

Misinformation Mitigation over Social Networks: a Control Approach

Nicolò Pagan

*Institut für Informatik
University of Zurich
Zurich, CH*

NICOLO.PAGAN@UZH.CH

Andreas Philippou

*Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands*

A.PHILIPPOU@STUDENT.TUE.NL

Giulia De Pasquale

*Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands*

G.DE.PASQUALE@TUE.NL

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Modern social networks rely on recommender systems that inadvertently amplify misinformation by prioritizing engagement over content veracity. We present a control framework that mitigates misinformation spread while maintaining user engagement by penalizing content characteristics commonly exploited by false information, specifically, extreme negative sentiment and novelty. We extend the closed-loop Friedkin-Johnsen model to incorporate the mitigation of misinformation alongside engagement maximization. Both model-free and model-based control strategies demonstrate up to 76% reduction in misinformation propagation across diverse network configurations, validated through simulations using the LIAR2 dataset with sentiment features extracted via large language models. Analysis of engagement-misinformation trade-offs reveals that in networks with radical users, median engagement improves even as misinformation decreases, suggesting content moderation enhances discourse quality for non-extremist users. The framework provides practical guidance for platform operators in balancing misinformation suppression with engagement goals.

Keywords: Misinformation mitigation, recommender systems, opinion dynamics, Friedkin-Johnsen model, model predictive control, social networks

1. Introduction

Modern social networks connect billions of users but simultaneously create conditions that facilitate rapid misinformation spread (Del Vicario et al., 2016). The societal consequences, affecting democratic processes, public health, and social cohesion (Cinelli et al., 2020; Persily, 2017), have intensified as research confirms misinformation spreads faster than truth (Vosoughi et al., 2018).

Current mitigation strategies focus on truthfulness through fact-checking and machine learning (Shu et al., 2017), yet overlook the psychology behind virality. Misinformation strategically exploits emotional triggers, particularly negative emotions (Brady et al., 2017), and novelty (Berger and Milkman, 2012) to achieve virality. Recommender systems designed to maximize engagement

can amplify such content, creating feedback loops that reinforce echo chambers and polarization (Del Vicario et al., 2015; Mansoury et al., 2020; Pagan et al., 2023; Lanzetti et al., 2023).

Recent advances model recommender systems as control inputs within opinion dynamics frameworks Dean et al. (2024); Rossi et al. (2022); Chandrasekaran et al. (2024); Dean and Morgenstern (2022); Mansoury et al. (2020); Sprenger et al. (2024). In particular, Sprenger et al. (2024) developed a closed-loop Friedkin-Johnsen model where engagement-maximizing recommendations fundamentally alter network opinion evolution. However, research explicitly addressing misinformation within this control framework remains limited.

We adapt this framework to model *sentiment propagation* rather than topical opinion, motivated by evidence that misinformation spreads through emotional manipulation. We modify the engagement objective to penalize extreme negative sentiment and novelty, characteristics misinformation exploits for virality, while maintaining user engagement. Both model-free and model-based strategies are developed with convergence guarantees, validated on the LIAR2 dataset (Xu and Kechadi, 2024) using large language models for sentiment extraction.

Our analysis demonstrates up to 76% misinformation reduction across network configurations including radicalized environments with stubborn extremist users. We reveal engagement-misinformation trade-offs: while mean engagement may decrease, median engagement in radical networks *improves*, indicating enhanced discourse quality for non-extremist majorities. The optimal region provides actionable guides for platform operators balancing content moderation with business goals.

2. Methods

We present a control framework for misinformation mitigation through recommender systems. The model dynamics and control-loop formulation (Section 2.1) follow Sprenger et al. (2024); we refer readers there for foundational details. Section 2.2 introduces our modified cost function incorporating psychological factors associated with misinformation spread. Section 2.3 formulates model-free and model-based control strategies. Section 2.4 provides convergence proofs.

2.1. Model Dynamics and Control Formulation

We adopt the closed-loop Friedkin-Johnsen framework from Sprenger et al. (2024), representing users as nodes, with system state $\mathbf{x}(t) \in [0, 1]^n$, at discrete time t , whose $i - th$ entry represents the opinion state of the i -th node (user). The recommender acts as an additional node influencing users through control input $u(t) \in [0, 1]$. This entails that the recommendation system suggests content in the range $[0, 1]$ on a given topic per time instant to all the connected users, which will influence their opinions. Recommendation systems of this type are widely used in real-world social media. In fact, users are grouped into pools based on shared preferences and demographic characteristics, and each group is shown the same set of recommended items. Importantly, even individuals who do not directly watch the channel may still be influenced indirectly through interactions with other users. While Sprenger et al. (2024) model opinion agreement, we employ the same structure to model sentiment propagation, motivated by evidence that emotional content drives misinformation virality (Brady et al., 2017; Vosoughi et al., 2018). In our formulation, $x_i(t) \in [0, 1]$ represents user i 's sentiment intensity, where $x_i(t) = 0$ corresponds to neutral/positive sentiment and $x_i(t) = 1$ to highly emotional content. The control input $u(t)$ represents recommended content sentiment.

The network is represented by row-substochastic adjacency matrix $\mathbf{W}_{\text{total}} \in [0, 1]^{(n+1) \times (n+1)}$, partitioned into user-to-user interactions $\mathbf{W} \in [0, 1]^{n \times n}$ and recommender-to-user influence $\mathbf{w}_{\text{rec}} \in$

$[0, 1]^n$. Dynamics evolve as:

$$\mathbf{x}(t+1) = (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{W}\mathbf{x}(t) + (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{w}_{\text{rec}}u(t) + \mathbf{\Lambda}\mathbf{x}(0),$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n) \in [0, 1]^{n \times n}$ is the stubbornness matrix with $\lambda_i \in [0, 1]$ quantifying user i 's resistance to influence ($\lambda_i = 0$: full susceptibility; $\lambda_i = 1$: complete adherence to $x_i(0)$). In compact form,

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) + \mathbf{\Lambda}\mathbf{x}(0), \quad (1)$$

where $\mathbf{A} = (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{W}$ and $\mathbf{B} = (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{w}_{\text{rec}}$. This linear time-invariant structure enables optimal control design for $u(t)$ balancing engagement with misinformation mitigation by penalizing high sentiment intensity.

2.2. Cost Function Modification for Misinformation Mitigation

We modify the engagement objective from [Sprenger et al. \(2024\)](#) to incorporate misinformation mitigation while acknowledging that platforms fundamentally rely on engagement. The original engagement cost the recommender system in [Sprenger et al. \(2024\)](#) minimizes, namely

$$\theta(\mathbf{x}(t), u(t)) = \|\mathbf{x}(t) - u(t)\mathbf{1}_n\|_2^2, \quad (2)$$

measures squared Euclidean distance between user states and recommendations and, based on confirmation bias, promotes recommended content $u(t)$ that more closely aligns with users' opinions. We augment this with penalties for extreme sentiment intensity $E(u(t)) = u(t)^2$ and novelty modulation $N(t, t_c) = e^{-\delta(t-t_c)}$ for $t - t_c \leq z$, where $\delta > 0$ controls decay rate, t_c is content creation time, and z defines the content window. Novel content, which misinformation exploits ([Berger and Milkman, 2012](#)), receives higher initial penalty that diminishes over time. The modified cost accounting for misinformation mitigation is:

$$\theta_M(\mathbf{x}(t), u(t)) = \theta(\mathbf{x}(t), u(t)) + \rho n \cdot u(t)^2 \cdot e^{-\delta(t-t_c)}, \quad t - t_c \leq z, \quad (3)$$

where $\rho \geq 0$ controls penalty strength and n ensures consistent scaling across network sizes. The recommender minimizes $\sum_{t=0}^{\infty} \theta_M(\mathbf{x}(t), u)$. For the sake of mathematical tractability, analytical results are developed under the assumption that the content creation time t_c , is fixed.

2.3. Model-free and Model-based Approaches

Model-free (MF). The MF approach $u_{\text{MF}}(t) = \arg \min_{u \in [0, 1]} \theta_M(\mathbf{x}(t), u)$ minimizes θ_M at time t using only the current users' sentiment vector $\mathbf{x}(t)$.

Model-based (MB). The MB approach adopt a model predictive control (MPC) strategy. The theoretical optimal steady-state $(\mathbf{x}_{\text{MB}}^*, u_{\text{MB}}^*) = \arg \min_{\mathbf{x}, u} \theta_M(\mathbf{x}, u)$ subject to $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}u + \mathbf{\Lambda}\mathbf{x}(0)$ and $u \in [0, 1]$ always exists (Section 2.4). The full MPC formulation is:

$$\mathcal{O}_t^* := \min_{\mathbf{x}_{\cdot|t}, u_{\cdot|t}} \sum_{k=0}^{T-1} \theta_M(\mathbf{x}_{k|t}, u_{k|t}) \quad (4)$$

$$\text{s.t. } \mathbf{x}_{k+1|t} = \mathbf{A}\mathbf{x}_{k|t} + \mathbf{B}u_{k|t} + \mathbf{\Lambda}\mathbf{x}(0), \quad \mathbf{x}_{0|t} = \mathbf{x}(t), \quad \mathbf{x}_{T|t} = \mathbf{x}_{\text{MB}}, \quad u_{k|t} \in [0, 1], \quad \forall k \in [0, T-1]$$

where \mathcal{O}_t^* is the general optimization cost function and T is the prediction horizon. The optimizer output $u_{\text{MB}}(t) = u_{0|t}$ is the first element of the MPC solution. The MF and MB approaches have significant differences and scopes of informational access. Unlike the MF approach, the MB approach must have access to the opinion dynamic dependencies \mathbf{A} , \mathbf{B} and the resilience matrix $\mathbf{\Lambda}$, which are hardly measurable. In our simulations, these parameters are treated as known design parameters set *a priori* to define the network scenarios under study (see Table 1), rather than being estimated from data. This reflects the idealized setting in which MB operates and deliberately places it at an advantage relative to real-world deployment conditions. The near-identical performance of MF and MB observed throughout Section 4 therefore strengthens the case for the MF strategy, which requires no such system knowledge.

2.4. Mathematical Analysis and Convergence Proofs

The new cost function definition in (3) requires revisiting the control strategies and their convergence properties. While the Friedkin-Johnsen model structure and graph properties from Sprenger et al. (2024) remain unchanged, the modified optimization problem necessitates new analysis.

2.4.1. CONVERGENCE ANALYSIS AND STEADY-STATE SOLUTIONS

We now derive the steady-state solutions for both the model-free and model-based approaches and establish convergence guarantees. The proofs closely follow Sprenger et al. (2024), with modifications to accommodate the penalty terms in θ_{M} .

Model-Free Steady State. The optimal MF control is obtained by minimizing $\theta_{\text{M}}(\mathbf{x}(t), u(t))$ with respect to u at each time step. Given θ_{M} is convex, taking $\frac{\partial \theta_{\text{M}}}{\partial u} = 0$ and solving yields:

$$u_{\text{MF}}(t) = \frac{\sum_{i=1}^n x_i(t)}{n(1 + \rho \cdot e^{-\delta(t-t_c)})}.$$

Substituting this into (1) gives the closed-loop dynamics:

$$\mathbf{x}(t+1) = (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{F}(t)\mathbf{x}(t) + \mathbf{\Lambda}\mathbf{x}(0),$$

where $\mathbf{F}(t) = \mathbf{W} + \frac{\mathbf{w}_{\text{rec}}\mathbf{1}_n^T}{n(1+\rho \cdot e^{-\delta(t-t_c)})}$. The matrix $\mathbf{F}(t)$ is sub-row stochastic and, given that the creation time t_c , is assumed to be fixed, the matrix $\mathbf{F}(t)$ becomes constant as the term $e^{-\delta(t-t_c)}$ vanishes, and convergence conditions established in Sprenger et al. (2024) apply. At steady state, $\mathbf{x}(t) = \mathbf{x}(t+1)$, which yields:

$$\mathbf{x}_{\text{MF}}^* = \left(\mathbf{I}_n - \mathbf{A} - \frac{\mathbf{B} \cdot \mathbf{1}_n^T}{n} \right)^{-1} \mathbf{\Lambda}\mathbf{x}(0).$$

Model-Based Steady State.

For the MB approach, we solve the constrained optimization problem (4) using the Karush-Kuhn-Tucker conditions. The interior solution ($0 < u < 1$) is:

$$u_{\text{MB}}(t) = \frac{\mathbf{1}_n^T \mathbf{y} - \mathbf{v}^T \mathbf{y}}{-\mathbf{1}_n^T \mathbf{v} + n + \mathbf{v}^T \mathbf{v} - \mathbf{v}^T \mathbf{1}_n + \rho n e^{-\delta(t-t_c)}} \rightarrow u_{\text{MB}}^* = \frac{\mathbf{1}_n^T \mathbf{y} - \mathbf{v}^T \mathbf{y}}{-2\mathbf{1}_n^T \mathbf{v} + n + \mathbf{v}^T \mathbf{v}},$$

$$\mathbf{x}_{\text{MB}}^* = \mathbf{v}u_{\text{MB}}^* + \mathbf{y},$$

where $\mathbf{v} = (\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}$ and $\mathbf{y} = (\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{\Lambda}\mathbf{x}(0)$.

Convergence Guarantees. The convergence proof for both approaches follows the same structure as [Sprengr et al. \(2024\)](#). The key modification is in the matrix:

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_n & -\mathbf{1}_n \\ -\mathbf{1}_n^T & n(1 + \rho e^{-\delta(t-t_c)}) \end{bmatrix},$$

which replaces their corresponding matrix in the Lyapunov stability analysis. All other proof steps remain identical, and we refer readers to [Sprengr et al. \(2024\)](#) for the complete argument. The penalty terms ρ and $e^{-\delta(t-t_c)}$ preserve the positive definiteness of \mathbf{H} required for convergence, provided $\rho \geq 0$ and $\delta > 0$.

3. Simulation Setup

This section details the simulation design, network configurations, and datasets used to evaluate the proposed misinformation mitigation framework. Two types of networks are considered: a large-scale synthetic network of 100 agents, and a small radicalized network of 6 agents adapted from [Sprengr et al. \(2024\)](#). Each network is tested using both synthetic continuous content and real-world data from the LIAR2 dataset.

3.1. Network Configurations

Network A — 100-agent synthetic network. This network models a general social platform with $n = 100$ users. Network parameters are given in Table 1. W is randomly generated with 25% edge density, weights drawn from $\mathcal{U}(0, 1)$ and row-normalized. The recommender is connected to 80% of users. Stubbornness values Λ are drawn from $\mathcal{U}(0, 0.05)$, yielding a weakly stubborn, moderately connected population. The initial opinion (sentiment) values are drawn from a beta distribution of parameters $\alpha = 7, \beta = 2$, which is skewed toward higher sentiment values (more emotional intensity), reflecting populations where mildly negative content dominates engagement.

Network B — 6-agent radicalized network. The network replicates the structure of [Sprengr et al. \(2024\)](#), with fixed weights and topology. User 5 is a radical agent ($\lambda_5 = 1$), fully anchored to their prior opinion and disconnected from the recommender. The remaining users have low stubbornness ($\lambda_i \leq 0.092$). The initial opinions are defined as the complement of those used in the original paper, so that the most stubborn user now holds an extreme negative opinion of 1. Specifically, $\mathbf{x}(0) = [0.33, 0.26, 0.17, 0.32, 1.00, 0.41]$. This modification ensures that the controller faces the more challenging task of mitigating an entrenched source of negativity. Unlike the large-scale network, this setup runs for $\tau = 50$ time steps, which is sufficient for convergence.

3.2. Simulation Scenarios

Each network is simulated under three configurations:

1. **Model-Free (MF)** without mitigation ($\rho = 0$), optimizing only user engagement θ , as in [Sprengr et al. \(2024\)](#).
2. **Model-Free (MF)** with mitigation ($\rho > 0$), introducing the misinformation penalty, θ_M .
3. **Model-Based (MB)** with mitigation ($\rho > 0$), using the predictive control formulation.

Table 1: Simulation Parameters for Both Network Configurations

Parameter	Description	Value
n	Number of users	100 (Network A) / 6 (Network B)
Λ_h	Highest stubbornness	0.05
Λ_l	Lowest stubbornness	0.00
κ_u	User-to-user connectivity	0.25
κ_r	Recommender-to-user connectivity	0.80
τ	Time steps	100 (A) / 50 (B)
ρ	Penalty strength regulator	[0.00, 5.50] (step 0.10)
T	MPC prediction horizon	50
z	Eligible content window	5
δ	Novelty decay rate	0.00

For all cases, we consider both synthetic continuous $u(t) \in [0, 1]$ and discrete data-driven content values described below.

3.3. Data-based Sentiment Extraction

To simulate realistic content, we use the LIAR2 dataset (Xu and Kechadi, 2024), which contains 4000 labeled statements (2000 true and 2000 false) from social media and news sources. For each statement, we compute a *emotional extremity score* $C(l) \in [0, 1]$ using a transformer-based natural language processing (NLP) model (Mistral NeMo). The model analyzes each text along six affective and linguistic dimensions—fear, disgust, anxiety, shock, overall negative sentiment, and subjectivity—and aggregates them into a single weighted score by giving equal weight (0.15) to each emotional dimension and slightly higher weight (0.20) to overall negativity and subjectivity. Higher $C(l)$ values correspond to content with stronger emotional tone or higher potential for emotional manipulation. The Mistral NeMo model is chosen for its efficiency, accuracy, and open-source availability. At each simulation time step, a random subset of the 4000 LIAR2 statements is made available to the recommender. On average, this corresponds to about 40 new pieces of content per step in the 100-agent network and about 80 in the 6-agent network. The appearance times are uniformly distributed to emulate a continuous stream of new posts. Consistent with prior research on misinformation virality, false statements exhibit higher average emotional intensity (mean $C(l) = 0.537$) compared to true statements (mean $C(l) = 0.379$).

3.4. Misinformation and Behavioral Metrics

To evaluate the effectiveness of the proposed mitigation strategies, we monitor three complementary quantities. *i*) First, we compute the *misinformation metric* as the ratio of false news, $\mathcal{M} = \frac{\#\text{falsenews}}{\#\text{news}}$, A lower value of \mathcal{M} indicates stronger suppression of misinformation exposure. *ii*) Second, we quantify the overall *sentiment shift* as the absolute change in emotional extremity between the final and initial states, i.e. the mean (and median) of $|x_i(\tau) - x_i(0)|$ across users. This captures how much individual sentiment evolves during the simulation.

iii) Finally, we track user *engagement*, defined as the per-user average of the instantaneous engagement cost introduced in (2), averaged over time. Higher engagement values correspond to stronger alignment between user sentiment and recommended content.

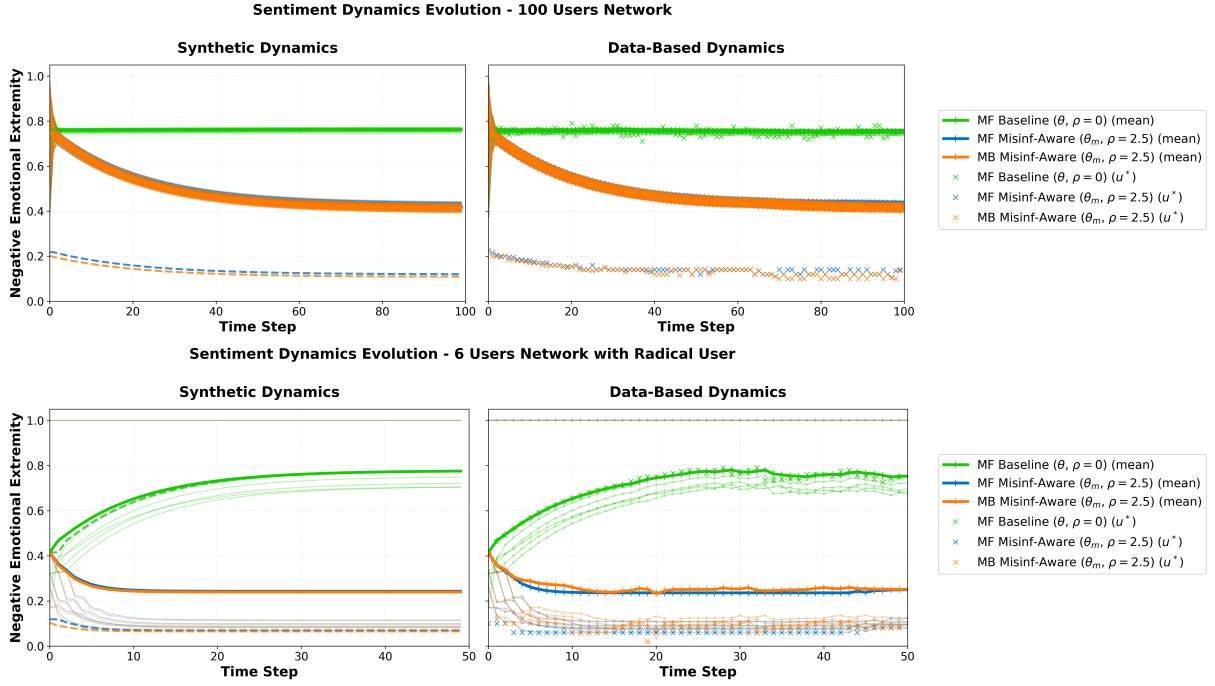


Figure 1: Sentiment Dynamics Evolution. **Top**: 100-User Network showing mean emotional extremity (solid) and recommender output (dashed) over 100 time steps, comparing baseline (θ , green), MF mitigation (θ_M , blue), and MB mitigation (θ_M , orange) at $\rho = 2.5$. Left: synthetic; Right: data-driven. Shaded regions indicate standard deviation. **Bottom**: 6-User Network with Radical User, same comparison over 50 time steps. Individual trajectories shown in light lines. Mitigation prevents negativity propagation beyond the stubborn radical user.

4. Results

4.1. Misinformation Mitigation Alleviates Users’Emotional Extremity

Figure 1 shows the evolution of user emotional extremity under different control strategies for both network configurations. In both networks, the baseline engagement-only model (θ) drives average sentiment toward more negative values, while the mitigation-aware controllers (θ_M) stabilize user states closer to neutrality. In the 100-agent network (Figure 1, top), the average emotional extremity converges rapidly to a steady moderate value. The MF and MB mitigation strategies exhibit nearly identical behavior, both successfully preventing the negative drift observed in the baseline case. This holds true for both synthetic continuous dynamics (left panels) and data-driven discrete content selection (right panels), demonstrating robustness across simulation modalities. This phenomenon was also observed in [Sprenger et al. \(2024\)](#), and is most likely due to the fact that the state matrices in the MB and MF strategies share the same dominant eigenvalue as they are both row stochastic. In the 6-agent network (Figure 1, bottom), the presence of a stubborn radical user anchored at $x_i = 1$ (maximum emotional extremity) causes the overall mean sentiment to remain relatively high, yet still substantially less extreme than without mitigation. Notably, the mitigation controllers prevent the radical user’s negativity from propagating to connected users, maintaining their sentiment at moderate levels despite the persistent influence of the extremist node.

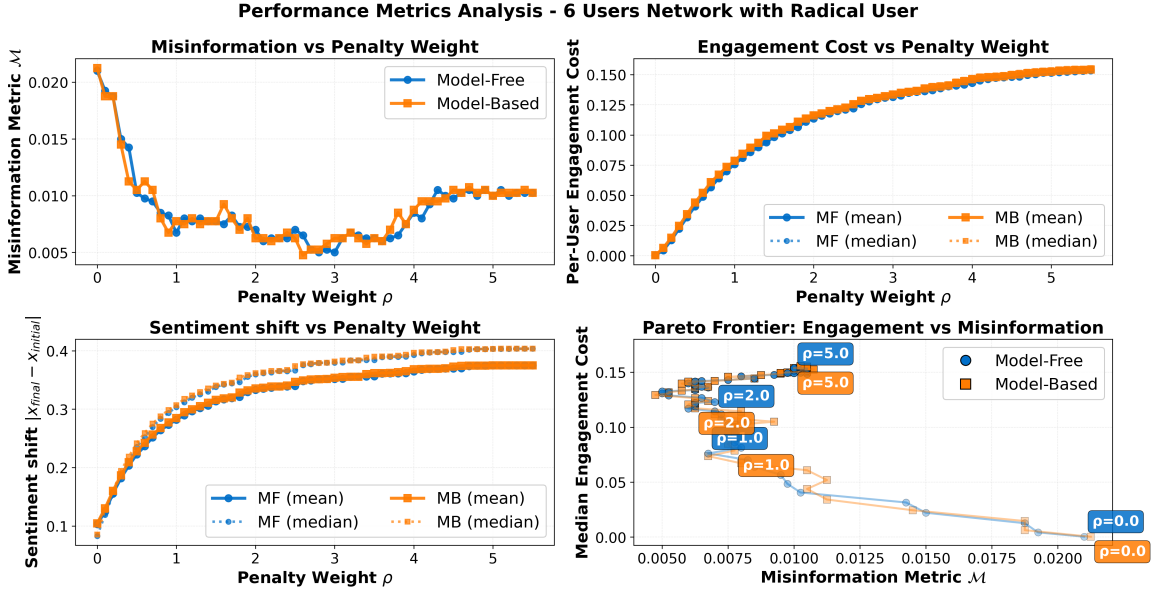


Figure 2: Performance Metrics Analysis for 100-User Network (Data-Driven). **Top-left:** Misinformation metric \mathcal{M} vs. penalty weight ρ . **Top-right:** Per-user engagement cost vs. ρ (mean and median). **Bottom-left:** Sentiment shift $|x_i(\tau) - x_i(0)|$ vs. ρ (mean and median). **Bottom-right:** Pareto frontier showing trade-off between median per-user engagement cost and misinformation (lower-left is better). Labeled points indicate ρ values. Blue: Model-Free; Orange: Model-Based.

4.2. Effect of trade-off Parameter ρ on Misinformation Spreading

Figure 2 illustrates how varying the penalty coefficient ρ influences misinformation spread and system behavior in the 100-user network under data-driven content selection. The mitigation metric \mathcal{M} (top-left) decreases monotonically as ρ increases from 0 to approximately 2.5, corresponding to a $\sim 76\%$ reduction in misinformation spread compared to baseline ($\rho = 0$). Beyond $\rho > 3$, performance slightly degrades, likely due to LLM misclassification of emotionally neutral yet false statements (discussed in Section 5). Per-user engagement cost (top-right) increases with ρ for both mean and median, but plateaus around $\rho \approx 3$, indicating users maintain substantial alignment with recommendations despite prioritization of less emotionally extreme content. Sentiment shift (bottom-left) increases with ρ , reflecting stronger moderation of initial emotional extremity. MF and MB approaches yield similar trajectories, with MB showing marginally higher shift at large ρ due to predictive optimization. The Pareto frontier (bottom-right) reveals that substantial misinformation reduction is achievable with modest engagement cost increases. The optimal operating region $\rho \in [1.0, 2.5]$ balances both objectives. MF and MB approaches trace nearly identical curves, indicating the simpler MF strategy suffices.

4.3. A Special Focus on Radical Users

Figure 3 presents the same analysis as in Section 4.2 for the 6-agent radicalized network, revealing important differences from the large network case. Misinformation mitigation (top-left) follows a similar trend, with optimal performance around $\rho = 1$, achieving up to $\sim 70\%$ reduction compared

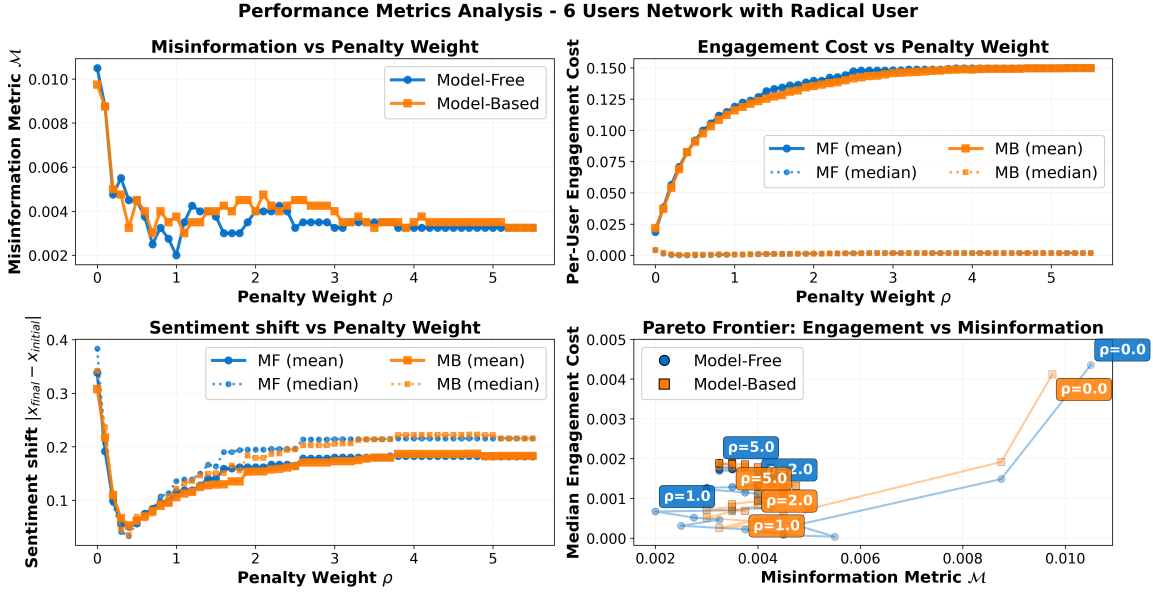


Figure 3: Performance Metrics Analysis for 6-User Radical Network (Data-Driven). **Top-left:** Misinformation metric \mathcal{M} vs. ρ . **Top-right:** Per-user engagement cost vs. ρ ; note that median (dashed) remains stable while mean (solid) increases, indicating improved engagement for non-radical users. **Bottom-left:** Sentiment shift vs. ρ . **Bottom-right:** Pareto frontier using median engagement, showing that mitigation can improve both objectives simultaneously in radicalized networks. Blue: Model-Free; Orange: Model-Based.

to the baseline that only accounts for engagement maximization. Engagement dynamics (top-right): Unlike the large network, median engagement remains nearly constant across all ρ values (dashed lines), while mean engagement cost increases (solid lines). This divergence suggests that while the radical user’s engagement decreases (pulling up the mean) with increasing ρ , the majority of users engage *more* with moderated content. Sentiment shift (bottom-left) exhibits a non-monotonic pattern, with a minimum around $\rho \approx 0.4$. This reflects the interplay between the radical user’s anchoring effect and the moderate initial opinions of the remaining users: at low ρ , engagement-only control pulls users toward the extremist’s sentiment, increasing shift; mild mitigation counteracts this drift, reducing shift; stronger mitigation then pushes sentiment below initial values, increasing shift again. Pareto frontier (bottom-right): In this case, when considering median engagement instead of mean, the trade-off curve inverts: higher ρ values simultaneously reduce misinformation *and* maintain or improve median engagement. This indicates that for networks with extremist users, mitigation strategies can enhance discourse quality for the majority of participants.

5. Discussion

Penalizing emotionally extreme recommendations through θ_M effectively reduces misinformation while maintaining engagement. Both MF and MB controllers converge to predicted equilibria, demonstrating robustness across network sizes. Their nearly identical performance suggests the simpler MF strategy suffices for deployment. The optimal penalty $\rho \in [1.0, 2.5]$ achieves up to

76% misinformation reduction, providing practical guidance for platforms balancing moderation with engagement. The radical network case reveals key insights: while misinformation control cannot override stubborn extremists, negativity propagation to other users is significantly reduced. Moreover, median engagement *improves* under mitigation, suggesting content moderation enhances discourse quality for non-radical majorities in polarized environments. A broader consideration is that θ_M penalizes emotionally extreme and novel content indiscriminately, without distinguishing between misinformation and legitimate content that is inherently emotional or urgent, such as breaking news, public health alerts, or politically charged but factually accurate speech. This could inadvertently suppress valid discourse, particularly from communities whose communicative norms involve stronger affective language. Mitigating this risk in practice would require complementary signals, such as source credibility scores or fact-checking labels, to condition the penalty on content veracity rather than emotional tone alone. Performance degradation at $\rho > 3$ stems from linguistically neutral false statements in LIAR2, weakening the emotion-truthfulness correlation. This emphasizes the need for datasets with fine-grained truth levels, temporal dynamics, and diverse linguistic styles targeting boundary cases where misinformation employs objective framing.

Several limitations point to directions for future work. The experiments rely on a single dataset (LIAR2); testing on alternative corpora would strengthen generalizability. Content arrival dynamics, including bursty or campaign-driven posting patterns, are not modeled, as LIAR2 lacks fine-grained timestamps; incorporating such dynamics would allow the novelty penalty $N(t, t_c)$ to be fully exercised. Network topology and connectivity parameters (κ_u, κ_r) represent plausible but generic social platform conditions rather than empirically calibrated ones; a sensitivity analysis over graph structure and density is a natural extension. These limitations reflect the inherent constraints of simulation-based evaluation without field data. Beyond these, future work should integrate ρ and \mathcal{M} into the closed-loop for adaptive control responding to misinformation surges during elections or crises, and specialized LLMs trained on misinformation corpora could enhance classification.

6. Conclusions

This paper presents a control framework for mitigating misinformation through sentiment-aware recommender systems. By adapting Friedkin-Johnsen dynamics to represent emotional extremity and penalizing characteristics misinformation exploits, we demonstrate up to 76% reductions in misinformation spread while maintaining engagement. Key contributions include: (1) a modified cost function θ_M penalizing misinformation-associated content characteristics; (2) convergence guarantees for both model-free and model-based strategies; (3) validation using LIAR2 dataset with LLM-extracted sentiment features; and (4) evidence that content moderation improves median engagement for non-extremists in radicalized networks. The framework provides foundation for next-generation recommender systems accounting for emotional and cognitive propagation dynamics. While challenges remain, e.g., dataset quality, adaptive tuning, results suggest algorithmic interventions can address misinformation without abandoning engagement-driven models. Future work should focus on real-world deployment, adaptive mechanisms, and integration with complementary strategies like fact-checking and user education.

Data and Code Availability

The code implementing the proposed framework and the processed LIAR2 dataset are publicly available at <https://github.com/paganick/misinformation-mitigation-model>.

References

- Jonah Berger and Katherine L Milkman. What makes online content viral? Journal of marketing research, 49(2):192–205, 2012.
- William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28):7313–7318, 2017.
- Sanjay Chandrasekaran, Giulia De Pasquale, Giuseppe Belgioioso, and Florian Dörfler. Network-aware recommender system via online feedback optimization, 2024.
- Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. Scientific reports, 10(1):1–10, 2020.
- Sarah Dean and Jamie Morgenstern. Preference dynamics under personalized recommendations. In Proceedings of the 23rd ACM Conference on Economics and Computation, pages 795–816, 2022.
- Sarah Dean, Evan Dong, Meena Jagadeesan, and Liu Leqi. Accounting for AI and users shaping one another: The role of mathematical models. Transactions on Machine Learning Research, pages 1–25, 2024.
- Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. Echo chambers in the age of misinformation. arXiv preprint arXiv:1509.00189, 2015.
- Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3):554–559, 2016.
- Nicolas Lanzetti, Florian Dörfler, and Nicolò Pagan. The impact of recommendation systems on opinion dynamics: Microscopic versus macroscopic effects. In 2023 62nd IEEE conference on decision and control (CDC), pages 4824–4829. IEEE, 2023.
- Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. In Proceedings of the 29th ACM international conference on information & knowledge management, pages 2145–2148, 2020.
- Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. A classification of feedback loops and their relation to biases in automated decision-making systems. In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pages 1–14, 2023.
- Nathaniel Persily. The 2016 us election: Can democracy survive the internet? Journal of democracy, 28(2):63–76, 2017.

Wilbert Samuel Rossi, Jan Willem Polderman, and Paolo Frasca. The closed loop between opinion formation and personalized recommendations. IEEE Transactions on Control of Network Systems, 9(3):1092–1103, 2022.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1):22–36, 2017.

Ben Sprenger, Giulia De Pasquale, Raffaele Soloperto, John Lygeros, and Florian Dörfler. Control strategies for recommendation systems in social networks. IEEE Control Systems Letters, 8: 634–639, 2024.

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

Chengcheng Xu and Madjid-Tahar Kechadi. An enhanced fake news detection system with fuzzy deep learning. IEEE Access, 12:88006–88021, 2024.