Dynamic Simulation Framework for Disinformation Dissemination and Correction With Social Bots

Boyu Qiao[1,2], Kun Li[1]*, Wei Zhou[1]†, Songlin Hu[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{qiaoboyu, likun2, zhouwei, husonglin}@iie.ac.cn

Abstract

In the "human-bot symbiotic" information ecosystem, social bots play key roles in spreading and correcting disinformation. Understanding their influence is essential for risk control and better governance. However, current studies often rely on simplistic user and network modeling, overlook the dynamic behavior of bots, and lack quantitative evaluation of correction strategies. To fill these gaps, we propose **MADD**, a **M**ulti-**A**gent-based framework for **D**isinformation **D**issemination. MADD constructs a more realistic propagation network by integrating the Barabási–Albert Model for scale-free topology and the Stochastic Block Model for community structures, while designing node attributes based on real-world user data. Furthermore, MADD incorporates both malicious and legitimate bots, with their controlled dynamic participation allows for quantitative analysis of correction strategies. We evaluate MADD using individual and group-level metrics. We experimentally verify the real-world consistency of MADD's user attributes and network structure, and we simulate the dissemination of six disinformation topics, demonstrating the differential effects of fact-based and narrative-based correction strategies. Our code is publicly available at https://github.com/QQQQQQBY/BotInfluence.

1 Introduction

In the era of human-bot symbiosis, social bots are prevalent on social networks and actively involved in information sharing (Alrhmoun and Kertész, 2023; Cresci, 2020). Malicious bots are strategically deployed to spread disinformation and disrupt healthy online discussions (Bhale and Thirupurasundari, 2024; Qiao et al., 2025). Conversely, legitimate bots are increasingly employed for the task of disinformation correction (Ferrara, 2023; Costello et al., 2024). Therefore, developing robust network propagation simulation frameworks that quantitatively analyze the interaction between these opposing bots in information dissemination can help us better govern disinformation.

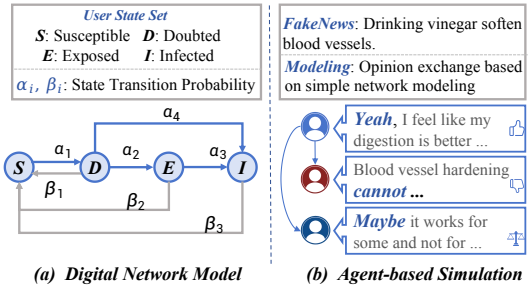


Figure 1: (a) Digital network model: Simulates dynamic user state transitions using probability parameters α and β. (b) Agent-based simulation model: Dynamically update and disseminate the opinions of user agents in a simple dissemination network.

Prior work in simulating disinformation propagation mainly follows two main paradigms: digital network-based propagation dynamics modeling and large language model (LLM)-based agent simulation. As shown in Figure 1(a), digital network-based models (Pastor-Satorras et al., 2015; Gopal et al., 2022; Govindankutty and Gopalan, 2023) simulate information diffusion through user state transitions. Although computationally efficient, this conventional approach oversimplifies reality by ignoring user demographics and individual behavioral differences, limiting its ability to capture key phenomena such as repeated exposure and group polarization. Figure 1(b) illustrates the emerging LLM-based agent simulation paradigm (Liu et al., 2024c,b), which improves user behavior modeling through semantic features and opinion propagating. However, it still faces notable limitations, redundant and irrelevant user attributes interfere with core propagation modeling, while inaccurate reconstruction of real-world network topologies causes

*Corresponding author: Kun Li
†Corresponding author: Wei Zhou

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 4688–4710
November 4-9, 2025

deviations from actual scenarios.

Furthermore, current simulation research on social bots' involvement in disinformation dissemination and correction still faces significant limitations (Qiao et al., 2024; Averza et al., 2022). For example, while Qiao et al. (2024) simulate bot-driven disinformation spread, they neglect the construction of intervention scenarios and lack systematic effectiveness evaluation. These shortcomings highlight the urgent need for a more comprehensive simulation framework that integrates core demographic features, realistic propagating network topologies, and a robust evaluation metrics, thereby enhancing the authenticity of dissemination modeling.

To address the aforementioned limitations, we propose **MADD**, a **M**ulti-**A**gent-based framework for **D**isinformation **D**issemination, designed to model both the dissemination and correction of disinformation by social bots. MADD integrates the scale-free property of Barabási-Albert Model (BAM, (Schweimer et al., 2022)) and the community structure of Stochastic Block Model (SBM (Abbe, 2018)) to construct a more realistic propagation network. It includes five node attributes, carefully designed based on real-world user data. Furthermore, we design three types of agents: regular users, malicious bots (spreading disinformation), and legitimate bots (correcting disinformation), enabling dynamic interactions. Using group and individual-level evaluation metrics, we validate the framework's effectiveness across six diverse disinformation topics and demonstrate the practical efficacy of fact-based and narrative-based correction strategies. Our study contributes the following:

(1) We introduce MADD, an innovative multi-agent framework for disinformation dissemination that incorporates refined real-user attributes and a propagation network that combines Barabási–Albert growth with Stochastic Block Model community structure to approximate real-world social graphs.

(2) To our knowledge, this is the first unified simulation to model the dynamic interplay between malicious and legitimate bots. We evaluate and compare fact-based versus narrative-based correction strategies, and MADD's modular design makes it straightforward to extend and test additional correction strategies.

(3) We meticulously design quantitative evaluation metrics at individual and group levels to systematically assess the impact of social bots on both the dissemination and correction of disinformation.

2 Related Work

2.1 Disinformation Dissemination Modeling

Disinformation dissemination modeling is theoretically significant for cybersecurity defense and information ecosystem security (López et al., 2024). Existing approaches include three main types: **digital-network models, agent-based simulations, and bot-assisted dissemination models.**

Digital-network modeling captures **macro-level propagation** by defining node states and transition probabilities (Cifuentes-Faura et al., 2022), often via epidemic-style models like SIR (Susceptible, Infected, Recovered) (Zhu and Wang, 2017), SIS (Susceptible, Infected, Susceptible) (Dong and Huang, 2018), SEIR (Susceptible, Exposed, Infected, Recovered) (Liu et al., 2018), and SEDIS (Susceptible, Exposed, Disseminated, Infected) (Govindankutty and Gopalan, 2022). However, these models **ignore memory effects** and **lack granular user attribute characterization**, limiting their ability to simulate phenomena like repeated exposure.

Agent-based simulation modeling (ABM) addresses upper limitations at the **micro level** by finely characterizing individual behavior and interactions, including user attributes and memory (Liu et al., 2024c,b; Muhammad and Kasahara, 2024). By integrating role settings with dynamic memory-feedback, ABM enables detailed simulations of user cognition and social interactions. However, ABM still faces challenges: **irrelevant and redundant user attributes** can reduce accuracy, and the **lack of a systematic network model** hinders capturing real-world topology and dynamics.

Bot-assisted dissemination modeling further considers **automated accounts' impact on disinformation.** For example, Qiao et al. (2024) simulate bot-driven spread but **lack intervention scenarios** and impact assessment. Aerza et al. (2022) quantify user susceptibility using belief scores but typically **overlook the interplay** of user attitudes and socio-environmental factors. Moreover, many such approaches focus primarily on exposure counts, neglecting how attitudes interact with context to shape dissemination and correction.

2.2 Disinformation Correction Strategies

Existing research primarily focuses on two correction strategies: **fact-based** and **narrative-based** (Song et al., 2025). **Fact-based correction** (Boukes and Hameleers, 2023; Nyhan et al., 2020)

widely used by professional fact-checkers, directly refutes disinformation with accurate facts and evidence through rational argumentation. **Narrative-based correction** (Dahlstrom, 2021; Vafeiadis and Xiao, 2021) conveys truth by sharing authentic eyewitness accounts or related stories, enhancing immersion and persuasiveness. This study we systematically evaluate the effectiveness of legitimate bots applying these strategies across six disinformation topics.

3 Methodology

In this work, we design a **M**ulti-**A**gent-based framework for **D**isinformation **D**issemination (MADD) to model the dynamic impact of social bots on disinformation dissemination and correction. As shown in Figure 2, MADD comprises three modules: **disinformation dissemination modeling, correction strategy, and propagation impact quantification**.

3.1 Disinformation Dissemination Modeling

This module consists of three components: user/bot agent attributes, disinformation and dissemination rules, and disinformation dissemination network.

3.1.1 User/Bot Agent Attributes

We design three agent types for the dissemination simulation: regular users, malicious bots (MBots), and legitimate bots (LBots). For efficiency and to reduce irrelevant attribute interference, we define five core attributes related to disinformation spread: interest community, trust threshold, dissemination tendency, social influence, and activation time.

User/Bot Agents: For regular users, we initialize their attributes based on real X/Twitter user data (data collection details in Appendix B.1). For MBots and LBots, their behavior is primarily controlled by automation, we configure their attributes using predefined procedures.

Interest Community (IC): To capture each user's communities of interest, we introduce interest community scores $\mathcal{IC}_{uj}$ for each user u in community j. This score, ranging from 1 to 10, measures attention to community-specific content using LLM evaluation (full prompts in Appendix B.2). It serves as a basis for community partitioning in subsequent propagation network construction and as a key feature for computing dissemination tendencies.

Trust Threshold (TT): To quantify user ability to identify disinformation in different communities, we use LLMs to evaluate the trust threshold $\mathcal{TT}_{uj}$ for each user u within community j. By analyzing users' historical info and basic attributes via prompt engineering (full prompt in Appendix B.3), we evaluate users' resistance to disinformation in different communities.

Dissemination Tendency (DT): We combine the truncated power law distribution (Cha et al., 2010; Kwak et al., 2010)) and user interest community scores $\mathcal{IC}_{uj}$ to quantify the user's dissemination tendency among communities. We fit the truncated power law distribution based on real user data, and we consider the user interest community score because user interest influences their likelihood to share information (Zhao and Cui, 2017). The final dissemination tendency $\mathcal{DT}_{uj}$ for user u in community j is a weighted linear combination:

$$\mathcal{DT}_{uj} = \Big[\theta \cdot \mathcal{CDF}\left(C \cdot (x_u)^{-\alpha} e^{-\lambda}\right) + (1-\theta) \cdot \frac{\mathcal{IC}_{uj}}{\max_j \mathcal{IC}_{uj}}\Big] \cdot e^{-\xi n}, \quad (1)$$

where $\theta \in [0,1]$ balances the truncated power-law component (capturing scale-free propagation patterns via its cumulative density function (CDF) with normalization constant C, share number $x_u > x_{min}$, exponent α, and cutoff λ) and the max-normalized $\mathcal{IC}_{uj}$ (which preserves relative differences across communities). The exponential term $e^{-\xi n}$ models the decreasing tendency to share as a user encounters the same information more frequently. (α, λ, C) are optimized against real-world data (fitting details in Appendix B.4).

Social Influence (SI): To quantify the user's influence in the disinformation dissemination network, we introduce a social influence attribute to calculate the node degree centrality in the subsequently network. For a community j with n users, we collect each user u's follower count f_u, forming the set $F = \{f_1, f_2, \ldots, f_n\}$. Through normalization, the social influence $\mathcal{SI}_{uj}$ of user u in community j is defined as: $\mathcal{SI}_{uj} = \frac{f_u}{\sum_{i=1}^{n} f_i}$. A higher $\mathcal{SI}_{uj}$ indicates greater centrality and a stronger role in the dissemination process.

Activation Time (AT): To capture the temporal dynamics of user participation in disinformation dissemination, we employ a discrete-time sequence design with L steps, forming a time sequence $T = \{t_1, t_2, ..., t_L\}$. Based on statistical analysis of real-world user activity times, we define a time-varying activation probability $\mathcal{AT}_{ut} \in [0,1]$ for each user u, where a higher value indicates an

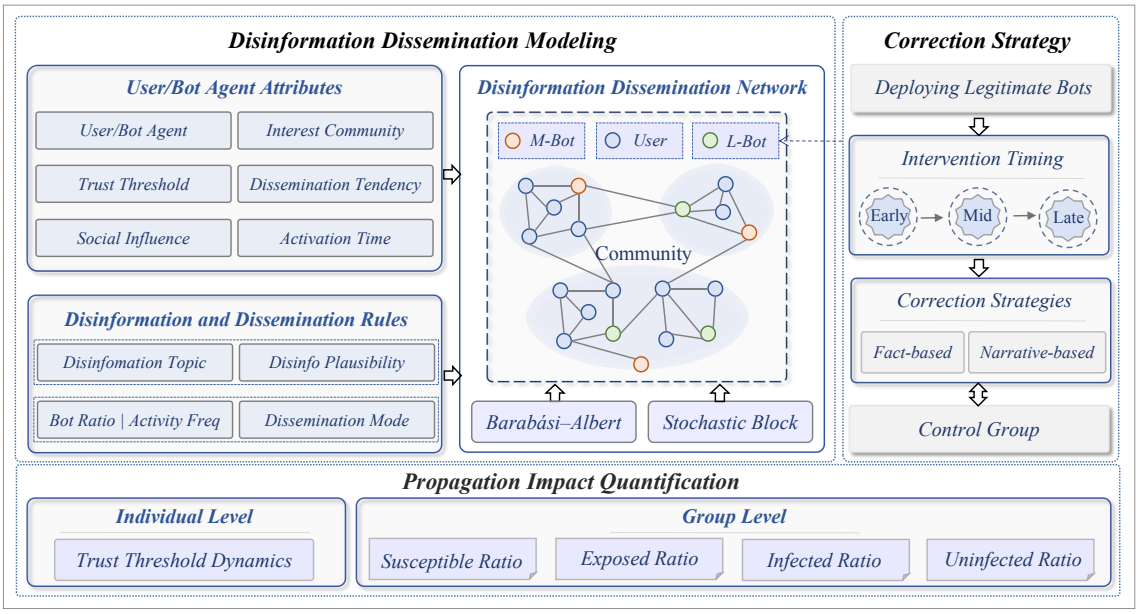


Figure 2: Architecture of our proposed MADD.

increased likelihood of exposure to disinformation at that time step.

3.1.2 Disinformation and Dissemination Rules

We set disinformation by topic and plausibility. Dissemination rules is defined by bot ratio and activity frequency, and dissemination mode.

Disinformation Topic: The topic type of disinformation is associated with the community structure in the propagation network. Given that users in different communities show variations in their trust thresholds and tendency to spread specific disinformation topics (Wei et al., 2025), we introduce the attribute of disinformation topic.

Disinformation Plausibility (DP): DP refers to the degree of logical coherence and argumentative structure of disinformation, where higher plausibility increases the likelihood of users misinterpreting it as truthful. Inspired by Wan et al. (2024), we quantify plausibility $\mathcal{DP}_j$ of disinformation j by evaluating emotional expression, propaganda strategies, information framing and logical consistency (prompt design in Appendix B.5).

Bot Ratio and Activity Frequency: To quantify the dynamic impact of social bots, we adjust malicious ($\mathcal{MR}_j$) and legitimate ($\mathcal{LR}_j$) bot injection ratios and their activity frequencies ($\mathcal{MF}_j$ and $\mathcal{LF}_j$) within each community j. This setup helps evaluate the potential impact of different bot manipulation strategies.

Dissemination Mode: To better model disinformation diffusion and reduce noise, we restrict regular-user actions to **Repost** and **Quote**, reserving original **Post** actions for MBots and LBots. Direct reposts are treated as implicit endorsements, and quoted content is analyzed with LLMs to infer users' stances. Detailed interaction and decision-making strategies are provided in Appendix B.10.

3.1.3 Disinformation Dissemination Network

We combine the Stochastic Block Model (SBM) (Abbe, 2018; Nowicki and Snijders, 2001) and Barabási-Albert Model (BAM) (Schweimer et al., 2022; Barabási and Albert, 1999) to construct disinformation propagation networks. The SBM captures community structure with dense intra-community and sparse inter-community connections, while BAM introduces the scale-free properties and preferential attachment in real networks. Our construction process first **assigns users to communities** using **interest community scores** $\mathcal{IC}_{uj}$, then **generate scale-free subgraphs** within communities based on **social influence scores** $\mathcal{SI}_{uj}$. Algorithm 1 outlines the network-construction procedure. The resulting network naturally emerges with influential users (high-degree) as dissemination hubs and ordinary users (low-degree) as peripheral participants. (Implementation details in Appendix B.6.)

Our propagation network represents users as nodes and potential information diffusion pathways as edges. By adjusting parameters such as com-

munity assignment and social influence of central nodes, we construct a network that simultaneously captures community structure and scale-free properties, providing a structured foundation for subsequent disinformation dissemination simulations. Appendix Figure 17 illustrates our extensions to disinformation-dissemination modeling, showing how user, information attributes, and the network model are incorporated into the simulation.

Algorithm 1 Building Disinformation Dissemination Networks Based on SBM-BAM

Specify the Inputs and Outputs
Inputs: Total nodes N. Initial number of nodes in each community m_0. Number of communities J. User interest community scores $\mathcal{IC}_{uj}$. Predefined threshold for community assignment τ. Number of edges each new node forms m ($m \leq m_0$).
Outputs: Constructed dissemination network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
Initialize sets $C_j \leftarrow \emptyset$ for all $j \in \{1, \ldots, J\}$
(1) Node Community Assignment (SBM Component):
for $u \leftarrow 1$ to N **do**
  Load u's interest community scores $\mathcal{IC}_{uj}$ for all $j \in \{1, \ldots, J\}$
  for $j \leftarrow 1$ to J **do**
    if $\mathcal{IC}_{uj} \geq \tau$ **then**
      Add u to node set of community C_j
    end if
  end for
end for
(2) Initialize Intra-Community Connections (BAM Component):
for $j \leftarrow 1$ to J **do**
  Initialize community j's graph $\mathcal{G}_j = \{\mathcal{V}_j, \mathcal{E}_j\}$, where $\mathcal{V}_j$ are the first m_0 users assigned to community j
  Create a complete graph on $\mathcal{V}_j$ and set $\mathcal{E}_j$ to these edges
end for
(3) Iterative Network Growth (BAM Component):
for $j \leftarrow 1$ to J **do**
  for u in C_j **do**
    for existing node e in $\mathcal{V}_j$ **do**
      Compute social influence score:

$$\mathcal{SI}_e = \frac{F_e}{\sum_{i \in \mathcal{V}_j} F_i}$$

      where F_e is the follower count of e
    end for
    Select m nodes from $\mathcal{V}_j$ based on $\mathcal{SI}_e$ to form $\mathcal{V}_{\text{selected}}$
    for each selected node n in $\mathcal{V}_{\text{selected}}$ **do**
      Add edge (n, u) to $\mathcal{E}_j$
    end for
    Add node u to $\mathcal{V}_j$
  end for
end for
return The network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

3.2 Correction Strategy

We design **fact-based** and **narrative-based** correction strategies, to assess intervention effectiveness across communities. Specifically, we deploy malicious bots to spread disinformation and legitimate bots to deliver correction information. By comparing against **control groups**, we evaluate the effectiveness of different **correction strategies** and **intervention timings**.

Correction Strategy: To systematically evaluate effectiveness of different correction strategies, we manually design guiding content for two types of corrective messages, which can subsequently be rewritten and disseminated by legitimate bots with the assistance of LLMs (examples of both strategies are provided in Appendix B.7):

(1) **Fact-based Correction**: Directly refuting disinformation by citing authoritative data, such as verified data and scientific evidence.

(2) **Narrative-based Correction**: Weakening the emotional appeal of disinformation through emotionally engaging narratives, such as personal stories and metaphors.

Intervention Timing: To examine the impact of intervention timing, we define three intervention stages: early, mid-stage, and late, corresponding to different time steps in the disinformation diffusion process (details in Appendix B.8).

Control Group: To evaluate the effects of the two correction strategies and three intervention timings, we establish a control group that does not implement any strategies. By comparing the propagation dynamics between the experimental and control groups, we quantify the effectiveness of different strategies across communities.

3.3 Propagation Impact Quantification

We assess the impact of disinformation dissemination at both individual and group levels. The simulation and evaluation algorithm process is presented in Appendix B.10.

3.3.1 Individual Level

Trust Threshold Dynamics: To evaluate the dynamic impact of repeated exposure to disinformation and corrective information on users' trust thresholds, we model changes in trust threshold using "enhancement" and "decay" terms. Specifically, we use the initial trust threshold $\mathcal{TT}_{uj}$ as a baseline and dynamically update the new trust threshold $\hat{\mathcal{TT}}_{uj}$ based on users' cumulative exposure to corrective and disinformation (Kemp et al., 2024). In modeling the dynamic adjustment of trust thresholds, we take into account both the user's individual interests and the influence of posts from

neighboring users with varying levels of social influence. Considering the common psychological principle of diminishing marginal utility, we further describe the evolution of trust thresholds after multiple exposures as a process driven by both exponential growth and exponential decay (Daley and Kendall, 1964; Loomba et al., 2021):

$$\begin{aligned} \mathrm{En}_{uj} &= \gamma \left(1 - e^{-\beta \sum_{k \in \mathcal{N}_{\mathrm{corr}}} \mathcal{SI}_k F'_{kj}}\right) \\ \mathrm{De}_{uj} &= (1-\gamma) \left(1 - e^{-\delta \sum_{k \in \mathcal{N}_{\mathrm{dis}}} \mathcal{SI}_k F_{kj}}\right) \\ \hat{\mathcal{TT}}_{uj} &= \mathrm{clip}\left(\mathcal{TT}_{uj} + \mathrm{En}_{uj} - \mathrm{De}_{uj}, 0, 1\right), \end{aligned} \tag{2}$$

where En_{uj} (Enhancement term) models the reinforcing effect from exposure to corrective content, while De_{uj} (Decay Term) captures the diminishing effect due to disinformation exposure. γ balances correction and disinformation influence, while β and δ control the changing rate. $\mathcal{N}_{\mathrm{corr}}$ and $\mathcal{N}_{\mathrm{dis}}$ are neighbors. $\mathcal{SI}_k$ is neighbor k's social influence score. F'_{kj} and F_{kj} are the persuasive strength of corrective and disinformation content (evaluated by LLMs, Appendix B.9). Subsequently, user u's disinformation discernment ability $\mathcal{DA}_{uj}$, which depends on their trust threshold $\hat{\mathcal{TT}}_{uj}$ and disinformation plausibility $\mathcal{DP}_j$, is calculated as follows:

$$\mathcal{DA}_{uj} = 1 - (1 - \hat{\mathcal{TT}}_{uj})\mathcal{DP}_j. \tag{3}$$

3.3.2 Group Level

At the group level, we assess the impact of disinformation using four user status ratios within community j at time t: (1) **Susceptible Ratio** ($\mathcal{SR}_t^j$): Proportion of users unexposed to disinformation. (2) **Exposed Ratio** ($\mathcal{ER}_t^j$): Proportion of users having encountered disinformation at least once. (3) **Infected Spreaders Ratio** ($\mathcal{IR}_t^j$): Proportion of spreaders exposed to and believing disinformation, actively propagating it. (4) **Uninfected Spreaders Ratio** ($\mathcal{UR}_t^j$): Proportion of spreaders exposed to but not believing disinformation, distributing corrective information.

4 Experiments

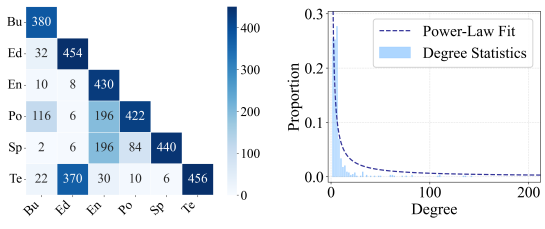
4.1 Experiment Setup

We employ DeepSeek-V3 (Liu et al., 2024a) as the base model, balancing response quality and computational efficiency. The experiment simulates six communities: **Business, Education, Entertainment, Politics, Sports, and Technology**. Regular-user attributes are drawn from real user data, with approximately 100 regular users distributed across communities. Bot agents mimic regular-user behaviors, including programmable activation times and dissemination tendencies. Furthermore, we set up 15% malicious bots (Ferrara et al., 2016) and 5% legitimate bots in each community network to spread false and corrective information, respectively. Detailed parameter definitions, value ranges, and data sources are available in Appendix A.

4.2 MADD and Real-World Consistency Evaluation

To confirm the MADD framework's trustworthiness, we perform a double verification: we first verify that user attributes and network configurations statistically match real social networks, and and then we validate our simulations by comparing the spread trends of disinformation with existing empirical studies.

Consistency of Attributes and Network Setup: Figure 3(a) shows that the distribution of user interest communities aligns with the structural characteristics of real-world social networks, exhibiting dense intra-community links and sparse inter-community links (Yang and Leskovec, 2012). Figure 3(b) reveals that the network's degree distribution, based on real user social influence, follows a power law pattern. Furthermore, Appendix C.1, Figure 9 indicates users' dissemination tendencies conform to the heavy-tailed distribution observed in real networks (Goel et al., 2016). Figure 10 shows user trust threshold scores approximately follow a normal distribution, consistent with the conclusion from previous studies that most users hold a neutral stance (Ecker et al., 2022).



(a) Interest Community. (b) Node Degree.

Figure 3: Interest Community and Node Degree Distributions. In (a), the X and Y axis labels represent the first two letters of each of the six communities.

Simulation Results of MADD: To validate the consistency of MADD's simulation with the real world, we simulate the diffusion of disinformation across six topics within the entire network constructed by six communities over 72 time steps.

Each topic initiates within its most relevant community. Figure 4 illustrates the changing proportions of user statuses within each community (recorded every 12 time steps), while Figure 5 shows the disinformation's spread across the entire network. Notably, our model simulates user activation timing, leading to lower activity at time steps 12, 36, and 48 compared to 24, 48, and 72, which explains some of the fluctuations observed in Figure 4.

Figure 4 shows that the proportion of "infected" spreaders in each community exhibits a trend of rapid initial increase followed by a gradual decline, peaking at approximately the 24-th time step. This aligns with existing empirical studies (Vosoughi et al., 2018; Shao et al., 2018; Tucker et al., 2018) and classical epidemic spreading models like SIR (Govindankutty and Gopalan, 2024; Gopal et al., 2022). Notably, the proportion of "uninfected" spreaders is higher than that of "infected" spreaders, indicating that a larger proportion of users can more successfully identify disinformation based on their trust thresholds and judgments about the plausibility of the disinformation. Furthermore, the peak time for the proportion of "uninfected" spreaders mostly occurs later than that of "infected" spreaders, suggesting that the spread of corrective information or rational responses is slower. In contrast, the "Politics" differs from other communities. The proportion of "infected" spreaders in this community does not exhibit the typical unimodal pattern but shows significant fluctuations at different time steps, likely due to the sensitive and debated nature of political information (Shin et al., 2018). This simulation highlights the challenge of predicting political topics on social networks.

Figure 5 illustrates the evolving proportions of infected, uninfected, and exposed users across the entire network for different disinformation topics. Most topics take 60-72 time steps for network-wide spread. In contrast, within closely related communities (Figure 4), spread is much faster, around 24 time steps. This suggests faster diffusion within communities but slower propagation across them due to the inherent community structure of the network. The increasing of all user states further suggests the ongoing diffusion of disinformation throughout the network. Capturing the critical turning point at which the "infected" rate begins to decline necessitates longer simulations. However, considering the computational cost of extended simulations, especially the LLM token usage for modeling user opinions in multi-level networks, subsequent experiments will focus on simulating spread within single communities to efficiently evaluate correction strategies.

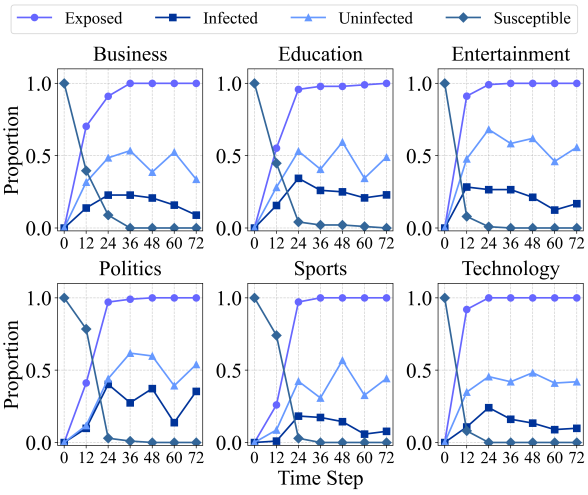


Figure 4: Proportion of user status in each community over time.

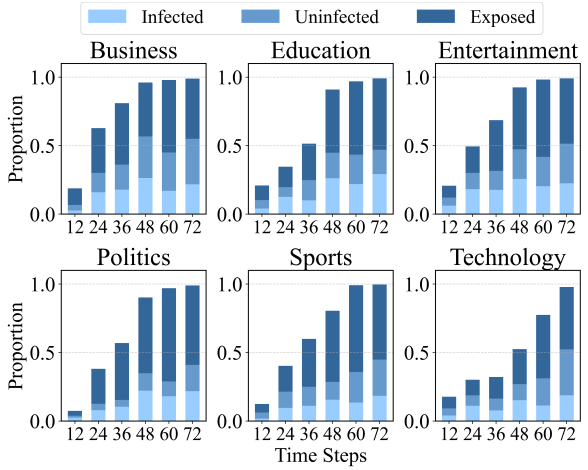


Figure 5: Proportion of user status change across the entire network over time.

4.3 Group-Level: Effectiveness Evaluation of Correction Strategies

To evaluate two correction strategies, we implement them for early propagation in Figure 6 and compare the resulting changes in the proportion of infected spreaders over time against a control group (without external correction strategies). Observing Figure 6, we find that fact-based correction is significantly more effective than narrative-based correction in 'Business', 'Politics', and 'Technology'. For 'Education' and 'Sports', both strategies show comparable effectiveness in curbing disinformation spread. However, neither strategy significantly impacts 'Entertainment', likely due to

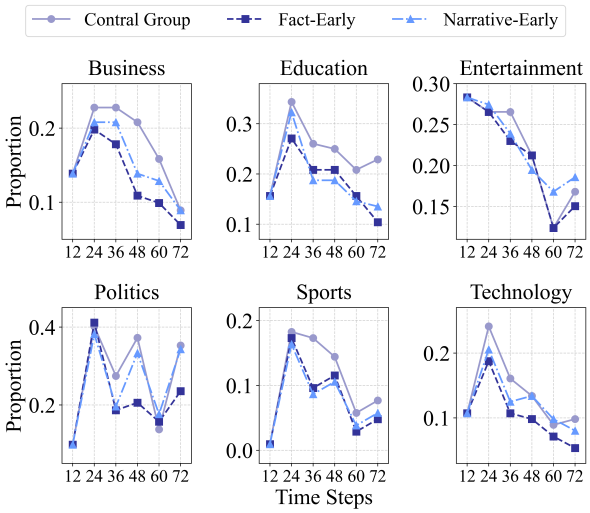


Figure 6: Comparison of early intervention effects of two correction strategies.

its inherent subjectivity and diverse interpretations, limiting the impact of debunking.

Appendix C.2, Figures 11 and 12 illustrate limited corrective effects of the two strategies in the mid and late propagation stages across most communities. Moreover, in certain communities like 'Politics', we even observe negative impacts, likely due to substantial prior disinformation exposure causing users to distrust debunking and strengthen their original views, leading to an echo chamber (Garimella et al., 2018).

4.4 Individual-Level: Dynamic Changes in User Trust Thresholds

To assess the potential impact of different correction strategies on user trust thresholds, Figure 7 visualizes the dynamic evolution of user trust thresholds over time steps in the early intervention scenario, specifically displaying the mean and standard deviation of all user trust thresholds at each time step. Observing Figure 7, we find that even in the control group where no external correction strategies are implemented, the user trust threshold for disinformation still exhibits a continuous upward trend, although its rate of increase is slower than in the simulations where correction strategies are applied. This observation indicates that even in the absence of intervention, a certain proportion of spontaneous debunking users exists within the network, who can mitigate the negative impact of disinformation to some extent.

Our further analysis comparing the impact of fact-based and narrative-based correction strategies on trust thresholds (relative to the control group) indicates that fact-based correction results in larger trust threshold increases across most communities. However, the "Politics" community displays a distinctive pattern, showing the smallest overall increase and the most significant fluctuations. Specifically, while fact-based corrective strategies significantly increased the trust threshold of this community initially, they eventually led to a downward trend in trust thresholds as disinformation continued to hedge against the spread of corrective information (Vosoughi et al., 2018). This phenomenon may stem from the highly controversial nature of political information itself, coupled with the often greater emotional appeal of disinformation, prompting users to shift their positions repeatedly. In addition, the persistence of cognitive bias and group polarization in political communities (Van Bavel et al., 2021) pose a substantial barrier to the effective dissemination and reception of corrective information.

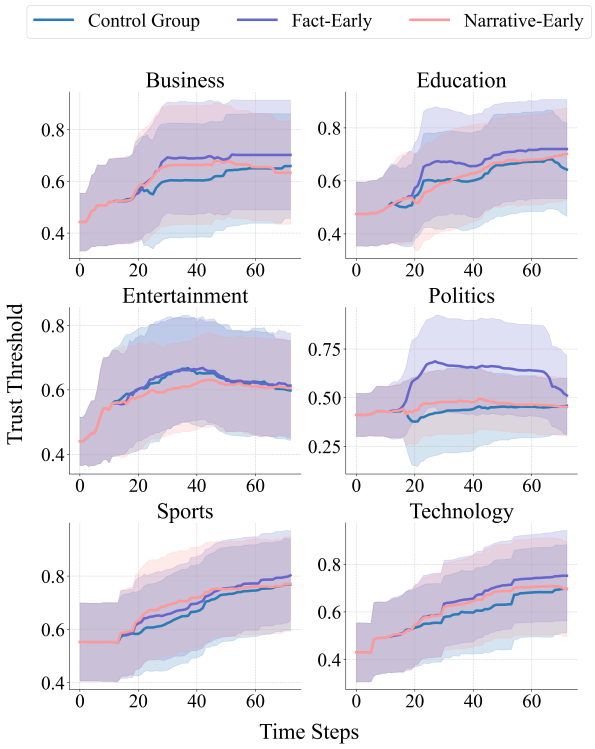
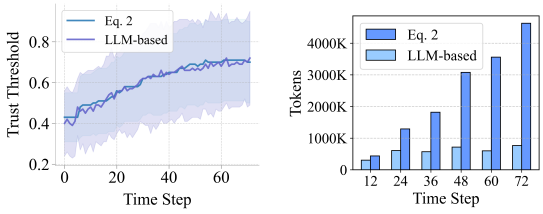


Figure 7: Effects of correction strategies on user trust thresholds during early-stage intervention

Appendix C.3, Figures 13 and 14 illustrate the changes in user trust thresholds when correction strategies are implemented in the mid and late stages. We observe that their effect on increasing user trust thresholds for evaluations is extremely limited. Furthermore, in the "Entertain-

ment," "Business," and "Politics" during the mid-stage, and in the "Sports" during the late-stage, the control group (without correction strategies) actually shows higher user trust thresholds than the experimental group (with correction strategies). This phenomenon likely stems from two main reasons: the intervention timing is delayed and echo chamber effects may have already formed within some user groups. Consequently, the aforementioned experimental results underscore the critical importance of intervening in the early stages of disinformation spread and implementing effective correction strategies.

Trust Threshold Validity Assessment: We validate the effectiveness of our proposed trust threshold calculation method (Equation 2) by comparing it with an LLM-based evaluation method. Figure 8(a) demonstrates a high degree of consistency in the trust threshold evolution trends of the both methods across different time steps. Furthermore, Figure 8(b) reveals that our approach achieves a 76% average reduction in token consumption compared to the LLM method. The experimental results indicate that MADD not only maintains evaluation accuracy comparable to LLMs but also significantly improves computational efficiency and effectively reduces resource overhead. These findings demonstrate the practical value of our proposed dynamic trust threshold estimation method in accurately characterizing user trust properties.



(a) Trust Thresholds. (b) Token Consumption.

Figure 8: Comparative Analysis of Trust Threshold Distributions and Token Consumption

5 Conclusion

In this work, we address key limitations in current disinformation dissemination research, including oversimplified user and network models and the lack of quantitative evaluation of correction strategies. To overcome these challenges, we propose MADD. MADD simulates realistic online social dynamics by incorporating five key user attributes relevant to disinformation spread and modeling network structure with scale-free and community properties. We then utilize individual and group level metrics to rigorously assess how MBots spreading disinformation and LBots providing corrections impact regular users. Our comparative experiments demonstrate the effectiveness of various correction strategies in curbing disinformation, offering valuable insights for future research.

Limitations

Our study has several limitations in the experimental design that need to be addressed.

First, when evaluating correction strategies, we validate within a single community rather than the entire network to avoid prohibitive token costs. While sufficient for preliminary insights, future work should build larger-scale network models and extend the simulation horizon to more fully assess the mechanisms by which social bots facilitate disinformation and how interventions propagate.

Second, there remains a potential concern that the LLM used as a user agent may exhibit topic-specific biases, which could influence simulation outcomes. For example, even under identical simulation settings, the extent of information spread may vary across topics due to the LLM's inherent preferences or assumptions. To mitigate such model-driven bias, future research should prioritize the selection of more neutral or balanced topics, thereby improving the objectivity and robustness of the simulation results.

Ethics Statement

This study follows research-ethics standards. First, the six disinformation topics and their corrections come from verified cases by Snopes, PolitiFact, and FactCheck.org; we do not create new disinformation. Second, concerning data protection, for the real user data involved in the study, we implement rigorous anonymization measures to ensure the full protection of user privacy.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. U24A20335) and by the Youth Innovation Promotion Association of the Chinese Academy of Sciences (CAS). The authors thank the anonymous reviewers and the meta-reviewer for their helpful comments.

References

Emmanuel Abbe. 2018. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86.

Abdullah Alrhmoun and János Kertész. 2023. Emergent local structures in an ecosystem of social bots and humans on twitter. *EPJ Data Science*, 12(1):39.

Aldo Averza, Khaled Slhoub, and Siddhartha Bhattacharyya. 2022. Evaluating the influence of twitter bots via agent-based social simulation. *IEEE Access*, 10:129394–129407.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science*, 286(5439):509–512.

Kanchan Bhale and DR Thirupurasundari. 2024. Malicious social bot detection in twitter: A review. In *2024 4th International Conference on Computer, Communication, Control & Information Technology (C3IT)*, pages 1–6. IEEE.

Mark Boukes and Michael Hameleers. 2023. Fighting lies with facts or humor: Comparing the effectiveness of satirical and regular fact-checks in response to misinformation and disinformation. *Communication Monographs*, 90(1):69–91.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 10–17.

Javier Cifuentes-Faura, Ursula Faura-Martínez, and Matilde Lafuente-Lechuga. 2022. Mathematical modeling and the use of network models as epidemiological tools. *Mathematics*, 10(18):3347.

Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814.

Stefano Cresci. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.

Michael F Dahlstrom. 2021. The narrative truth about scientific misinformation. *Proceedings of the National Academy of Sciences*, 118(15):e1914085117.

Daryl J Daley and David G Kendall. 1964. Epidemics and rumours. *Nature*, 204(4963):1118–1118.

Carlos Diaz Ruiz and Tomas Nilsson. 2023. Disinformation and echo chambers: how disinformation circulates on social media through identity-driven controversies. *Journal of public policy & marketing*, 42(1):18–35.

Suyalatu Dong and Yong-Chang Huang. 2018. Sis rumor spreading model with population dynamics in online social networks. In *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 1–5. IEEE.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Emilio Ferrara. 2023. Social bot detection in the age of chatgpt: Challenges and opportunities. *First Monday*.

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

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922.

Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2016. The structural virality of online diffusion. *Management science*, 62(1):180–196.

Greeshma N Gopal, G Sreerag, and Binsu C Kovoor. 2022. The sepns model of rumor propagation in social networks. In *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021*, pages 695–707. Springer.

Sreeraag Govindankutty and Shynu Padinjappurath Gopalan. 2024. Epidemic modeling for misinformation spread in digital networks through a social intelligence approach. *Scientific Reports*, 14(1):19100.

Sreeraag Govindankutty and Shynu Padinjappurathu Gopalan. 2022. Sedis—a rumor propagation model for social networks by incorporating the human nature of selection. *Systems*, 11(1):12.

Sreeraag Govindankutty and Shynu Padinjappurathu Gopalan. 2023. From fake reviews to fake news: A novel pandemic model of misinformation in digital networks. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(2):1069–1085.

Paige L Kemp, Aaron C Goldman, and Christopher N Wahlheim. 2024. On the role of memory in misinformation corrections: Repeated exposure, correction durability, and source credibility. *Current Opinion in Psychology*, 56:101783.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Xiongding Liu, Tao Li, and Mi Tian. 2018. Rumor spreading of a seir model in complex social networks with hesitating mechanism. *Advances in Difference Equations*, 2018(1):1–24.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024b. From skepticism to acceptance: simulating the attitude dynamics toward fake news. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7886–7894.

Yuhan Liu, Zirui Song, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2024c. From a tiny slip to a giant leap: An llm-based simulation for fake news evolution. *arXiv preprint arXiv:2410.19064*.

Sahil Loomba, Alexandre De Figueiredo, Simon J Piatek, Kristen De Graaf, and Heidi J Larson. 2021. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348.

Alejandro Buitrago López, Javier Pastor-Galindo, and José A Ruipérez-Valiente. 2024. Frameworks, modeling and simulations of misinformation and disinformation: a systematic literature review. *arXiv preprint arXiv:2406.09343*.

Radifan Fitrach Muhammad and Shoji Kasahara. 2024. Agent-based simulation of fake news dissemination: the role of trust assessment and big five personality traits on news spreading. *Social Network Analysis and Mining*, 14(1):75.

Krzysztof Nowicki and Tom A B Snijders. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087.

Brendan Nyhan, Ethan Porter, Jason Reifler, and Thomas J Wood. 2020. Taking fact-checks literally but not seriously? the effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political behavior*, 42:939–960.

Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. 2015. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925–979.

Boyu Qiao, Kun Li, Wei Zhou, Shilong Li, Qianqian Lu, and Songlin Hu. 2024. Botsim: Llm-powered malicious social botnet simulation. *arXiv preprint arXiv:2412.13420*.

Boyu Qiao, Wei Zhou, Kun Li, Shilong Li, and Songlin Hu. 2025. Dispelling the fake: Social bot detection based on edge confidence evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):7302–7315.

Christoph Schweimer, Christine Gfrerer, Florian Lugstein, David Pape, Jan A Velimsky, Robert Elsässer, and Bernhard C Geiger. 2022. Generating simple directed social network graphs for information spreading. In *Proceedings of the ACM web conference 2022*, pages 1475–1485.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787.

Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287.

Yunya Song, Yuanhang Lu, Stephanie Jean Tsang, Jingwen Zhang, and Kelly YL Ku. 2025. When corrections fail: Effects of misinformation targets, repeated exposure, and partisanship on misinformation beliefs. *International Journal of Communication*, 19:25.

Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Michail Vafeiadis and Anli Xiao. 2021. Fake news: How emotions, involvement, need for cognition and rebuttal evidence (story vs. informational) influence consumer reactions toward a targeted organization. *Public Relations Review*, 47(4):102088.

Jay J Van Bavel, Elizabeth A Harris, Philip Pärnamets, Steve Rathje, Kimberly C Doell, and Joshua A Tucker. 2021. Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*, 15(1):84–113.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2637–2667.

Lingwei Wei, Dou Hu, Wei Zhou, Philip S Yu, and Songlin Hu. 2025. Structure-adaptive adversarial contrastive learning for multi-domain fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9739–9752.

Jaewon Yang and Jure Leskovec. 2012. Community-affiliation graph model for overlapping network community detection. In *2012 IEEE 12th international conference on data mining*, pages 1170–1175. IEEE.

Narisa Zhao and Xuelian Cui. 2017. Impact of individual interest shift on information dissemination in modular networks. *Physica A: Statistical mechanics and its applications*, 466:232–242.

Liang Zhu and Youguo Wang. 2017. Rumor spreading model with noise interference in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 469:750–760.

A Experimental Parameter Settings

We present the parameter variables, their value ranges, definitions, and sources necessary for replicating the current simulation in Table 2, facilitating reproducibility for researchers. We construct six distinct community networks— **Entertainment, Technology, Sports, Business, Politics, and Education**, which collectively form the integrated propagation network. Correspondingly, we define six categories of disinformation topics, each closely aligned with one specific community. Thus, both the number of communities and disinformation categories are uniformly denoted by J, while an individual community or disinformation type is indexed by j.

B Methodology

B.1 Real-World Data

Data collection sources: We select six representative vertical communities from the 'Communities' module on the X/Twitter platform as data sources: Entertainment, Technology, Sports, Business, Politics, and Education. For each community, we systematically collect:

(1) Users' basic profile information: user name, self-introduction of the user, number of posts by the user, number of followers, number of follows, creation time, whether verified or not.

(2) A sample of 200 content posts, including original posts, retweets, and quotes.

Data filtering strategies: Given that some users exhibit a high posting frequency—reaching the 200-post threshold within a short period—we apply a time window from February 7 to February 10, 2025 (a total of three days) to filter the data. This approach allows for the analysis of time-series characteristics such as user sharing frequency and activation times while ensuring consistency. Furthermore, to enhance simulation efficiency, we randomly sample approximately 100 users from the collected data of each community to participate in the subsequent simulations. As shown in Table 1, we provide detailed statistics of each community after applying the time-based filtering and user sampling, including:

(1) The number of users in each community.

(2) The number of valid posts, retweets and quotes within the specified time frame.

This data processing strategy ensures the reliability and comparability of our subsequent analysis.

Potential limitations: The collected dataset size is relatively small compared to real-world social networks, and our analysis covers only six communities, whereas the actual number of online communities is much larger. In addition, we filter only three days of user activity and limited the dataset to 200 text entries per user, which allows us to capture only short-term interests rather than long-term user preferences.

B.2 Interest Community Prompt

Although Appendix B.1 describes how users are collected from the "Communities" module on the X platform, this doesn't imply that the collected users are solely interested in the community from which they are sampled. Gathering users from the six major communities is to ensure the presence of active users within those domains, but it doesn't restrict their interests to just one community. Therefore, it's necessary for us to further evaluate the diversity of their interest communities. This step ensures that the subsequently constructed propagation network topology has multiple communities, and that a subset of users with cross-community interests can serve as a path for information transfer between these communities.

We design a prompt to evaluate users' interest communities based on their historical posts, retweets, quotes, and basic profile information. Below is the detailed prompt.

Interest Community Intensity Prompt

Please evaluate the user's interest intensity in different communities based on their [User Data], including historical posts, retweets, quotes, and basic profile information. The interest communities include Entertainment, Technology, Sports, Business, Politics, and Education. Based on the provided user data, assign a score (1-10) for each domain and briefly explain the reasoning behind the rating. If certain domains lack sufficient data support, please mark them as "Insufficient Data" and explain the reason. Your response must strictly follow the [Output Format].

[User Data:]
Personal Description: $\{personal\_description\}$
Historical Posts: $\{historical\_posts\}$

Table 1: Statistical information in each community. We present the mean and standard deviation of users' original posts, retweets, and quotes.

Community	Entertainment	Technology	Sports	Business	Politics	Education
Number of Users	113	112	104	101	102	96
Number of Posts	24.05±21.54	14.38±13.09	31.70±24.25	34.17±21.17	36.01±28.96	23.96±15.65
Number of Retweets	32.63±21.54	11.35±4.31	24.72±12.95	36.64±14.59	48.23±31.65	14.03±6.25
Number of Quotes	10.95±7.87	6.08±3.30	11.42±7.53	4.82±2.62	23.51±12.85	8.03±2.15

Historical Retweets:{$historical\_retweets$}
Historical Quotes:{$historical\_quotes$}

[Evaluation Guidelines:]
1. Assign a rating of 1-10 points to each community
2. Provide a brief rationale for each rating
3. If there is insufficient data in a community, mark it as "Insufficient Data" and explain why.
[Output Requirements:]
1. Strictly valid JSON format only
2. No additional text outside the JSON structure
3. All 6 communities must be included
4. Use double quotes for all strings
5. Escape special characters properly
6. Response must strictly follow the [Output Format]
[Output Format:]
Required JSON Output Format:

```
{"Interest Community Scores": [
    {"Community": "Entertainment",
    "Score": "Score 1",
    "Reasoning": "Reasoning 1"},
    {"Community": "Technology",
    "Score": "Score 2",
    "Reasoning": "Reasoning 2"},
    ...
    {"Community": "Education",
    "Score": "Score 6",
    "Reasoning": "Reasoning 6"}
]}
```

B.3 Trust Threshold Prompt

We assess users' trust thresholds when encountering disinformation across different communities based on the semantic features of their historical posts, reposts, and comments, as well as their publicly available basic attributes. Our prompt design fully considers multiple dimensions of user-generated content, including linguistic style and sentiment tendencies, source citations, interaction patterns, and fundamental user attributes, ensuring a comprehensive and precise evaluation. The following is a detailed Prompt case.

Trust Threshold Evaluation Prompt

Please evaluate the user's trust threshold when encountering disinformation across different communities — Entertainment, Technology, Sports, Business, Politics, and Education — based on [User Data], including historical posts, reposts, quotes, and basic attributes. The trust threshold represents a user's expectation regarding the truthfulness of received information. Provide a trust threshold score (0-1 scale) for each community, where:
- [0.8-1.0] = Highly resilient to disinformation
- [0.5-0.79] = Moderately resilient to disinformation
- [0.0-0.49] = Potentially vulnerable to disinformation
[User Data:]
- Basic Attributes:
• Follower Count:{$follower\_count$}
• Following Count: {$following\_count$}
• Personal Description: {$personal\_description$}
- Content Data:
• Historical Posts: {$historical\_posts$}
• Historical Retweets: {$historical\_retweets$}
• Historical Quotes: {$historical\_quotes$}
[Evaluation Guidelines:]
1. Linguistic Style and Sentiment Tendencies:
- Critical thinking markers (e.g., "requires verification", "source needed")

Table 2: Parameter Definitions, Values, and Sources. In the "Value" column, values before the "/" represent the parameter range, and values after the "/" indicate the selected setting.

Module	Parameter	Value	Definition	Source
Attributes	$\mathcal{IC}_{uj}$	[1, 10]	The degree of interest of user u in community j	LLM
	$\mathcal{TT}_{uj}$	[0, 1]	The probability that user u perceives information j as disinformation.	LLM
	$\mathcal{DT}_{uj}$	[0, 1]	The tendency of user u to disseminate information j	Equation 1
	$\mathcal{SI}_{uj}$	[0, 1]	The social influence of user u in community j	Real data calculation
	$\mathcal{AT}_{ut}$	{0.1, ...}	The probability that user u is activated at time step t	Real data statistics
	θ	(0, 1)/0.5	The balance parameter of interest community and truncated power-law score	Predefined
	α	(1, $+\infty$)	The exponential part of the truncated power-law distribution	Real data fitting
	λ	(0, $+\infty$)	The truncation intensity of the truncated power-law distribution in Eq. 1	Real data fitting
	T	[1, 72]	The length of the time step in the simulation	Predefined
Disinfo	$\mathcal{DP}_j$	[0, 1]	The plausibility score of the disinformation j	LLM
Bot List	$\mathcal{MB}_j$	0.15 * N	The number of malicious bots in the dissemination network	Predefined
	$\mathcal{LR}_j$	0.05 * N	The number of legitimate bots in the dissemination network	Predefined
	$\mathcal{MF}_j$	[1, 18]	The number of times the malicious bot is activated within the total simulated time step	Predefined
	$\mathcal{LF}_j$	[1, 12]	The number of times the legitimate bot is activated within the total simulated time step	Predefined
Network	N	689	The number of dissemination network nodes	Real data
	m_0	5	The initial number of fully connected nodes in the BAM process	Predefined
	m	[1, 4]	The number of edges introduced by newly added nodes in the BAM process	Predefined
	τ	8	The threshold for the allocation of communities that users are interested in	Predefined
Correct	$\mathcal{EI}$	[12, 72]	The time step range of early intervention	Predefined
	$\mathcal{MI}$	[36, 72]	The time step range for mid-term intervention	Predefined
	$\mathcal{LI}$	[48, 72]	The time step range of late intervention	Predefined
Quantify	γ	(0, 1)/0.5	The balance parameter between enhancement and decay term in Eq. 2	Predefined
	β	(0, 1)/0.5	The rate of change of the enhancement term in Eq. 2	Predefined
	δ	(0, 1)/0.5	The rate of change of the decay term in Eq. 2	Predefined
	F'_{kj}	[0, 1]	The persuasiveness of corrective info j towards user k in Eq. 2	LLM
	F_{kj}	[0, 1]	The persuasiveness of disinfo j towards user k in Eq. 2	LLM
	$\hat{\mathcal{TT}}_{uj}$	[0, 1]	The updated trust threshold of user u's repeated exposure in community j	Equation 2
	$\mathcal{DA}_{uj}$	[0, 1]	The probability that user u successfully identifies disinfo j	Equation 3
	$\mathcal{SR}_t^j$	[0, 1]	Proportion of users unexposed to disinfo j at time t	Simulation changing
	$\mathcal{ER}_t^j$	[0, 1]	Proportion of users having encountered disinfo j at least once at time t	Simulation changing
	$\mathcal{IR}_t^j$	[0, 1]	Proportion of spreaders believe disinfo j at time t	Simulation changing
	$\mathcal{UR}_t^j$	[0, 1]	Proportion of spreaders unbelieve disinfo j at time t	Simulation changing

- Sentiment polarity distribution (skepticism vs. credulity markers)
- Hedging language frequency

2. Education Level:

- Technical terminology accuracy
- Complex sentence ratio
- Academic reference frequency

3. Source Reliability:

- Authoritative source citation rate (gov/academia)
- Low-credibility source retweets

4. Basic Attributes:

- Personal descriptions, follower-to-following ratio analysis
- Disinformation report history

[Output Requirements:]

1. Strictly valid JSON format only
2. No additional text outside the JSON structure
3. All 6 communities must be included
4. Use double quotes for all strings
5. Escape special characters properly
6. Trust threshold (0-1 scale)
7. Response must strictly follow the [Output Format]

[Output Format:]

Required JSON Output Format:

```
{"Trust Threshold Scores": [
    {"Community": "Entertainment",
    "Score": "Score 1",
    "Reasoning": "Reasoning 1"},
    {"Community": "Technology",
    "Score": "Score 2",
    "Reasoning": "Reasoning 2"},
    ...
    {"Community": "Education",
    "Score": "Score 6",
    "Reasoning": "Reasoning 6"}
```

]}

B.4 Truncated Power-Law Distribution Fitting

Based on real-world data, we fit truncated power-law distribution functions $P_{uj} = C \cdot (x_u)^{-\alpha} e^{-\lambda}$, where x_u represents the total number of retweets and quotes by a user. We employ maximum likelihood estimation (MLE) to optimize the truncated power-law exponent α and the lower bound x_{min}, subsequently calculating the normalization constant C as $C = \frac{\alpha-1}{x_{min}^{1-\alpha}}$. The exponential term $e^{-\lambda}$ acts as a cutoff to prevent infinite values. Our analysis of the collected real user data yielded a fitted distribution $P_{uj}(x_u) = 1.006 \cdot x_u^{-1.146} e^{-0.006}$, $x_u \geq 16$. Recognizing biases in the real-world data, the fitted truncated power-law distribution function achieves higher accuracy for $x_i \geq 16$. To mitigate the impact of estimation errors for $x_i < 16$, we reduce the θ parameter in Equation 1.

B.5 Disinformation Plausibility Prompt

Disinfo Plausibility Evaluation

Excluding all background knowledge, rigorously evaluate the plausibility of the [Input] below by analyzing:
1. **Emotional Expression** (e.g., exaggerated language, manipulative appeals)
2. **Propaganda Strategies** (e.g., cherry-picking, strawman arguments)
3. **Information Framing** (e.g., bias via omission, misleading context)
4. **Logical Consistency**: (e.g., contradictions, unsupported claims, or logical fallacies)
[Input]: $\{Disinformation Text\}$
[Output Requirements:]
1. Strictly valid JSON format optionally
2. No additional text outside the JSON structure
3. Use double quotes for all strings
4. Response must strictly follow the [Output Format]
[Output Format]:
Required JSON Output Format:
{ "PlausibilityScore": [0-1],
"Reasoning": "brief explanation"}

B.6 Disinformation Dissemination Network Construction

We employ the Stochastic Block Model (SBM) and the Barabási–Albert Model (BAM) to construct a disinformation dissemination network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ that exhibits both scale-free properties and community structure. This approach ensures that the network follows to the preferential attachment mechanism while also reflecting the characteristic of dense intra-community connections and sparse inter-community links. The construction process consists of the following steps:

(1)**Node Community Assignment** (SBM Component): Given a network with N users and J communities, each user u is assigned to one or more communities based on their community interest scores $\mathcal{IC}_{uj}$, which are evaluated by LLMs, and a predefined threshold τ. Specifically:

- If $\mathcal{IC}_{uj} \geq \tau$, user u is assigned to community j, and thus belongs to the set of assigned communities $\{C_j\}$.
- Since a user's interest community scores $\mathcal{IC}_{uj}$ may exceed the threshold for multiple communities, they can be assigned to more than one community.

The threshold τ is designed to reflect the structural characteristic of strong intra-community connectivity and limited inter-community links, ensuring a realistic simulation of information diffusion patterns.

(2)**Initialize Intra-Community Connections** (BAM Component): Within each community j, we first construct a fully connected subgraph $\mathcal{G}_j = \{\mathcal{V}_j, \mathcal{E}_j\}$ for each community j, consisting of an initial set of m_0 seed nodes. New nodes u are then iteratively introduced into the community, forming connections according to:

- The preferential attachment rule of the BAM.
- Add new nodes u to community j based on the social influence $\mathcal{SI}_{ej}$ of the existing nodes $e \in \mathcal{V}_j$ in the propagation network, where $\mathcal{SI}_{ej}$ is defined as:

$$\mathcal{SI}_{ej} = \frac{F_e}{\sum_{i \in \mathcal{V}_j} F_i}$$

F_e represents the follower count (or another influence metric) of node e.

This process leads to the formation of a scale-free network in which high-degree nodes (i.e., influential users) act as hubs for information dissemination, while low-degree nodes resemble ordinary users' social interaction patterns.

(3) **Iterative Network Growth**:

The iterative addition of new nodes u continues until the community j reaches its designated size. The complete procedure for constructing the disinformation dissemination network is detailed in Algorithm 1.

B.7 Corrective Strategy Example

We use an example of disinformation in the political community to illustrate how two correction strategies can be employed to curb the spread of disinformation and guide the public toward accurate information.

Disinformation Case: "Trump said the House select committee that investigated the Jan. 6, 2021, Capitol attack deleted and destroyed all of the information that they collected over two years."

Fact-based Corrective Strategy: Correcting disinformation by conveying officially released statements.

Fact-based Corrective Strategy

Case 1" The House Select Committee that investigated the Jan. 6, 2021, attack on the U.S. Capitol publicly released a final 845-page report and more than 100 transcripts of testimony, along with memos, depositions and documents. Much of the committee's work remains available online.

Case 2: House Republicans in March 2024 released a report accusing the committee of suppressing and deleting some pieces of evidence, but they have not gone as far as Trump in claiming that "all" evidence was destroyed.

Narrative-based Corrective Strategy: Correcting disinformation by sharing personal experiences of relevant individuals.

Narrative-based Corrective Strategy

Case 1: The House committee members, consisting of seven Democrats and two Republicans, stated that certain videos and sensitive materials were excluded from the official record to protect witnesses. All other materials were properly preserved, including over 100 testimony transcripts, written depositions, and documents that remain publicly available online.

Case 2: "As a member of the committee, we were under tremendous pressure in those days," a Democratic MP recalled, with a hint of exhaustion. "In order to protect those who bravely stood up to testify, we had to make difficult decisions, and rule out extremely sensitive videos and materials. I always remember a young witness who trembled and burst into tears when she described the experience. Imagine what secondary harm would have caused her if those images were revealed? We must protect her, which is our most basic responsibility. As for other records, I can assure you that, including those over a hundred testimonies, and those publicly written records and documents, each kept intact, and anyone can check and learn the truth."

B.8 Intervention Timing Setting

Early Intervention: Early intervention initiates corrective actions shortly after the onset of disinformation spread, aiming to limit its initial propagation and exposure before it gains significant traction.

Mid-Stage Intervention: Mid-stage intervention implements corrective actions after some disinformation spread. This allows observation of initial diffusion patterns and targeted correction of affected individuals and influential spreaders to mitigate further amplification.

Late Intervention: Late intervention deploys corrective measures when disinformation has already been circulating within the network for an extended period. Although the disinformation may have established a stronger foothold by this stage, the late intervention still aims to: reduce its persistence, correct the beliefs, and mitigate long-term negative impacts.

B.9 Persuasiveness Evaluation Prompt

Persuasiveness Evaluation Prompt

You are a social network user. The summary of the content you have posted or forwarded and quoted in the past three days is as follows:
[history Info]: $\{history\_info\}$
Evaluate the persuasiveness of the given [Input] information for you. The evaluation must result in a normalized score between [0,1] with clear justification.
[Input]: $\{text\_information\}$
Your assessment should synthesize the following dimensions and give normalized scores between [0–1] and brief reasons:
1. Relevance: how well it fits with your interests, career, or topics of concern
2. Bias: Whether the information is highly consistent with your views or is deliberately biased
3. Novelty & Informativeness: Whether the new ideas or data are provided compared to existing knowledge
[Output Requirements:]
1. Assign a score of [0,1] points
2. Include both quantitative score and brief reasoning
3. Strictly valid JSON format only
4. No additional text outside the JSON structure
5. Use double quotes for all strings
6. Escape special characters properly
7. Response must strictly follow the [Output Format]
[Output Format:]
Required JSON Output Format:
{ "Score": [0,1],
"Reasoning": "brief_explanation" }

B.10 Simulation and Evaluation Algorithm

We present the simulation workflow for disinformation in the MADD framework in Algorithm 2. The susceptible ratio (SR), exposed ratio (ER), infection spreader ratio (IR), and uninfection spreader ratio (UR) are obtained by calculating the proportion of susceptible users (SU), exposed users (EU), infection spreaders (IS), and uninfection spreaders (US) as determined by Algorithm 2, out of the total number of users.

B.11 Interaction and Decision-Making of Agents

Interaction and Decision-Making Strategy of Malicious and Legitimate Bots: We assign each bot agent the same five categories of attribute information as human users, including a predefined interest community. In terms of propagation behavior, the bots exhibit the following characteristics: their trust threshold and dissemination tendency are both set to 1, ensuring that they always trust and actively propagate either disinformation (for malicious bots) or corrective information (for legitimate bots). Their social influence is randomly sampled to match the influence levels of certain human users, and their active time is determined based on the predefined time range specified in Table 2. The bots' action policy is straightforward: once their activation time is reached, malicious bots will proactively disseminate misinformation, while legitimate bots will distribute corrective content.

Interaction and Decision-Making Strategy of Human Users: During their activation time, human users receive and read information propagated by their neighboring nodes. Their decision-making process is governed: (1) their dissemination tendency determines whether they are willing to share the received information; and (2) their trust threshold is used to evaluate whether they trust the content. On the premise of confirming whether the information is trusted, users will choose one of two dissemination methods—either retweeting or quoting—to share the content. This strategy is designed to realistically simulate human interaction behaviors and cognitive judgment processes in the context of information diffusion.

C Experimental Results

C.1 MADD and Real-World Consistency Evaluation

Figure 9 demonstrates that the probability distribution of users' dissemination tendencies exhibits a distinct heavy-tailed characteristic: while the majority of users show low propagation activity, a small subset displays exceptionally high dissemination propensity. Although the relatively limited sample size may introduce some outliers, the overall distribution pattern remains consistent with previously documented user dissemination behaviors in the literature (Goel et al., 2016).

Figure 10 presents the distribution of users' trust thresholds toward disinformation. For all topic cat-

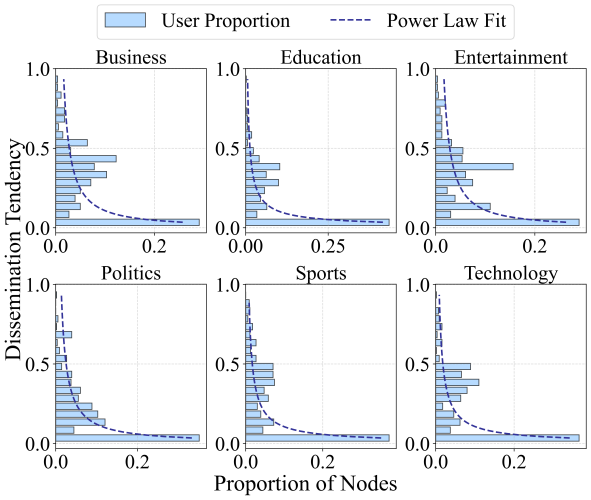


Figure 9: User dissemination tendency exhibits a heavy-tailed pattern.

egories, the trust threshold distribution follows an approximately normal pattern, indicating that most users adopt a neutral stance of "neither fully believing nor completely rejecting" towards disinformation, with trust threshold scores predominantly clustered around 0.5. This finding is consistent with established research outcomes (Ecker et al., 2022).

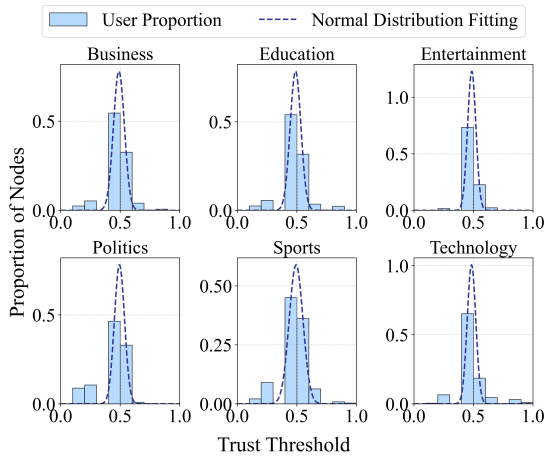


Figure 10: Trust threshold follow normal distribution fitting.

C.2 Group-Level: Effectiveness Evaluation of Correction Strategies

Figures 11 and 12 respectively illustrate the comparative intervention effects of two correction strategies (fact-based and narrative-based correction) implemented during the mid (T=[36-72]) and late (T=[48-72]) stages of disinformation spread. The results indicate that the intervention effects of both correction strategies are not significant for most communities. Notably, in some communities, such as "Politics" and "Business," we observe negative intervention effects, manifested as an increase in the spread of related disinformation after correction. This phenomenon may be because prolonged exposure to disinformation leads users to form stable incorrect cognitive frameworks, and the introduction of corrective information is interpreted by some users as a "threat to their original viewpoints," thereby triggering a psychological reactance effect (Diaz Ruiz and Nilsson, 2023) and forming an echo chamber effect (Garimella et al., 2018).

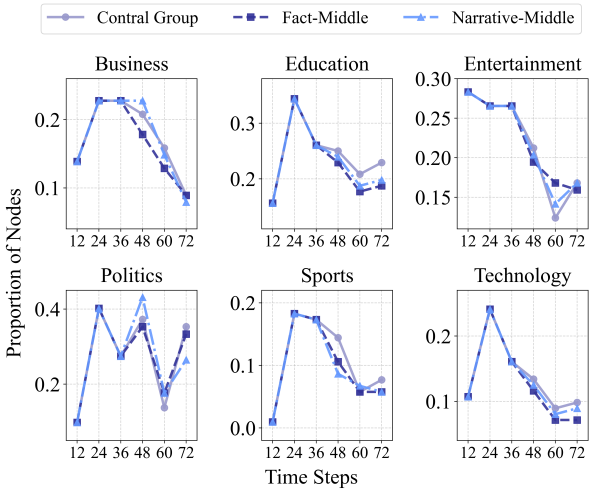


Figure 11: Comparison of mid intervention effects of two correction strategies.

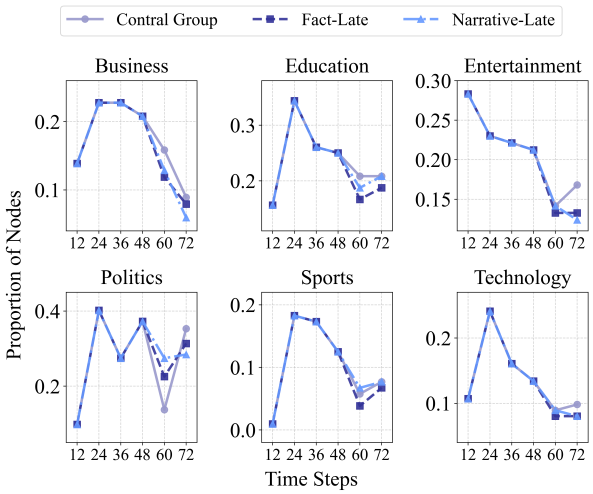


Figure 12: Comparison of late intervention effects of two correction strategies.

C.3 Individual-Level: Dynamic Changes in User Trust Thresholds

Figures 13 and 14 illustrate the changes in user trust thresholds when correction strategies are implemented in the mid and late stages. We observe that their effect on increasing user trust thresholds for evaluations is extremely limited. Furthermore, in the "Entertainment," "Business," and "Politics" communities during the mid-term intervention phase, and in the "Sports" community during the late-term intervention phase, the control group (without correction strategies) actually shows higher user trust thresholds than the experimental group (with correction strategies). This phenomenon likely stems from two main reasons: firstly, the intervention timing is delayed, and secondly, echo chamber effects may have already formed within some user groups. When disinformation has already spread widely across the network and deeply influenced users' cognitive belief systems, correction strategies introduced at this point often struggle to effectively reverse established negative perceptions. Therefore, these findings strongly confirm the critical importance of intervening in the early stages of disinformation spread and implementing effective correction strategies.

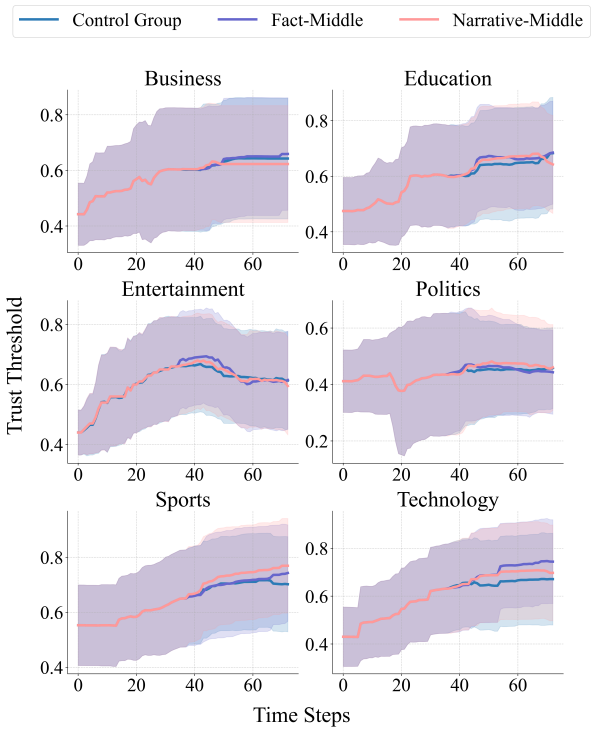


Figure 13: Effects of Correction Strategies on User Trust Thresholds During Mid-Stage Intervention

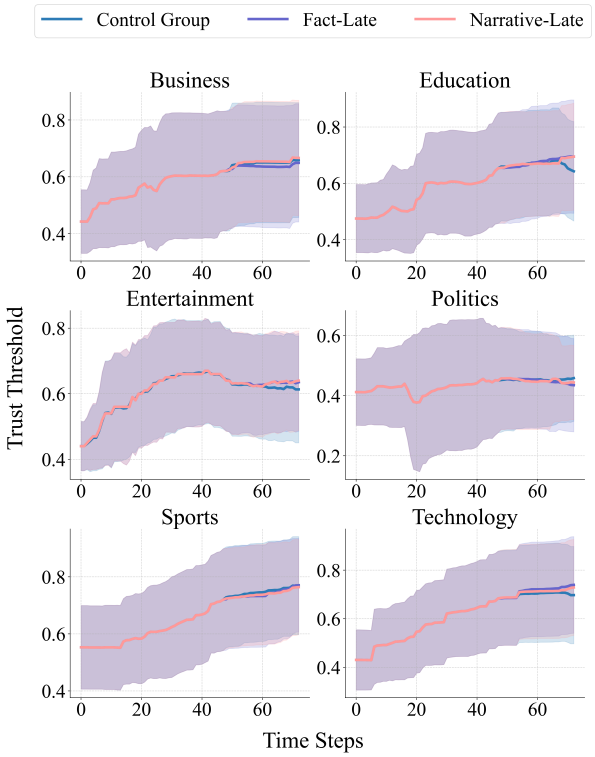


Figure 14: Effects of Correction Strategies on User Trust Thresholds During Late-Stage Intervention

C.4 Case Study: Change of the Trust Threshold of the Case User

In Figure 15, we show an example of a user's Trust Threshold over time under a fact-based early intervention strategy. Overall, the trust threshold undergoes a dynamic change: first rising sharply, then declining, and finally stabilizing.

The initial rapid increase is primarily because the user receives highly credible corrective information from high-influence nodes, which significantly enhances their ability to discern information. However, as disinformation gradually increases in the network, the user's trust threshold begins to decline slowly, reflecting some cognitive interference. As legitimate bots continue their correction efforts in the later stages, the proportion of true information in the user's environment increases, causing the trust threshold to rise again and eventually stabilize. This indicates that the user has developed a strong resistance to disinformation on this topic.

C.5 Effect of Legitimate Bot Ratio on Correction

Figure 16 illustrates the infection rate dynamics from time steps 12 to 72 under a fact-based early intervention strategy with varying proportions of

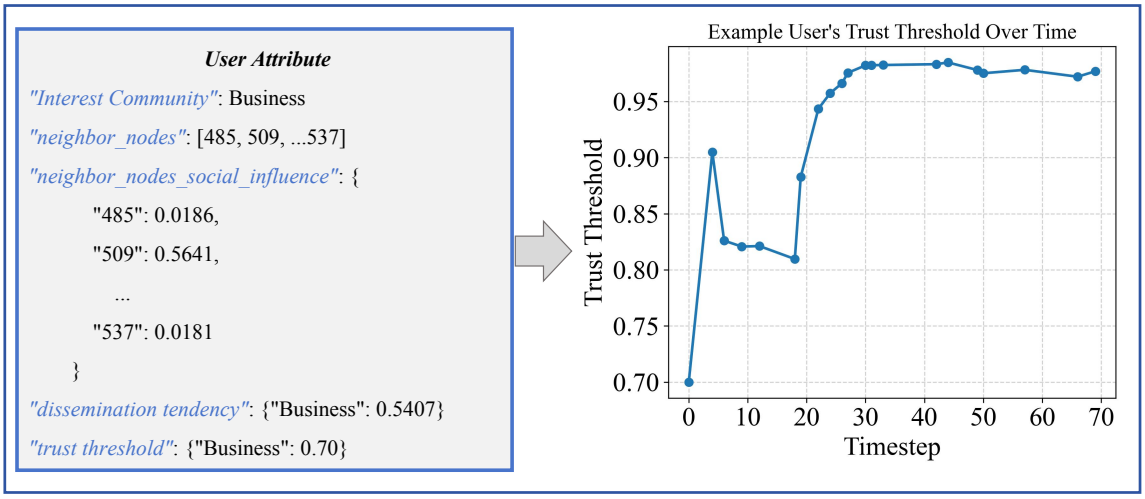


Figure 15: The changing trend of the trust threshold of the case users

legitimate bots.

Overall, the infection rate consistently declines over time, but the speed and scale of this decline are directly tied to the strength of the intervention.

- **No Intervention (0% L-Bots):** Without any legitimate bots, the infection rate remains high (around 22.7%) and declines slowly. This highlights that users' natural ability to correct disinformation is limited when there is no active intervention.
- **With Intervention (5% to 30% L-Bots):** As the proportion of legitimate bots increases from 5% to 30%, the infection rate drops much faster during the middle stages (time steps 24 to 48), proving the effectiveness of stronger corrective intervention.

Interestingly, the infection rates across all three intervention scenarios converge in the later stages (time steps 48 to 72). This is likely because the remaining infected users have low initial confidence, making them less susceptible to change. Furthermore, since bot intervention is concentrated in the early stages, the later dynamics are driven more by users' own cognitive adjustments, leading to a similar rate of change across all groups.

C.6 System Resources Analysis

We present in Table 3 the key metrics for the control group (simulation involving only malicious bots), including the number of LLM calls, inference time, and token usage. Since the simulation relies on external LLM APIs, response speeds may vary,

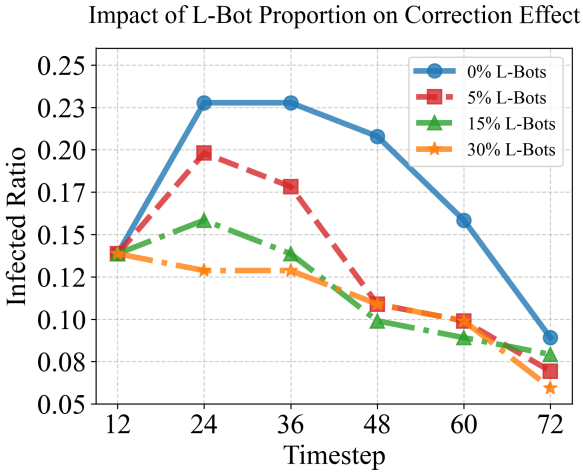


Figure 16: Effect of Legitimate Bot Ratio on Correction

and inference time is not strictly proportional to the number of calls.

D Licenses

All source code and scripts associated with this paper are released in our public repository at github.com/QQQQQQBY/BotInfluence. The repository includes a standard MIT License, reproduced in the `LICENSE` file. This license permits use, modification, and redistribution under the specified terms.

Algorithm 2 Simulation and Evaluation Process

Inputs: Time steps $\mathcal{T}$, Disinfo set $\mathcal{J}$, Corrective info set $\mathcal{C}$, Malicious Bots $\mathcal{MB}$, Legitimate Bots $\mathcal{LB}$, Regular Users $\mathcal{RU}$, Active users at time t, $\mathcal{AC}_t$, Initial received/shared info ($\mathcal{R}_u$, $\mathcal{S}_u$ for each user u).
Outputs: Trust thresholds $\mathcal{TT}$, Susceptible users $\mathcal{SU}$, Exposed users $\mathcal{EU}$, Infection Spreaders $\mathcal{IS}$, Uninfection Spreaders $\mathcal{US}$.
Initialize: $\mathcal{SU} \leftarrow \mathcal{RU}, \mathcal{EU} \leftarrow \emptyset, \mathcal{IS} \leftarrow \emptyset, \mathcal{US} \leftarrow \emptyset$
Start Simulation:
for disinfo $j \in \mathcal{J}$ **do**
  for each time $t \in \mathcal{T}$ **do**
    for each active user $u \in \mathcal{AC}_t$ **do**
      1. Load Share Info:
      if $u \in \mathcal{MB}$ **then**
        Spread disinfo $\hat{info} = j$ to neighbors $\mathcal{N}_u$
      else if $u \in \mathcal{LB}$ **then**
        Spread corrective info $\hat{info} = c, c \in \mathcal{C}$ to neighbors $\mathcal{N}_u$
      else if $(u \in \mathcal{RU}) \wedge (\mathcal{R}_u \neq \emptyset)$ **then**
        $info \leftarrow$ Lastest received info ($\mathcal{R}_u$)
        if random( ) < $\mathcal{DT}_{u,j}$ **then**
          Action $\hat{info} \leftarrow \mathcal{LLM}(info, history, ActionList)$
        end if
        2. Share Info:
        Share $\hat{info}$ to neighbors $\mathcal{N}_u$, update $\mathcal{S}_u \leftarrow \mathcal{S}_u \cup \{\hat{info}\}$
      end if
      for each neighbor $v \in \mathcal{N}_u \cap \mathcal{RU}$ **do**
        3. Receive Info:
        Update $\mathcal{R}_v \leftarrow \mathcal{R}_v \cup \hat{info}$
        Discern Ability $\mathcal{DA}_{v,t}^{\hat{info}} \leftarrow \mathcal{LLM}$(Check Belief($\mathcal{TT}_u, \mathcal{DP}_{\hat{info}}$)) based on Eq. 3
        Update $\mathcal{TT}_{v,j}$ via:
          $\mathcal{TT}_{v,j} \leftarrow$ Eq. 2 Dynamic adjustment
          Update state sets $\mathcal{SU}_{t,j}, \mathcal{EU}_{t,j}, \mathcal{IS}_{t,j}, \mathcal{US}_{t,j}$
      end for
    end for
  end for
end for

return $\mathcal{SU}, \mathcal{EU}, \mathcal{IS}, \mathcal{US}, \mathcal{TT}$

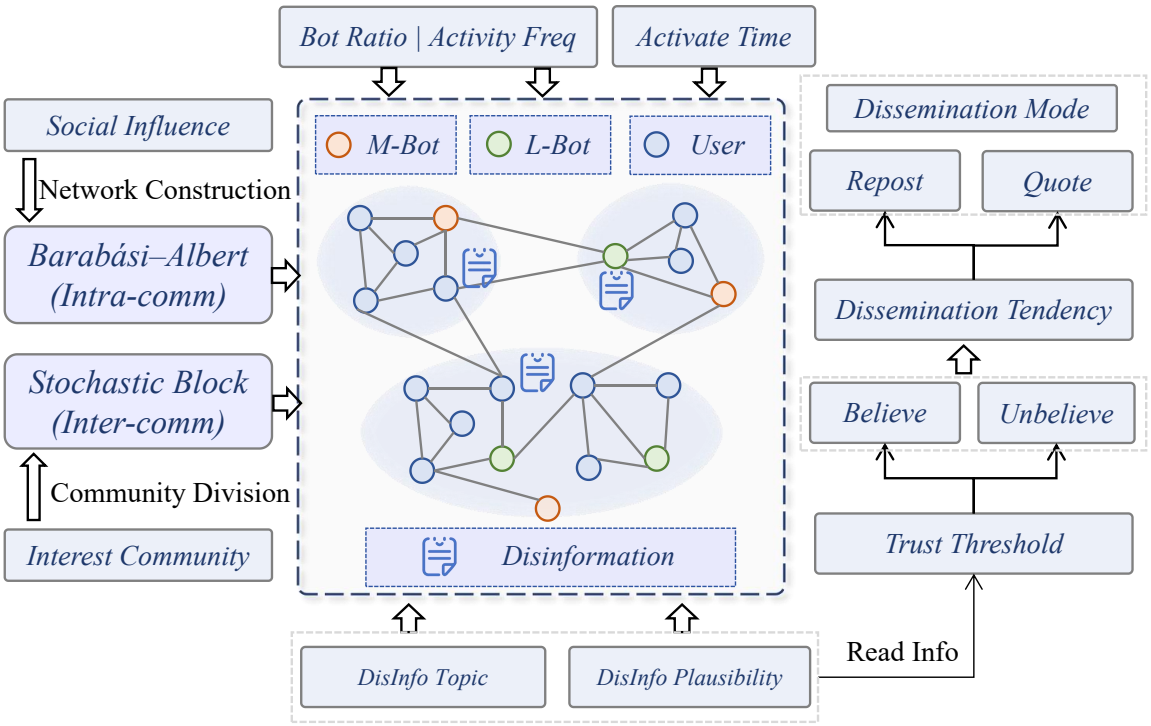


Figure 17: Supplementary expansion of Figure 2. Attribute-driven network dynamics and disinformation dissemination in MADD.

Table 3: Statistics on the consumption of simulated resources for each community (LLM call volume, inference time, Token consumption)

Metric	Politics	Business	Education	Entertainment	Sports	Technology
LLM Calls / Avg. per User	374,212 / 366.9	275,478 / 272.8	259,344 / 2,702	325,702 / 2,882	256,483 / 2,466	234,723 / 2,096
Inference Time / Avg. per User	19.02 h / 671 s	17.53 h / 625 s	16.82 h / 631 s	14.73 h / 469 s	12.59 h / 435 s	12.27 h / 394 s
Token Usage / Avg. per User	9,983,935 / 97,881	8,773,767 / 86,869	7,684,069 / 80,042	9,893,845 / 87,556	7,115,933 / 68,422	7,503,570 / 66,996