Sedative: Pacify the Online Social Networks under Hijacking-based Troll Attacks

Wen Ma, Zhiyi Liu, Haohua Du*, and Xiaoyu Liang
School of Cyber Science and Technology, Beihang University, Beijing, China
Email: {marvin2022, 21371234, duhaohua, xiaoyuliang}@buaa.edu.cn

***Abstract*—In online communities, trolling disrupts constructive discussion, fuels conflicts, and even triggers offline violence. While many defense techniques assume trolling arises from endogenous user behavior, our research unveils a different dimension: attackers control a group of accounts through hacking or payoffs, and intentionally sow discord by posting specifically designed statements. This attack is considered as a *hijacking-based troll attack*. Based on the Fridkin–Johnson opinion dynamics model, we formalize how such attacks propagate and accordingly propose Sedative, a real-time defense framework grounded in the concept of attraction basins in complex systems. Sedative identifies and protects structurally and dynamically critical nodes that constitute the backbone of the stable attraction basin to constrain perturbation spread under limited defense budgets. By focusing on localized, real-time influence rather than global computation, Sedative achieves high scalability and responsiveness. Experiments on real and synthetic networks show that Sedative suppresses up to 98% of attack-induced conflict, significantly outperforming the baselines.**

***Index Terms*—social networks, conflict, polarization**

I. Introduction

Online social networks have become breeding grounds for hate speech and extremism, which lead to abhorrent real-world events. Prior studies propose conflict mitigation strategies centered on restricting discourse to like-minded communities [1], altering network topology [2], or filtering harmful content via keyword blocking [3]. However, recent evidence highlights a new class of threats—hijacking-based troll attacks—that undermine the assumptions of these strategies [4]–[6].

A hijacking-based troll attack intrudes on different accounts to post statements that cause division and polarization. By strategically selecting the types of accounts to hijack, attackers can manipulate network dynamics at scale. For instance, hacking out-of-community accounts can provoke intergroup clashes [7]; compromising highly connected accounts breaks recommendation safeguards [8]; and orchestrating large-scale account takeovers circumvents content filters. The emergence of account black markets further lowers the barrier for executing such attacks [9], making them a potent tool for malicious actors to escalate online discord into offline consequences [10].

Despite the growing threat of these attacks, limited research has addressed how to respond effectively. To the best of our knowledge, only Gaitonde et al. [11] have proposed a defense mechanism against such discord-inducing perturbations from a game-theoretic perspective. However, their method focuses on the network's final state and incurs high computational complexity, making it unsuitable for real-time defense in online social networks.

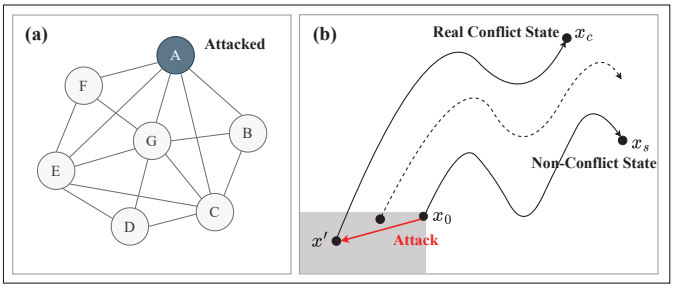


Fig. 1. (**a**) Online social network topology, node A under attack. (**b**) Perturbations caused by attack in (a) to an initial state x_0 of the network. The perturbations drive the network into a non-self-healing conflict state x_c.

To counter hijacking-based troll attacks in dynamic networks, we propose a real-time defense framework, Sedative, inspired by the concept of the basin of attraction in complex systems [12], [13]. In this context, the basin of attraction of a network state refers to the set of initial configurations from which the system dynamics will converge to that state. As shown in Fig. 1, we consider a hijacking-based troll attack as driving the network from a stable, low-conflict state toward a high-conflict one by perturbing a small set of accounts. An attack succeeds if the induced perturbation enters the conflict state's basin of attraction. The core idea of Sedative is to proactively identify and protect a set of structurally and dynamically critical nodes that constitute the backbone of the stable attraction basin. This is achieved by estimating how localized perturbations originating from potential hijacked nodes may propagate through the network and deform the basin boundary. We design an efficient mechanism to assess each node's contribution to the potential expansion of the conflict basin, thereby enabling selective protection of the most influential nodes under a limited defense budget.

To ensure responsiveness in dynamic environments, Sedative employs a localized evaluation strategy that focuses on the real-time impact of emerging perturbations, rather than computing global influence over long horizons. This approach not only aligns with the time-sensitive nature of online interactions but also significantly reduces computational overhead, making the framework scalable to large, evolving networks.

* Haohua Du is the corresponding author. This research is supported by the National Key R&D Program of China under Grant No. 2022YFB3104400.

979-8-3315-4940-4/25/$31.00 © 2025 IEEE

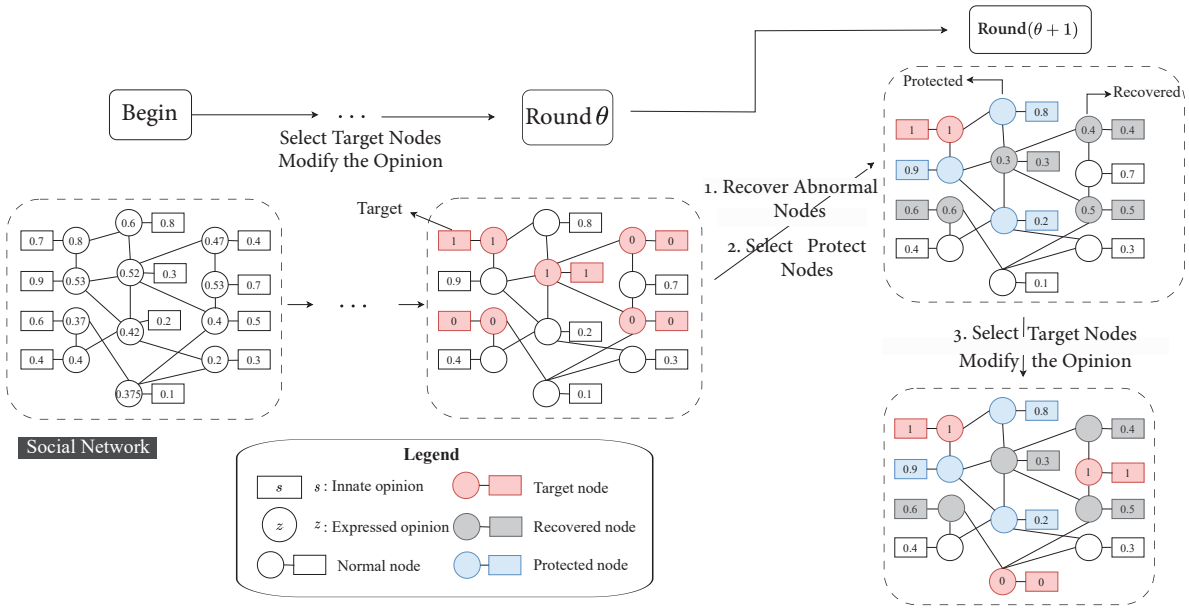


Fig. 2. **Workflow**: The attacker perturbs up to γ nodes per round over θ rounds before detection. Upon detection at round θ, the defender identifies abnormal nodes, restores their states, and protects critical nodes to contain further impact.

We summarize our contributions as follows:

- We propose the first real-time defense framework against hijacking-based troll attacks by modeling them as dynamic shifts between attraction basins in online networks. This novel perspective enables a principled understanding of how adversarial perturbations can drive the system from a low-conflict to a high-conflict state.
- We develop an efficient influence-based node assessment mechanism to identify and prioritize protection for dynamically critical nodes under a limited budget. This mechanism effectively limits the spread of perturbations toward conflict states by safeguarding nodes with high propagation potential.
- We implement a scalable, adaptive system that supports local evaluation and historical learning, achieving up to 98% attack suppression in real and synthetic datasets. The system dynamically adjusts to attacker behavior while maintaining real-time responsiveness.

II. Problem Formulation

An online social network is represented by a weighted graph $G = (V, E, w)$, where V represents users, and E represents social connections. The w_{ij} is the edge weight between nodes i and j that describes the strength of the ties between user i and j. When $(i, j) \notin E$, $w_{ij} = 0$. Let L be the weighted Laplacian matrix, and L is positive semidefinite and symmetric.

1) Opinion Dynamics Model under the Troll Attack: We adopt the Fridkin-Johnson model [14] to characterize opinion dynamics. For a specific event and all users involved in it, each user i holds a constant innate opinion denoted as $s_i \in [0, 1]$ and expresses evolving opinions $z_i^t \in [0, 1]$ at time t, which is updated with the changing opinions of neighbors as follows:

$$z_i^{t+1} = \frac{s_i + \Sigma_{j \in V} w_{ij} z_j^t}{1 + \Sigma_{j \in V} w_{ij}} \tag{1}$$

2) Key Network Metrics: We use network's disagreement value $\mathcal{D}^t(z^t)$ and polarization $\mathcal{P}^t(z^t)$ for quantifying network states, which can be calculated by

$$\mathcal{D}^t(z^t) \stackrel{\text{def}}{=} \Sigma_{(i,j) \in E} w_{ij} (z_i^t - z_j^t)^2 = {z^t}^T L z^t$$

$$\mathcal{P}^t(z^t) \stackrel{\text{def}}{=} \Sigma_{i \in V} (z_i^t - \overline{z}^t)^2$$

where $z^t = [z_1^t, z_2^t, \cdots, z_n^t]$, $z^t \in [0, 1]^n$. The $\overline{z}^t$ indicates the mean opinions at time t which can be calculated by $\overline{z}^t := 1/n \sum_{i \in V} z_i^t$. Naturally, the smaller $\mathcal{D}^t(z^t)$ and $\mathcal{P}^t(z^t)$ are, the more stable the network is, and vice versa. When considering both indicators together, we use their weighted sum, denoted as $\mathcal{D}^t + \lambda m/n \mathcal{P}^t$. Here, $|V| = n$ and$|E| = m$.

3) Objectives of Attackers and Defenders: The **attacker's goal** is to perturb the current network state from a stable region into the attraction basin of a conflict state, thereby increasing both hate and polarization. This is achieved by hijacking up to γ nodes at each time step to inject divisive content or extreme opinions. Specifically, the attacker needs to solve the following optimization problem:

$$\begin{aligned} &\max_{z_a^t} \mathcal{D}^t(z_a^t),\ \mathcal{P}^t(z_a^t),\ \mathcal{D}^t(z_a^t) + \lambda \frac{m}{n} \mathcal{P}^t(z_a^t) \\ &s.t.\ \ z_a^t \in [0, 1]^n \\ &\quad\ |V_a^t| - |V_a^{t-1}| \leq \gamma \end{aligned}$$

where $V_a \subseteq V$ represents the set of nodes controlled by the attacker and z_a^t indicates the opinions after the attacker's controlling behavior at time t. And γ is budget of the attacker. Under troll attacks, we assume that the variation of z_i^{t+1} will follow the settings of the attacker rather than being calculated

by Equation 1 for the attacked node i. That is, expressed opinions are updated according to the Equation

$$z_i^{t+1} = \begin{cases} \dfrac{s_i + \Sigma_{j\in V} w_{ij} z_j^t}{1 + \Sigma_{j\in V} w_{ij}} & \text{with normal } i \\ z_i^a \quad \text{with attacked node } i \end{cases} \tag{2}$$

where z_i^a denotes the opinion value set by the attacker for the hijacked account.

The **defender's goal** is to maintain the system within the attraction basin of a low-conflict state by strategically protecting nodes whose perturbation would lead to large-scale shifts toward conflict. That is:

$$\min_{V_d^t} \max_{z_a} \mathcal{D}^t(z_a^t),\ \mathcal{P}^t(z_a^t),\ \mathcal{D}^t(z_a^t) + \lambda \frac{m}{n} \mathcal{P}^t(z_a^t)$$
$$s.t.\ |V_d^t| - |V_d^{t-1}| \le \eta$$

where $V_d \subseteq V$ represents the set of nodes protected by the defender and η is budget of the defender. The process of attack and defense in network evolution is shown in the Fig. 2. Before the attack is detected (assumed to occur within θ time slots), the attacker selects nodes and modifies their opinions. Upon detecting the attack, the defender scans for anomalous nodes, restores their states, and protects critical nodes in the network.

III. Methodology

In this section, we will describe the design details of how our work realizes the idea discussed in Section I. The defender's strategies primarily solve two main problems: **identifying the riskiest perturbation** and **finding the top-η associated nodes**. However, they are NP-hard since they all can be reduced to the network dismantling problem [15]. Thus, we propose heuristic algorithms respectively.

A. Risky Perturbation Identification

Getting real-time accurate numerical modeling in social networks is difficult, so we approximate the dynamics of the network by

$$\begin{cases} \dfrac{d\mathcal{D}^t(z_a^t)}{dt} = \alpha \mathcal{D}^t(z_a^t)\Theta(t) - \varepsilon \mathcal{D}^t(z_a^t) \\ \dfrac{d\mathcal{P}^t(z_a^t)}{dt} = \alpha \mathcal{P}^t(z_a^t)\Theta(t) - \varepsilon \mathcal{P}^t(z_a^t) \end{cases} \tag{3}$$

The two ordinary differential equations describe the time evolution of the disagreement and polarization. Take the first equation as an example, the derivative of $\mathcal{D}^t(z_a^t)$ with respect to time is the changing rate of $\mathcal{D}^t(z_a^t)$. The changing part is the nodes affected by the t^{th} round attacking with a part of $\alpha\mathcal{D}^t(z_a^t)\Theta(t)$ and recovered by the defender with $\varepsilon\mathcal{D}^t(z_a^t)$. α indicates the success rate of the attack and ε is the proportion of the suspected conflicting nodes that were successfully recovered at time t. The other equation obeys similar rules.

When a node has been successfully controlled with probability α, it can increase the disagreement and polarization with $\Theta(t)$, which equals,

$$\Theta(t) = \frac{1}{\langle k \rangle} \sum_{i=1}^{n} c(k_i) \Pr(k_i) N_{k_i}(t)$$

Algorithm 1 Risky Perturbation Identification

Input: $\mathbf{x}^*, \mathbf{x}_0, \mathbb{P} = \varnothing$
Output: $\mathbb{P} = \{\forall\delta | |\mathbf{M}(t_a)\cdot\delta\mathbf{x}_{t_0} - \mathbf{x}^*| \le \tau\}$

```
t_a ≡ arg min |x* − x_0|
while t < t_a do
    if x'_0 ∈ x*'s basin of attraction then
        P = P ∪ δx'_0
    end if
    t ← t + t'
    x'_0 ← x'_0 + δx_0
end while
return P
```

where $\Pr(k_i)$ is the probability that the node has degree k_i and $\langle k \rangle$ is the average node degree of the network. $c(k_i)$ indicates the contradiction of the node, and $N_{k_i}(t)$ gives the number of nodes that can be reached in a single round. Thus, the attack perturbation can be approximated by

$$\delta^t = \begin{cases} \delta\mathcal{D}^t(z_a^t) = \mathbf{M}(\mathcal{D}^t(z_a^t), t)\cdot\delta\mathcal{D}^{t-1}(z_a^{t-1}) \\ \delta\mathcal{P}^t(z_a^t) = \mathbf{M}(\mathcal{P}^t(z_a^t), t)\cdot\delta\mathcal{P}^{t-1}(z_a^{t-1}) \end{cases}$$

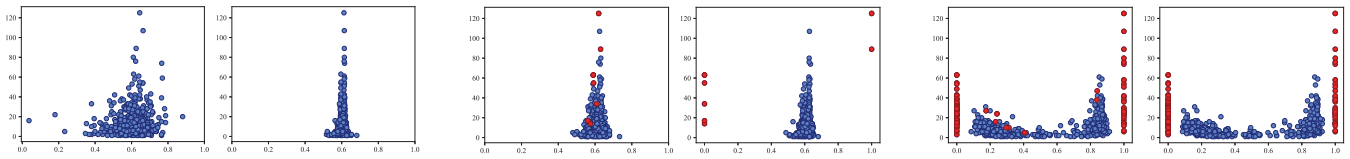
and the matrix $\mathbf{M}$ is the solution of the Eq. 3 with a fixed initial condition.

We assume that in each round the attacker moves first, then in round t there exist two system states, $\mathbf{x}_t$ and $\mathbf{x}'_t$, with and without a defender action, respectively. Since the upper bound of the attacking perturbation is of more concern, we use only the state of the system without defense intervention here. Given the state of the network $\mathbf{x}_t$ at t, we identify the risky attack perturbations iteratively as follows.

To identify the time of the attacker's approach to the predefined conflicting state $t_a \equiv \arg\min|\mathbf{x}^* - \mathbf{x}_t|$, we first integrate the system dynamics over a time window T. Notably, the T increases with the upper bound of attacking perturbation. We then integrate the variational equation up to this time to obtain the corresponding variational matrix $\mathbf{M}(t_a)$, which maps a small change $\delta\mathbf{x}_{t_0}$ to the initial state of the network to a change $\delta\mathbf{x}(t_a)$ in the resulting of attacking perturbation orbit according to $\delta\mathbf{x}(t_a) = \mathbf{M}(t_a)\cdot\delta\mathbf{x}_{t_0}$. This mapping is used to select the risky incremental perturbation $\delta\mathbf{x}_t$ to the current initial state that minimizes the distance between the perturbed orbit and the target at time t_a within an error interval τ, as well as additional constraints on $\delta\mathbf{x}_{t_0}$ to ensure the validity of the variational approximation. The detailed information is shown in Algorithm 1.

B. Defense Node Selection

To mitigate risky perturbations, the defender's next task is to identify nodes that contribute most to the perturbation. The contribution of hijacked nodes to perturbations is governed by two key factors: (1) the magnitude of node-level perturbation, dictated by how much a node's state deviates due to hijacking; and (2) the topology and semantic contrast between neighboring nodes, which governs how perturbations propagate and



(a) Opinion evolution without attacks (b) Attacked network initial status (c) Attacked network status after a period

Fig. 3. Opinion Evolution With Structure-Aware Attacks. The x-axis is opinion, and the y-axis is degrees. (a) shows opinion convergence without attacks. (b) and (c) depict structure-aware attackers modifying opinions of target nodes. After several attack rounds, opinions split into two distinct groups.

amplify through the network. To quantify the impact of a single node's hijacking, we define the perturbation sensitivity of node i as:

$$S_i = (\hat{z}_i - \bar{z})^2$$

where $\bar{z}$ is the global opinion mean, and $\hat{z}_i \in \{0, 1\}$ denotes the opinion that maximizes local divergence within the node's L-hop neighborhood. Given the convex nature of the attacker's optimization objective, pushing hijacked nodes toward binary extremes (0 or 1) is the optimal strategy. To capture how perturbations diffuse through the network, we define a direction-aware weighted adjacency matrix:

$$W_{ij} = \omega_{ij} = w_{ij} \cdot (\hat{z}_i - z_j)^2$$

This encodes both structural connectivity and opinion divergence between adjacent nodes. Using S_i and W, we compute each node's contribution v_i via a biased PageRank formulation:

$$v_i = \alpha \sum_{j \in N_{out}(i)} v_j \cdot \frac{\omega_{ij}}{\sum_{k \in N_{in}(j)} \omega_{kj}} + (1 - \alpha) S_i$$

In a single-round game with no historical information, the defender chooses the top-η nodes with the highest v_i.

While maintaining real-time responsiveness, Sedative dynamically adjusts to attacker behavior. We analyze historical attacks using kernel density estimation (KDE) over the features z_i and degrees $\deg_i$ of detected hijacked nodes $i \in V_a$ to estimate hijack risk r_i. The estimated risk reflects the likelihood that node i will be targeted based on its similarity to previously hijacked nodes. Finally, the protection set V_t is constructed by selecting the top-η nodes with the highest joint score $v_i(r_i + b)$, and we set $b = 1$. This adaptive strategy ensures that Sedative prioritizes both high-impact and high-risk nodes, optimizing defense resource allocation and enhancing robustness against evolving hijacking strategies.

IV. Experiments

A. Experiment Setup

Datasets. The validation datasets are sourced from both real-world and synthetic environments. We synthesized two networks using the Barabási-Albert model [16](abbreviated as BA model) and Erdős–Rényi model [17](abbreviated as ER model), respectively. In real-world datasets, user opinions are normalized to values in [0,1]. For synthetic networks, opinions are sampled from a Gaussian distribution with mean 0.5 and variance 0.2, followed by several rounds of evolution to establish correlation between opinions and network topology. Table I provides basic information about the network graphs.

TABLE I
Networks for Evaluation

Dataset	Vertices	Edges	Disagreement	Polarization
Twitter [18]	548	3638	2.25	0.53
Reddit [6]	556	8969	0.88	0.05
Facebook [19]	4039	88234	66.64	9.72
BA [16]	10000	99900	27.23	7.16
ER [17]	10000	199991	15.66	0.43

Although Twitter is now called X, we use its original name to match the dataset at the time of collection.

Attacker Setting. We design a set of attacker models with diverse behavioral preferences to simulate realistic hijacking strategies:

- **Structure-aware attackers:** prioritize high-degree nodes, inspired by the Acquaintance Immunization Algorithm [20]. They randomly sample a node and select the highest-degree neighbor as the target, repeating this process until selecting γ distinct nodes.
- **Feature-aware attackers:** target nodes whose opinion values fall within a specific range (e.g., [0.4, 0.6] around the network mean of 0.5), reflecting a preference for moderates or swing users.
- **Hybrid attackers:** jointly consider node degree and opinion value by selecting neighbors of random nodes whose opinions fall in the desired range.
- **Random attackers:** serve as a baseline, selecting γ target nodes uniformly at random from unselected nodes.

To reflect realistic uncertainty in attacker behavior, all strategies incorporate stochastic sampling based on their preference distributions. We set the budget γ of the attacker to be 10 in the Twitter network and Reddit network, 15 in the Facebook network, and 30 in Synthetic networks, considering the scale of networks.

The illustration of structure-aware attackers' attacks on the Twitter dataset is shown in Fig. 3.

Defender Setting. We evaluate our proposed defense algorithm alongside four baseline methods to validate its effectiveness. The compared algorithms are as follows:

- **Sedative**: Our proposed method, which jointly considers perturbation contribution and hijack risk in an adaptive protection strategy.

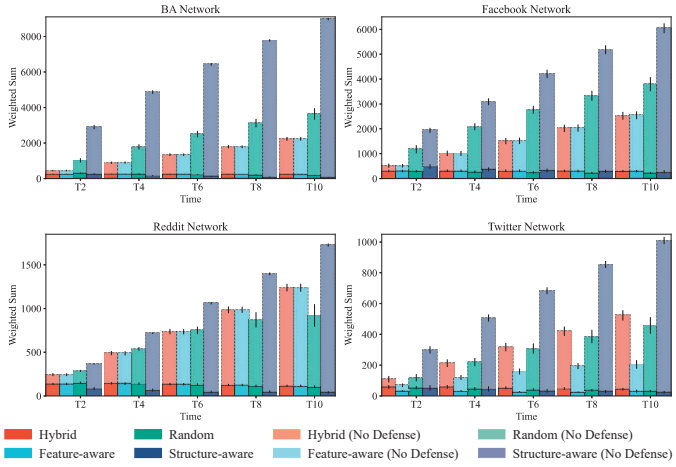


Fig. 4. Defense Effectiveness of Sedative Against Various Attacks on Real and BA Networks

- **Degree-Based**: Ranks nodes by degree centrality and selects the top-η nodes with the highest combined degree and risk.
- **Greedy**: Iteratively selects nodes whose hijacking would cause the largest increase in network-level loss. This process repeats until η nodes are selected.
- **Risk-Only**: Selects the top-η nodes purely based on hijack risk.
- **Random**: Randomly selects η nodes as the protection set.

To simulate anomaly detection in social networks, we assume that after an attack, the defender can detect and restore a fraction ε of hijacked nodes to their original states. This setting represents real-world mechanisms such as account recovery and malicious account removal. We set the budget η of the attacker to be 10 in the Twitter and Reddit network, 15 in the Facebook network, 60 in Synthetic networks, and $\varepsilon = 0.8$.

B. Results

Due to the stochastic nature of attacks, we use the Monte Carlo simulation by repeating each experiment 500 times and report the average results.

Fig. 4 illustrates the network states of three real-world datasets and a BA network under four types of attacks, comparing scenarios without defense (represented by the upper semi-transparent dashed bars) and with the Sedative defense mechanism applied (represented by the lower solid bars). The difference in bar heights directly reflects Sedative's ability to suppress attack-induced perturbations—the larger the gap, the more effective the defense. The results demonstrate that Sedative significantly reduces the extent of network disruption and enhance overall network stability. Furthermore, attackers employing structure-aware target selection cause the most severe damage, underscoring the critical need for defense strategies specifically designed to counter such attacks.

Fig. 5 presents the trends of network disagreement and weighted polarization from evolution cycles T_0 to T_{10} under structure-aware and hybrid attacks on Facebook and ER networks. The bar charts represent disagreement values under different defense methods (left y-axis), while the line plots depict the weighted polarization levels (right y-axis). Notably, Sedative demonstrates strong defense performance from the early stages of attack. At T_1, both disagreement and polarization metrics under Sedative are significantly lower than those of the baseline methods, indicating its effectiveness in curbing early-stage fragmentation and polarization in social networks under adversarial perturbations.

TABLE II
DEFENSE EFFECTIVENESS UNDER VARYING BUDGET (BA NETWORK)

Attack	Budget	Disagreement			Polarization		
		$\varepsilon = 0.8$	$\varepsilon = 0.6$	$\varepsilon = 0.4$	$\varepsilon = 0.8$	$\varepsilon = 0.6$	$\varepsilon = 0.4$
Structure	3γ	98.66%	98.26%	97.45%	97.27%	96.37%	94.67%
	2γ	97.71%	96.76%	95.35%	95.63%	94.01%	91.30%
	γ	94.44%	92.68%	89.51%	91.71%	88.76%	83.66%
Feature	3γ	81.90%	77.06%	69.03%	79.14%	73.27%	63.17%
	2γ	82.44%	75.93%	67.37%	80.35%	72.82%	62.56%
	γ	79.88%	74.35%	64.84%	78.46%	72.45%	61.89%
Hybrid	3γ	90.93%	88.50%	83.74%	84.19%	78.75%	69.98%
	2γ	90.07%	86.48%	81.40%	84.65%	77.16%	68.07%
	γ	86.75%	83.57%	77.92%	81.42%	75.47%	66.03%
Random	3γ	81.94%	77.23%	68.42%	79.18%	73.31%	63.13%
	2γ	82.38%	75.85%	67.53%	80.31%	72.83%	62.58%
	γ	79.54%	74.53%	65.54%	78.44%	72.47%	61.93%

Tabel II shows the percentage of attack effects mitigated by our defense under different detection rates (ε) and defense budgets (η) in the BA network. A lower ε and smaller η (e.g., $\varepsilon = 0.4$, $\eta = \gamma$) represent the most constrained defense setting. Even in this case, Sedative reduces disagreement by 89.51% under structure-prioritized attacks, showing strong robustness. In the optimal setting ($\varepsilon = 0.8$, $\eta = 3\gamma$), the mitigation rate reaches 98.66%, confirming the method's high effectiveness when resources are sufficient.

V. CONCLUSION

This paper addresses the issue of hijacking-based troll attacks in online communities, where attackers control multiple accounts to sow discord and fuel conflicts. We propose a dynamic model based on the Friedkin-Johnsen framework to describe these attacks and design a defense mechanism, Sedative. Our experimental results demonstrate that Sedative can reduce network conflict by up to 98%. Future work will explore the use of graph neural networks to identify risky perturbations in large-scale social networks.

REFERENCES

[1] A. Zareie and R. Sakellariou, "Similarity-based link prediction in social networks using latent relationships between the users," *Scientific Reports*, vol. 10, no. 1, p. 20137, 2020.

[2] W. Alorainy, P. Burnap, H. Liu, M. Williams, and L. Giommoni, "Disrupting networks of hate: characterising hateful networks and removing critical nodes," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 27, 2022.

[3] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of hate speech in online social media," in *Proceedings of the 10th ACM Conference on Web Science*, ser. WebSci '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 173–182.

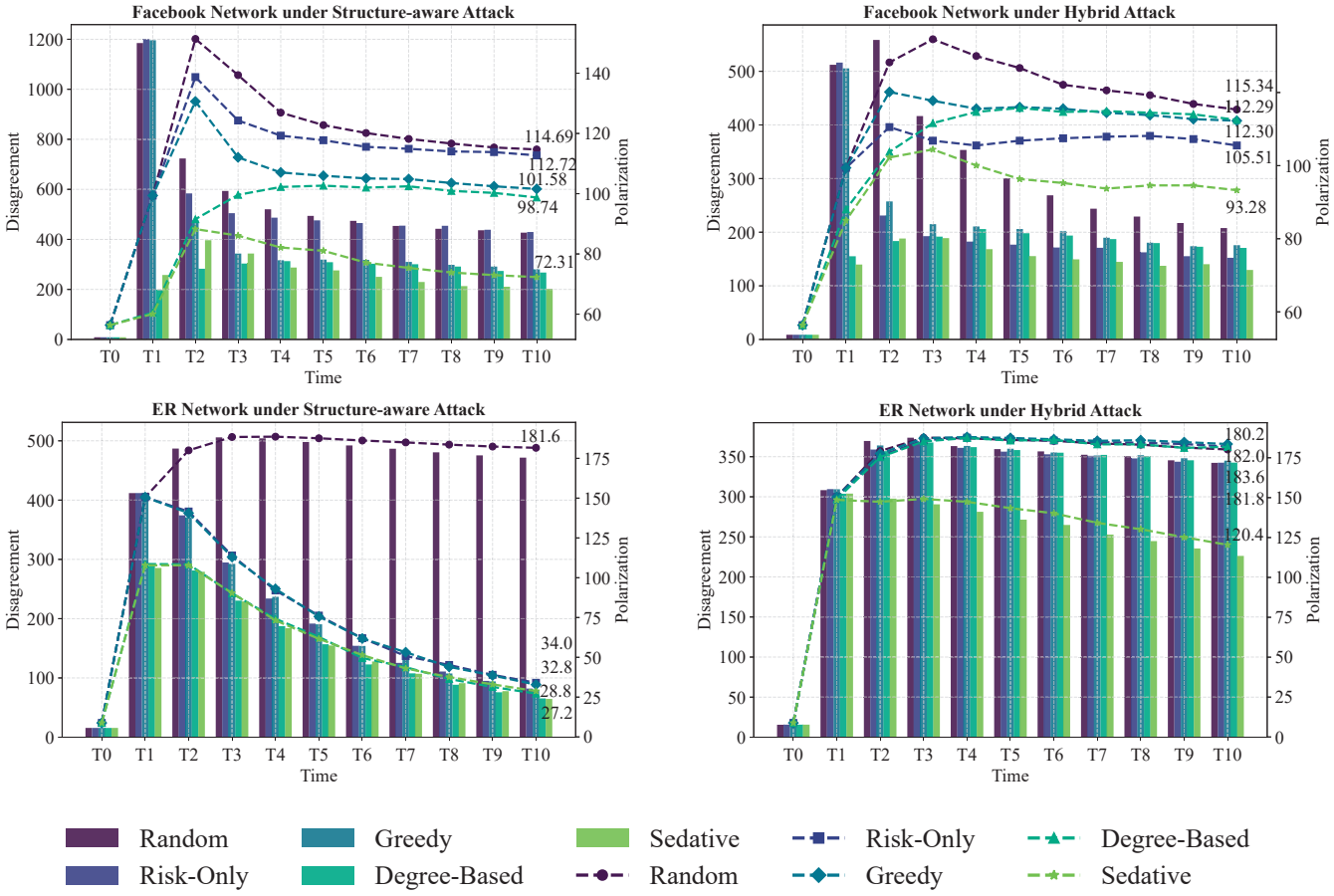


Fig. 5. Disagreement and Weighted Polarization under Structure-Aware and Hybrid Attacks on Facebook and ER Networks

[4] H. Ira and E. C. Shawn, "Hammering hirak: Exposing a multi-year effort to manipulate algeria's online political discourse and suppress dissident voices," 2021. [Online]. Available: https://public-assets.graphika.com/reports/graphika_report_hammering_hirak.pdf

[5] Twitter, "Disclosing networks of state-linked information operations," 2021. [Online]. Available: https://blog.twitter.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations

[6] M. F. Chen and M. Z. Racz, "An adversarial model of network disruption: Maximizing disagreement and polarization in social networks," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 2, pp. 728–739, 2022.

[7] J. C. Wong, "Twitter ceo jack dorsey account hacked," San Francisco, Aug. 2019. [Online]. Available: https://www.theguardian.com/technology/2019/aug/30/twitter-ceo-jack-dorsey-account-hacked

[8] S. Gatlan, "New york post hacked with offensive headlines targeting politicians," Oct. 2022. [Online]. Available: https://www.bleepingcomputer.com/news/security/new-york-post-hacked-with-offensive-headlines-targeting-politicians

[9] B. Toulas, "Over 15,000 hacked roku accounts sold for 50¢ each to buy hardware," Mar. 2024. [Online]. Available: https://www.bleepingcomputer.com/news/security/over-15-000-hacked-roku-accounts-sold-for-50-each-to-buy-hardware

[10] M. Ayad, "The propaganda pipeline: The isis fuouaris upload network on facebook," *Elaborada por Istitute for Strategic Dialog (ISD). Disponível em: https://WWW.isdglobal. org/isd-publications/the-propaganda-pipeline-the-isis-fuouaris-upload-network-on-facebook/[Acesso em 14 dez. 2021]*, 2020.

[11] J. Gaitonde, J. Kleinberg, and É. Tardos, "Adversarial perturbations of opinion dynamics in networks," in *Proceedings of the 21st ACM Conference on Economics and Computation (EC '20)*, 2020, Conference Paper, p. 471–472.

[12] S. M. Huttegger and K. J. Zollman, "Dynamic stability and basins of attraction in the sir philip sidney game," *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, no. 1689, pp. 1915–1922, 2010.

[13] S. P. Cornelius, W. L. Kath, and A. E. Motter, "Realistic control of network dynamics," *Nature communications*, vol. 4, no. 1, p. 1942, 2013.

[14] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *The Journal of Mathematical Sociology*, vol. 15, no. 3-4, pp. 193–206, 1990.

[15] O. Artime, M. Grassia, M. De Domenico, J. P. Gleeson, H. A. Makse, G. Mangioni, M. Perc, and F. Radicchi, "Robustness and resilience of complex networks," *Nature Reviews Physics*, vol. 6, no. 2, pp. 114–131, 2024.

[16] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[17] P. Erdös and A. Rényi, "On random graphs i," *Publ. math. debrecen*, vol. 6, no. 290-297, p. 18, 1959.

[18] A. De, S. Bhattacharya, P. Bhattacharya, N. Ganguly, and S. Chakrabarti, "Learning a linear influence model from transient opinion dynamics," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 401–410.

[19] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," Jun. 2014. [Online]. Available: http://snap.stanford.edu/data

[20] P. Liu, H. Miao, and Q. Li, "A common acquaintance immunization strategy for complex network," in *8th IEEE/ACIS International Conference on Computer and Information Science*. LOS ALAMITOS: IEEE Computer Soc, 2009, Conference Proceedings, pp. 713–717.