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

Wen Ma, Zhiyi Liu, Haohua Du*, and Xiaoyu Liang

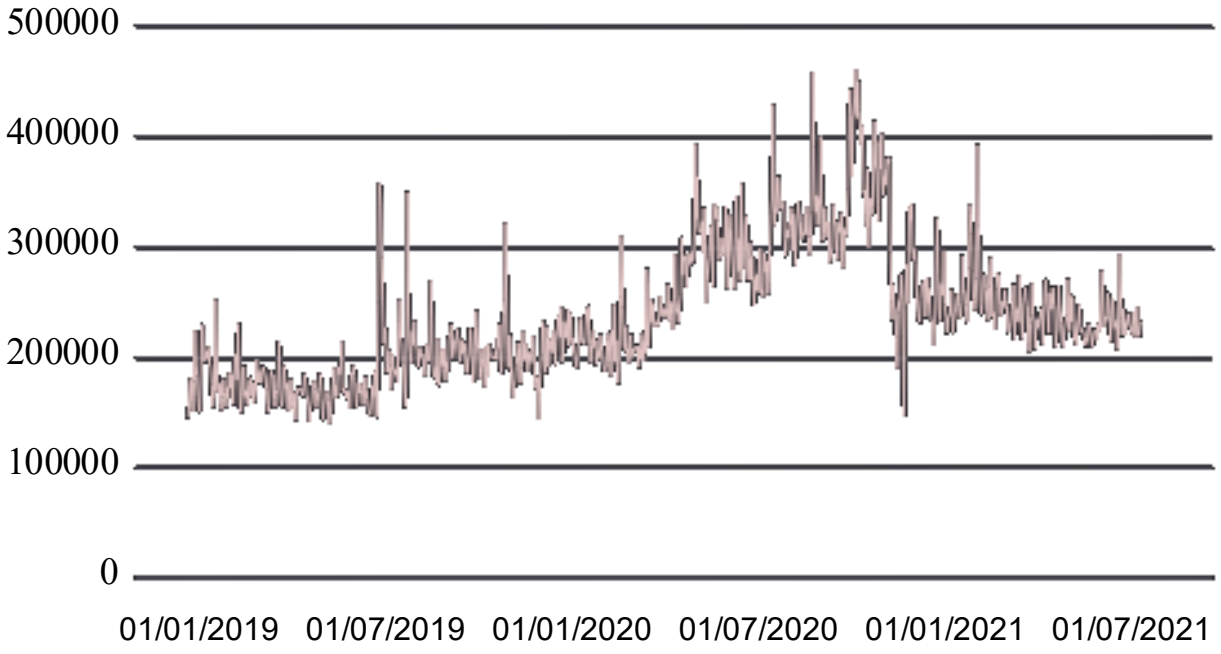
School of Cyber Science and Technology, Beihang University

Presenter: Wen Ma

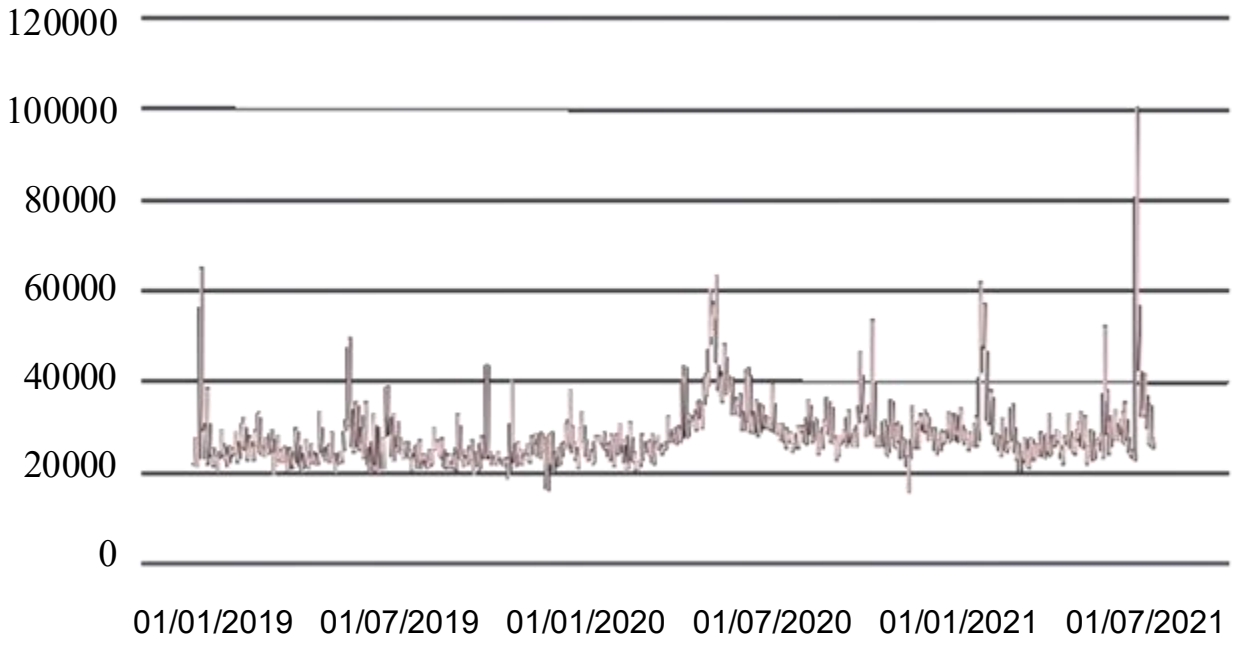
Email: marvin2022@buaa.edu.cn

Background

Discussion around and examples of hate speech posted online (US)



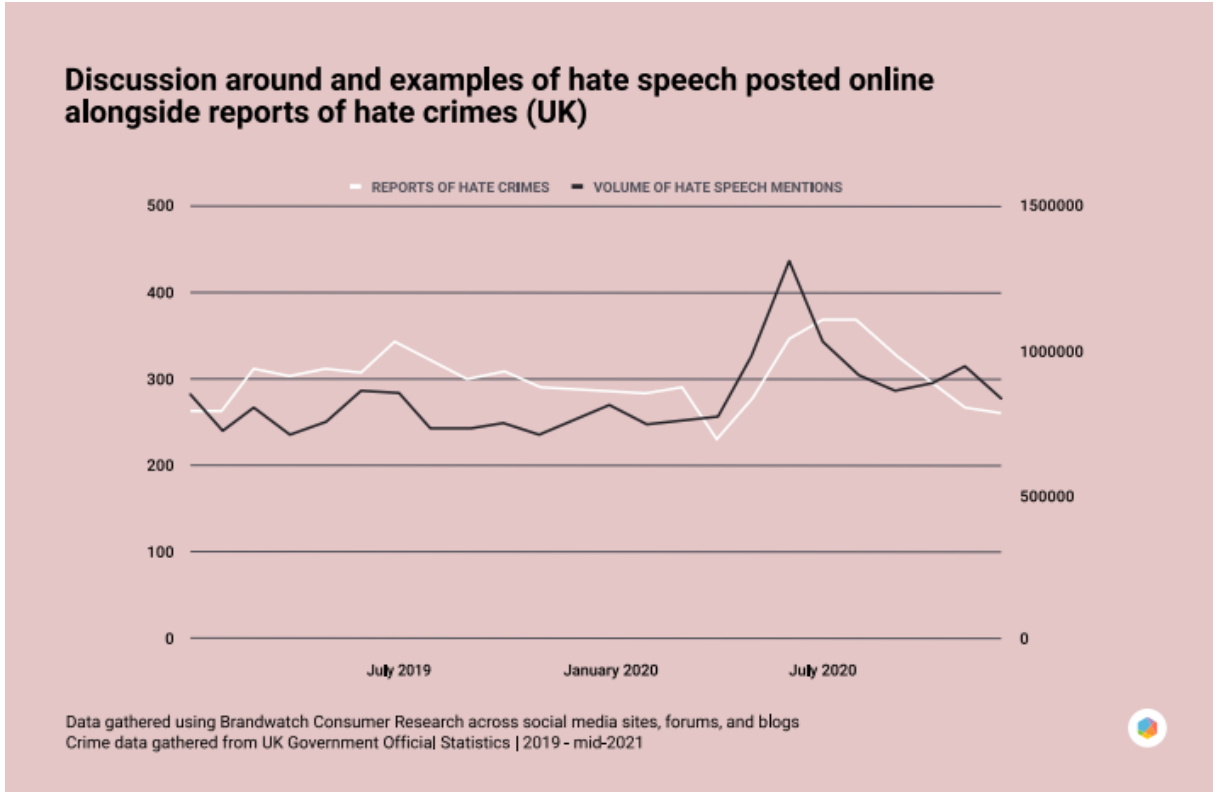
Discussion around and examples of hate speech posted online (UK)



Between 2019 and mid-2021, on average there was a new post about **race** or **ethnicity-based hate** speech **every 1.7 seconds.** **But it doesn't stop online.**

https://dtl-beta-website-assets.s3.amazonaws.com/Uncovered_Online_Hate_Speech_DT_Lx_BW_V2_1_6aef9e5c5f.pdf

Background



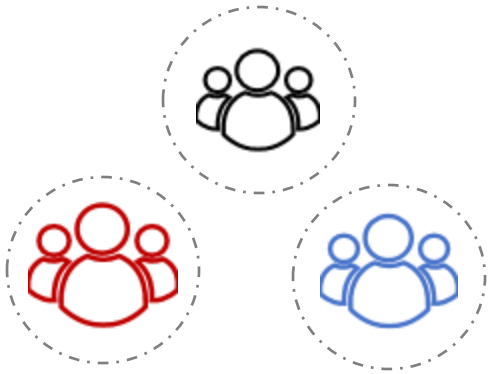
"**hate crime** and **online discussion** around hate speech rose roughly in tandem in both countries **suggests a troubling link between online words and 'real-world' action.** "

《Uncovered: Online Hate Speech in the COVID Era》

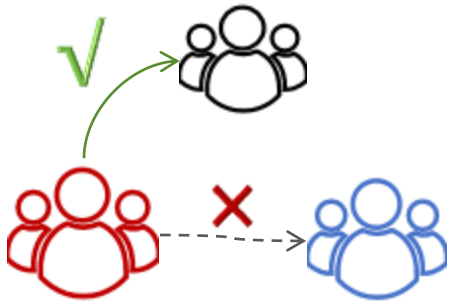
How are current defense strategies trying to solve this problem?

Existing Defense Strategies

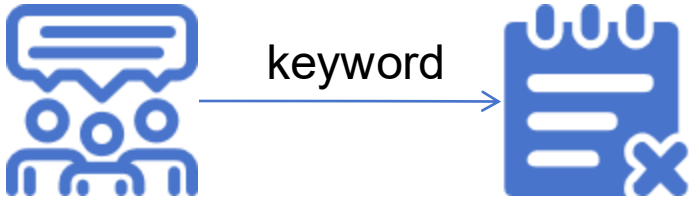
- Keeping users in like-minded communities



- Rewiring parts of the network



- Filtering harmful content using keyword detection



A New Type of Threat

Twitter Support@ @TwitterSupport. 23，July

To recap:

130 total accounts targeted by attackers

45 accounts had Tweets sent by attackers

36 accounts had the DM inbox accessed

8 accounts had an archive of "Your Twitter Data" downloaded, none of these are Verified

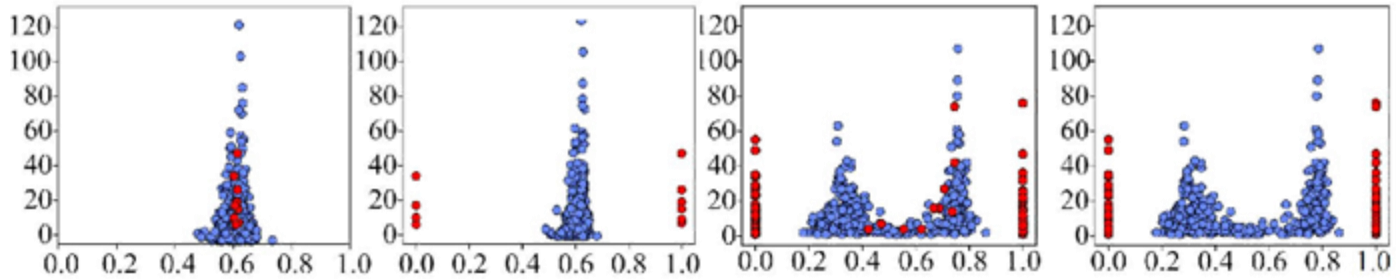
isolated
asynchronous

coordinated
synchronized

Trolling may also be organized and planned.

Limitations of Existing Approaches

➢ **hijacking-based troll attack:** intrudes on different accounts to post statements that cause division and polarization.

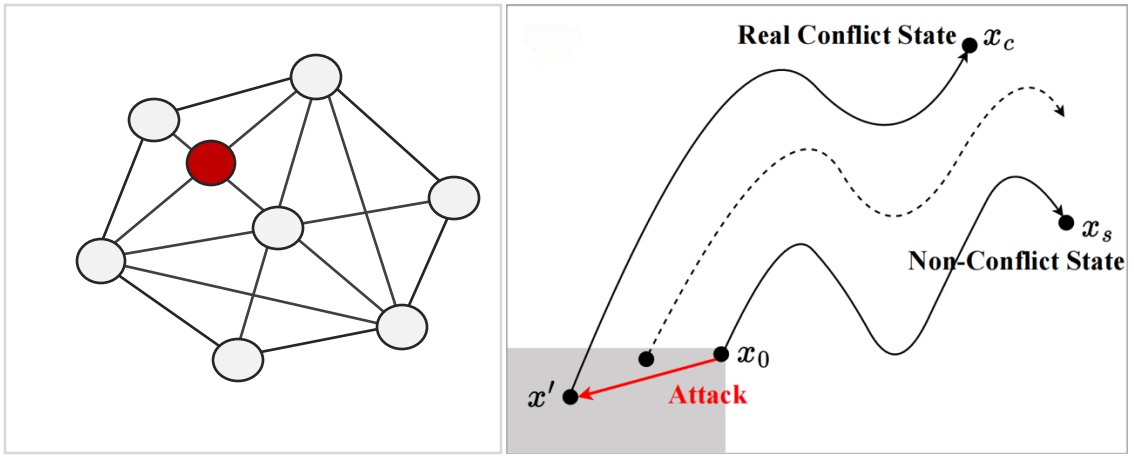


$t=0$ $t=1$ $t=5$ $t=10$

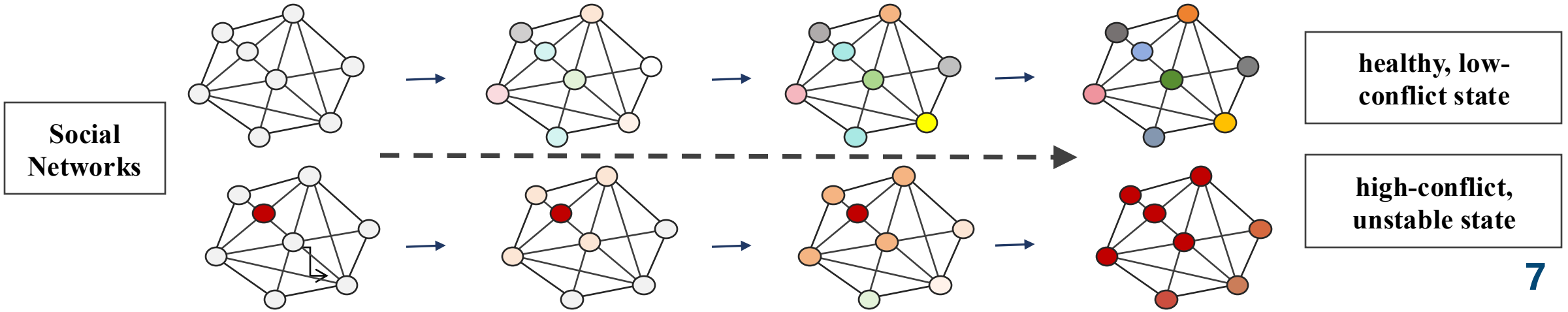
- hacking out-of-community accounts can provoke intergroup
- compromising highly connected accounts breaks recommendation safeguards
- orchestrating large-scale account takeovers circumvents content filters

Our Target

➢ **Target:** Pacify the Online Social Networks under Hijacking-based Troll Attacks



We consider a troll attack as making a network change from the stable state to the undesired conflict state by intentionally perturbing at least γ nodes in it.



The Challenge

- Unknown Attack Origins:
  - It's unclear which nodes are perturbed and how local changes affect the global network state.

- Hard Protection Decisions:
  - Selecting nodes to protect under budget is NP-hard and requires careful prioritization.

- Real-Time Adaptation Needed:
  - The network evolves rapidly due to simultaneous attacker and defender actions.

Sedative, a real-time defense framework

- **Basin of attraction:** 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

- Sedative protects critical nodes that maintain network stability
- Estimates how local perturbations can deform the basin boundary
- Enables targeted protection under limited defense budget

Problem Formulation

social network

$$G = (V, E, w)$$

innate opinion z_i

expressed opinion s_i

Laplacian matrix

$$L = D - A \qquad D_{ii} = \sum_{(i,j)\in E} w_{ij} \qquad A_{ij} = w_{ij}$$

Fridkin-Johnson model

$$z_i^{(t+1)} = f(s, z^{(t)}, w) = \frac{s_i + \sum_{j\in V} w_{i,j} z_j^{(t)}}{1 + \sum_{j\in V} w_{i,j}}$$

Key Metrics

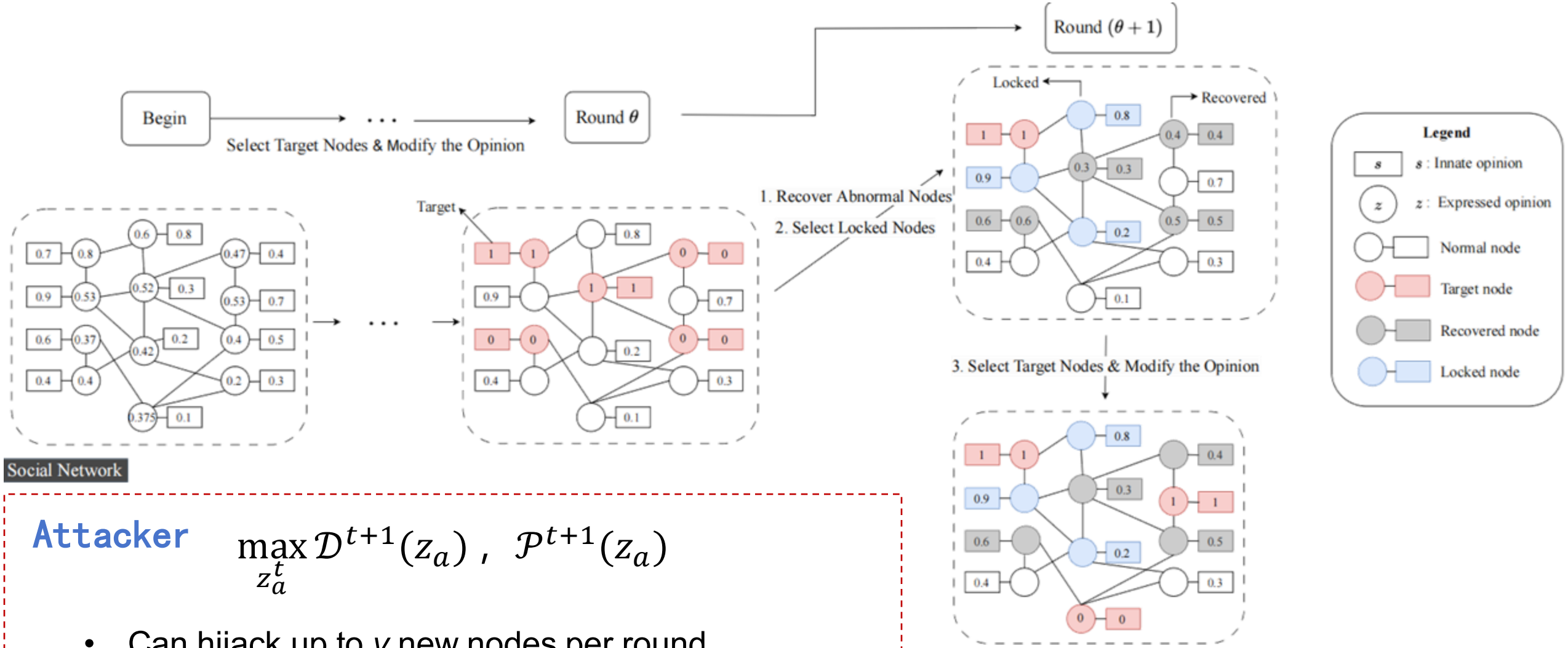
Disagreement

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

Polarization

$$\mathrm{P}^t(z^t) \overset{\text{def}}{=} \Sigma_{i\in V}(z_i^t - \bar{z}^t)^2$$

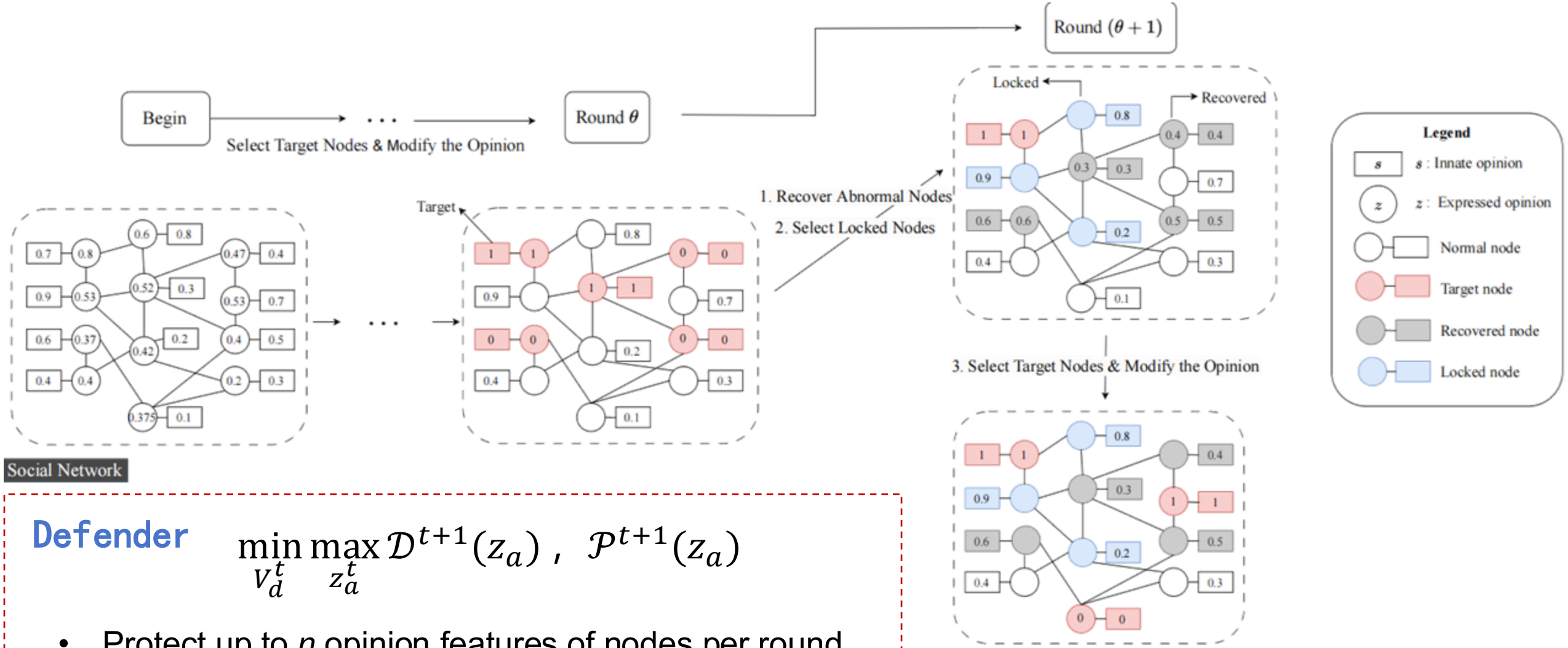
Modeling the Game



Attacker $\max_{z_a^t} \mathcal{D}^{t+1}(z_a)$, $\mathcal{P}^{t+1}(z_a)$

- Can hijack up to γ new nodes per round
- Controlled set V_a^t modifies opinions directly

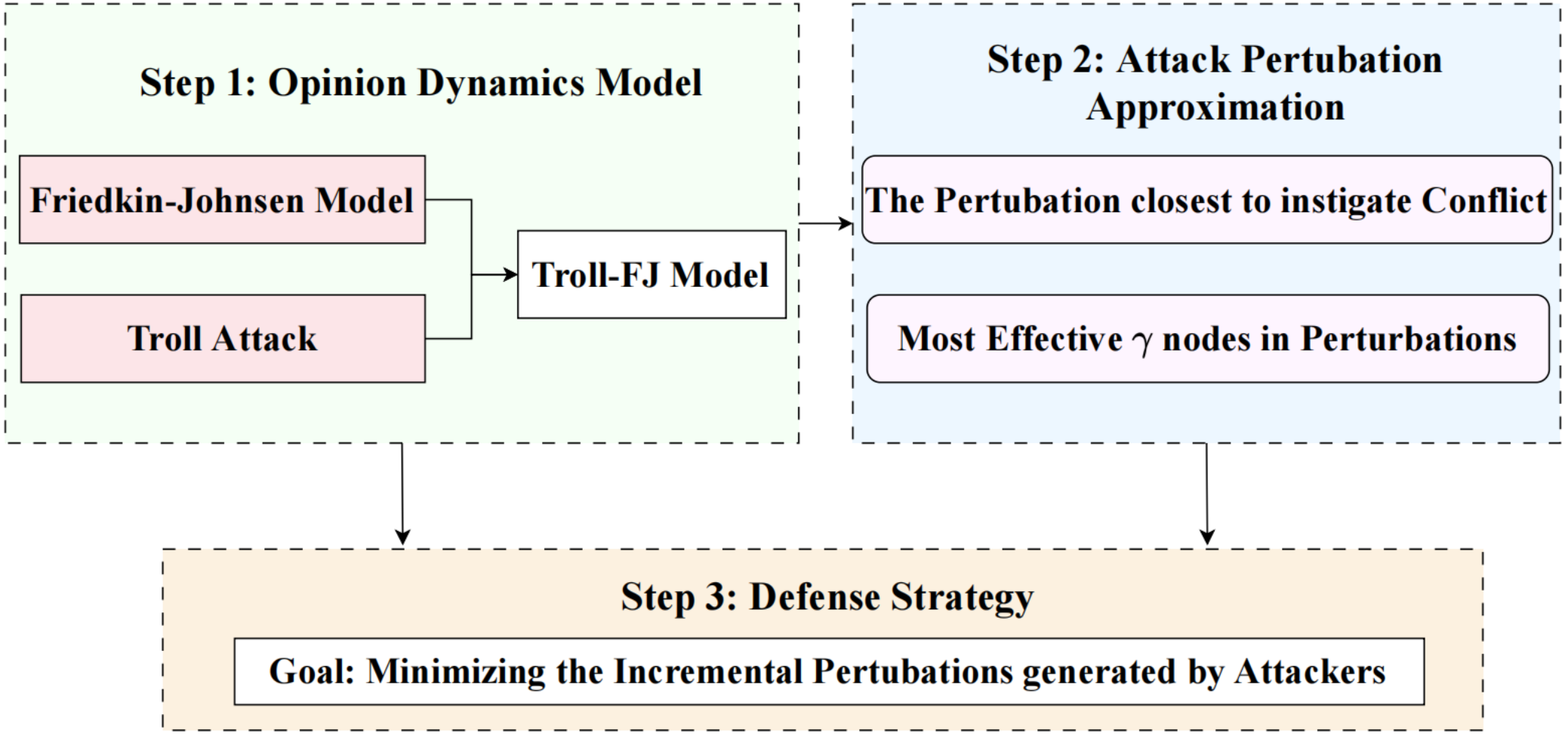
Modeling the Game



Defender $\min_{V_d^t} \max_{z_a^t} \mathcal{D}^{t+1}(z_a)\ ,\ \mathcal{P}^{t+1}(z_a)$

- Protect up to η opinion features of nodes per round
- The larger η is, the greater the defender's ability

Solution Sketch



Risky Perturbation Identification

Algorithm 1 Risky Perturbation Identification

Input: $\mathbf{x}^*, \mathbf{x}_0, \mathbb{P} = \varnothing$

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

1: $t_a \equiv \arg\min |\mathbf{x}^* - \mathbf{x}_0|$
2: **while** $t < t_a$ **do**
3: **if** $\mathbf{x}'_0 \in \mathbf{x}^*$'s basin of attraction **then**
4: $\mathbb{P} = \mathbb{P} \cup \delta\mathbf{x}'_0$
5: **end if**
6: $t \leftarrow t + t'$
7: $\mathbf{x}'_0 \leftarrow \mathbf{x}'_0 + \delta\mathbf{x}_0$
8: **end while**
9: **return** $\mathbb{P}$

- ✓ Predict attack path by integrating opinion dynamics over time window T
- ✓ Estimate when the system is most likely to reach a conflict state:

  $t_a \equiv \arg\min |x* - x_t|$

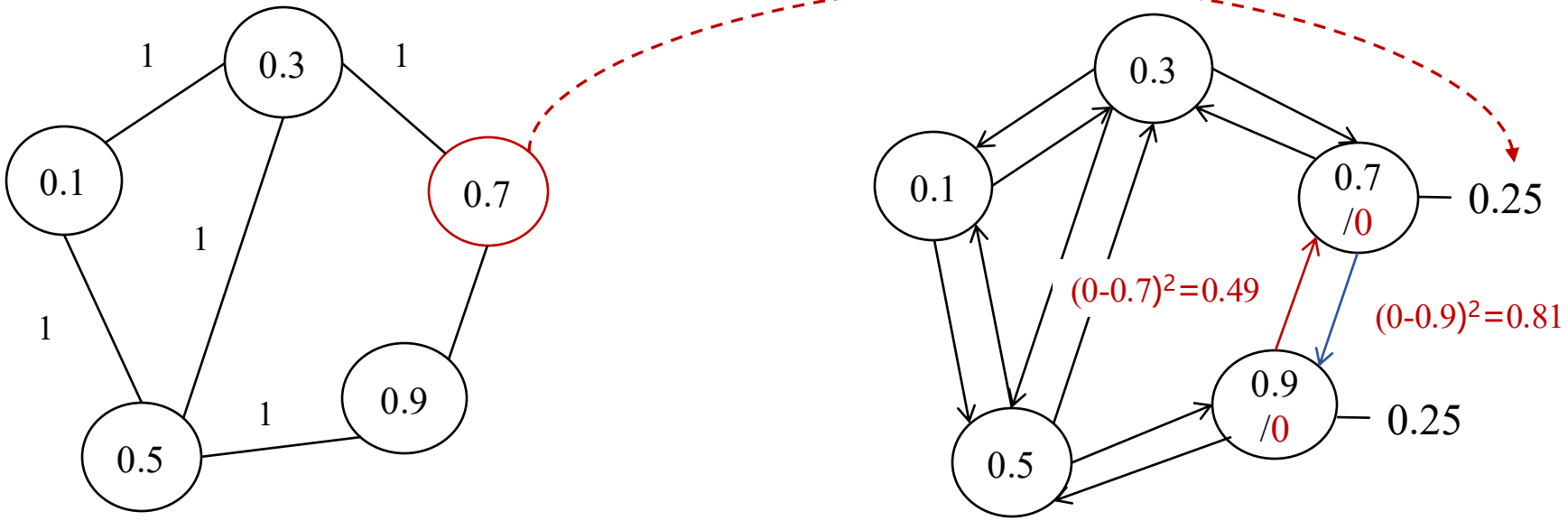
- ✓ Compute variational matrix $M(t_a)$:
  maps small initial perturbation δx_{t0} to system shift at t_a
- ✓ Select worst-case perturbation δx_t that drives network closest to conflict

Defense Node Selection

- Goal: Identify and protect nodes most likely to spread harmful perturbations
  - Perturbation Sensitivity: $S_i = (\hat{z}_i - \bar{z})^2$

The optimal perturbation is modified to 0;
$S_i=(0-0.5)^2=0.25$

$\bar{z}=0.5$



Use the increment of the perturbation-induced edge divergence value to assign weights to the edges.

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

$$g(v) = \alpha \sum_{u \in N_{\text{out}}(v)} g(u) \frac{\omega_{vu}}{\sum_{j \in N_{in}(v)} \omega_{vj}} + (1-\alpha) S_v$$

Defense Node Selection

- Goal: Identify and protect nodes most likely to spread harmful perturbations

  - Node's contribution:

$$g(v) = \alpha \sum_{u \in N_{\text{out}}(v)} g(u) \frac{\omega_{vu}}{\sum_{j \in N_{in}(v)} \omega_{vj}} + (1-\alpha) \mathrm{S}_v$$

  - Node's risk:

$$P(Z) = \frac{1}{|\mathbf{V}_a| h^d} \sum_{v \in \mathbf{V}_a} K\left(\frac{X - X_v}{h}\right)$$

  - Select:

$$p(v) = g(v)(rv + b)$$

Evaluation

Dataset

	Nodes	Edges	Init D	Init P
Twitter	548	3638	2.25	0.53
Reddit	556	8969	0.88	0.05
Facebook	4039	88234	8.74	2.58
Barabási-Albert	10000	99900	27.23	1.95
Erdős–Rényi	10000	199991	15.66	0.43
WS	10000	100000	30.60	7.48

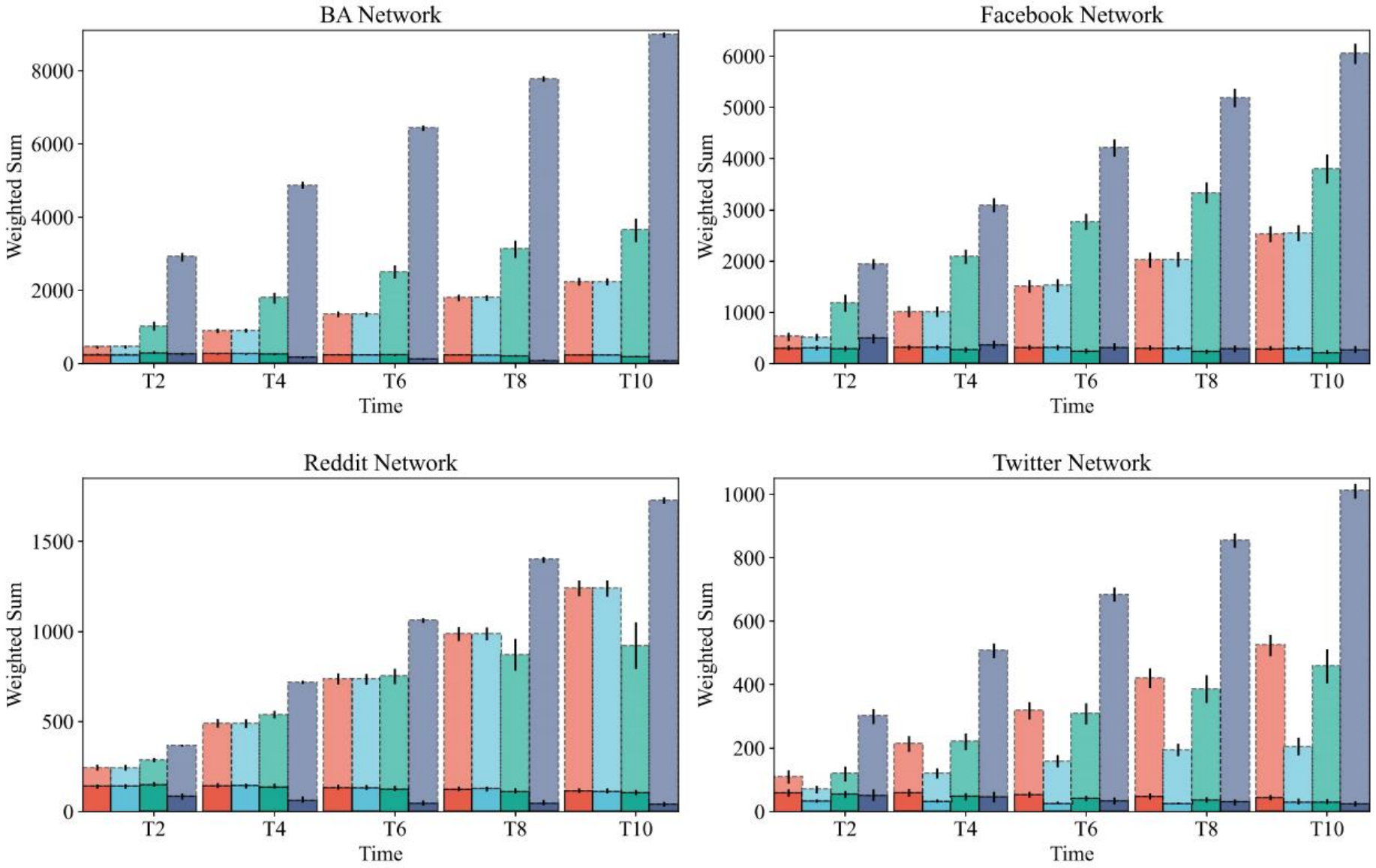
Baseline

- **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.
- **Risk-Only:** Selects the top-η nodes purely based on hijack risk.
- **Random:** Randomly selects η nodes as the protection set.

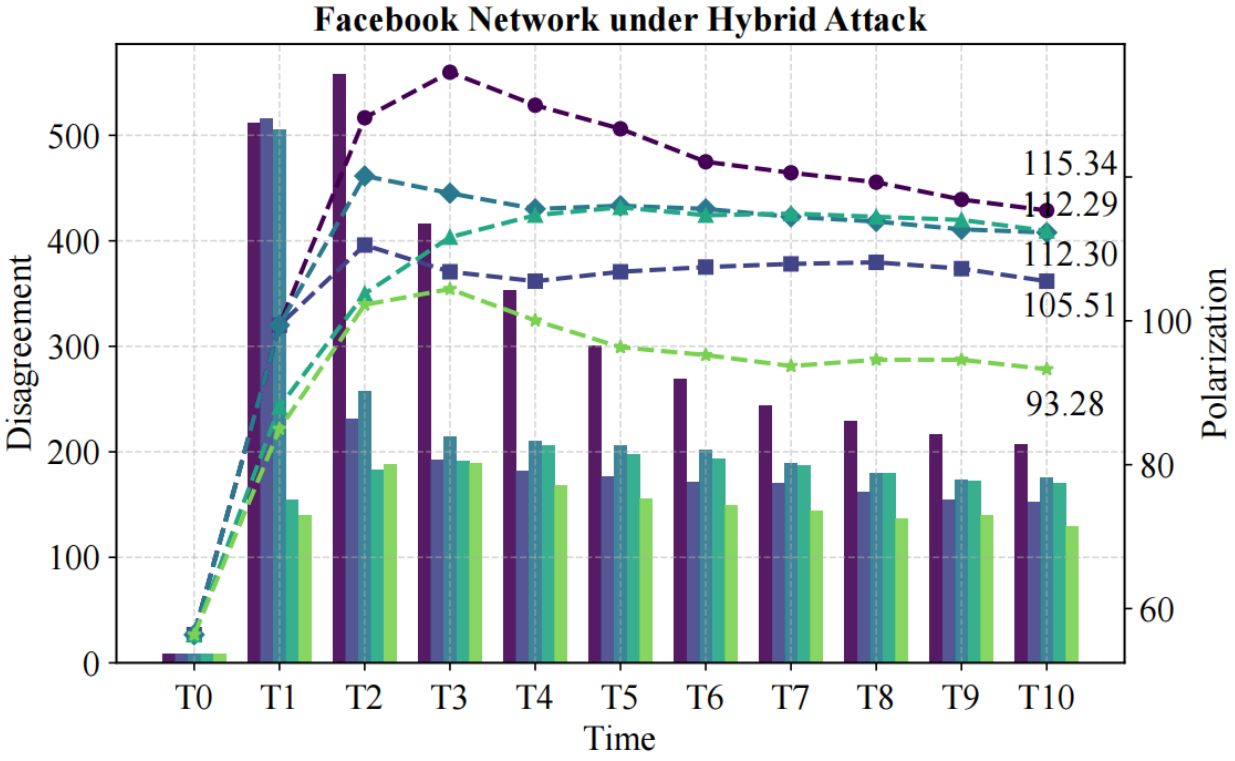
Attacker Setting

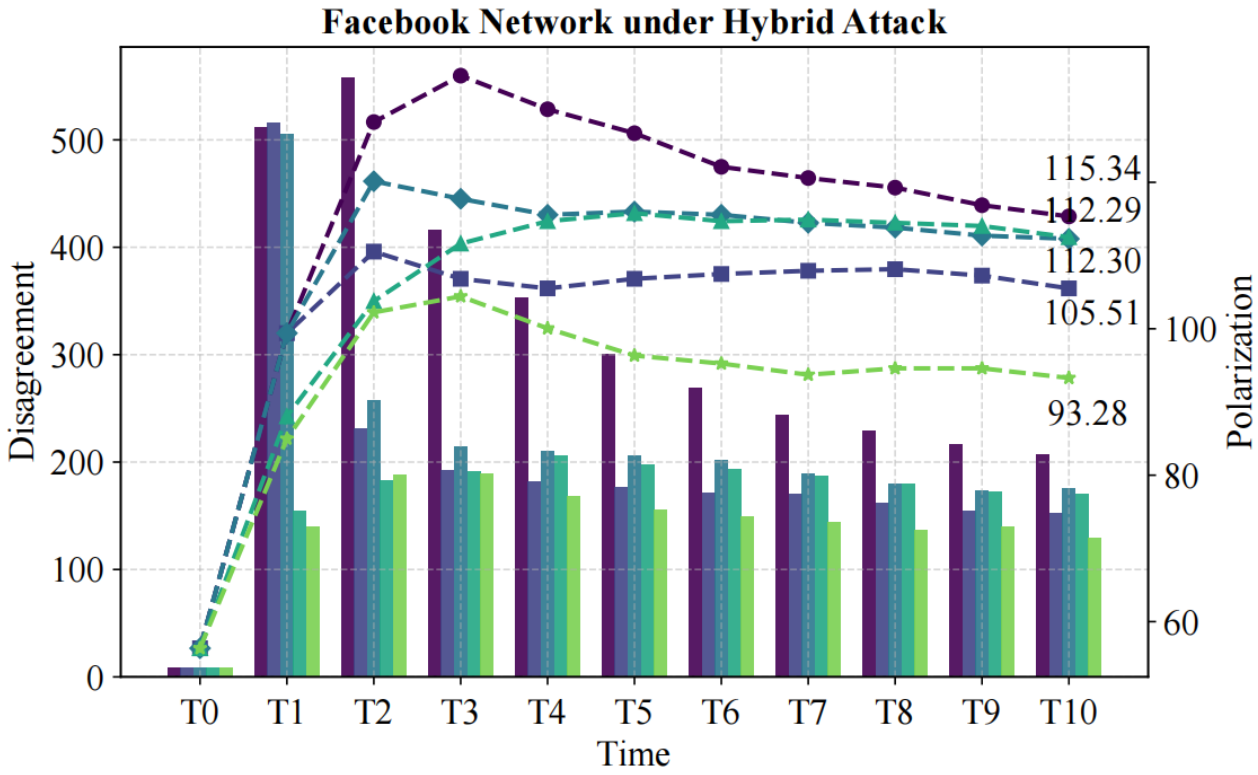
- **Structure-aware attackers:** prioritize high-degree nodes.
- **Feature-aware attackers:** target nodes whose opinion values fall within a specific range (e.g., [0.4, 0.6] around the network mean of 0.5).
- **Hybrid attackers:** jointly consider node degree and opinion value.
- **Random attackers:** serve as a baseline, selecting γ target nodes uniformly at random from unselected nodes.

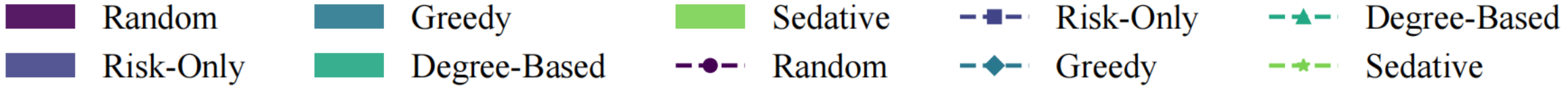
Overall Performance



Effect comparison







Robustness

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%

- ✓ structure-aware attacks are the most damaging ones — so this level of mitigation is quite impressive

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

Presenter: Wen Ma

Email: marvin2022@buaa.edu.cn