection elimination under SEIZ-BM coincides with convergence to skepticism, while SEIZ-SM maintains non-zero infection under message-driven and profile-aware intervention. Moderation mechanisms should therefore be interpreted through the full state evolution and the rules that generate it, rather than through the final infected count alone.

These conclusions apply to the tested synthetic settings and do not estimate how platform moderation affects toxicity diffusion. Future work will fit and validate the model using empirical OSN data. Repeated simulations will provide uncertainty estimates, while sensitivity analyses will vary network topology, size, density, and initial conditions. Further work will compare user-profile representations, aggregation rules, and moderation transitions.

Acknowledgments. This work has been partially supported by the Italian Project Fondo Italiano per la Scienza FIS00001966 “MIMOSA”, and by the European Union under the scheme HORIZON-INFRA-2025-01-DEV-02, Grant Agreement No. 101292277, “SoBigData IP: SoBigData Implementation Phase”.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bettencourt, L.M., et al.: The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. Physica A: Statistical Mechanics and its Applications **364**, 513–536 (2006)
2. Bonetti, A., et al.: Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. Applied Sciences **13**(10) (2023). https://doi.org/10.3390/app13106038
3. Castiello, M., et al.: Using epidemiological models to predict the spread of information on twitter. Algorithms **16**(8), 391 (2023)
4. Cresci, S., et al.: Personalized interventions for online moderation. In: Proceedings of the 33rd ACM Conference on Hypertext and Social Media. pp. 248–251 (2022)
5. Jiang, J., et al.: Social approval and network homophily as motivators of online toxicity. CoRR **abs/2310.07779** (2023). https://doi.org/10.48550/ARXIV.2310.07779
6. Jin, F., et al.: Epidemiological modeling of news and rumors on twitter. In: Proceedings of the 7th workshop on social network mining and analysis. pp. 1–9 (2013)
7. Keeling, M.J., Rohani, P.: Modeling infectious diseases in humans and animals. Princeton university press (2008)
8. Kermack, W.O., et al.: A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character **115**(772), 700–721 (1927)
9. Maleki, M., et al.: Applying an epidemiological model to evaluate the propagation of toxicity related to covid-19 on twitter. In: HICSS. pp. 1–10. ScholarSpace (2022)
10. Maleki, M., et al.: A comparative evaluation of the SIR and SEIZ epidemiological models to describe the diffusion characteristics of COVID-19 polarizing viewpoints on online social networks. In: HICSS 2025. pp. 1–10. ScholarSpace (2025). https://doi.org/10.24251/HICSS.2025.302
11. Mansal, F., Faye, I.: Mathematical modeling and optimal control of untrue information : Dynamic SEIZ in online social networks. CoRR **abs/2309.13058** (2023). https://doi.org/10.48550/ARXIV.2309.13058
12. Mathur, A., Gupta, C.P.: Dynamic seiz in online social networks: epidemiological modeling of untrue information. International Journal of Advanced Computer Science and Applications **11**(7) (2020)
13. Obadimu, A., et al.: Developing an epidemiological model to study spread of toxicity on youtube. In: SBP-BRiMS. pp. 266–276. Springer (2020)
14. Paulhus, D.L., et al.: The dark triad of personality: Narcissism, machiavellianism, and psychopathy. Journal of research in personality **36**(6), 556–563 (2002)
15. Sheth, A., et al.: Defining and detecting toxicity on social media: Context and knowledge are key. Neurocomputing **490**, 312–318 (2022)
16. Tambuscio, M., et al.: Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In: Proceedings of the 24th international conference on World Wide Web. pp. 977–982 (2015)
17. Tambuscio, M., et al.: Network segregation in a model of misinformation and fact-checking. Journal of Computational Social Science **1**, 261–275 (2018)
18. Yousefi, N., et al.: Examining the impact of toxicity on community structure in social networks. In: the Proceeding of SOTICS 2024 (2024)