

BOCLOAK: Optimal Transport-Guided Adversarial Attacks on Graph Neural Network-Based Bot Detection

Kunal Mukherjee, Zulfikar Alom, Tran Gia Bao Ngo, Cunevt Gurcan Akcora, Murat Kantarcioglu



Motivation

- Social media bots threaten public discourse, yet the effectiveness of GNN-based bot detectors remains unclear.
- Existing graph attacks often assume **unrealistic powers**, such as arbitrary rewiring or full graph access, unlike real bots.
- **Real attackers face hard constraints**: limited observability, no forced human follow-backs, temporal plausibility, and only local edits around the bot.
- There is a **need for an attack framework** that **reflects these real-world constraints** while still being effective.
- **Optimal transport** offers a principled way to compare neighborhood structures and guide sparse, plausible edits.

Problem Definition

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, and a labelled nodes-subset with $y_v \in \{\text{human}, \text{bot}\}$. A bot detector perform binary classification for each node.

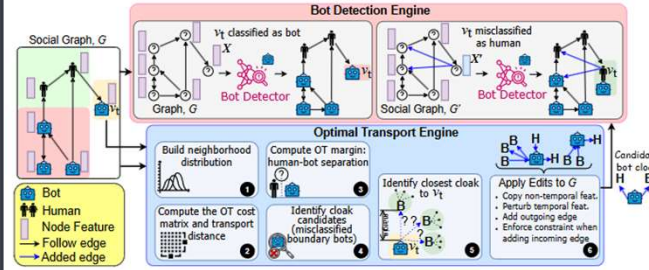
The attacker operates at test time by injecting or editing a existing bot under constraints:

$$\begin{aligned} \mathcal{F}(B) &= \{\Delta\mathcal{E} : |\Delta\mathcal{E}| \leq B, \Delta\mathcal{E} \subseteq (\{v_t\} \times \mathcal{V}) \cup (\mathcal{V} \times \{v_t\}), \\ &\Delta\mathcal{E} \cap (\mathcal{V}_{\text{hum}} \times \{v_t\}) = \emptyset, \\ &\Psi_{\text{dir}}(\Delta\mathcal{E}) \leq 0, \Psi_{\text{time}}(\Delta\mathcal{E}) \leq 0\}. \end{aligned}$$

Black box setting:

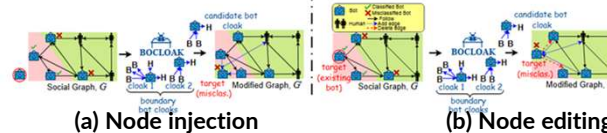
- ✓ Training graph + node labels + node features
- ✗ Victim model architecture, parameters, gradients or logit outputs

BOCLOAK Framework



BOCLOAK learns an optimal transport geometry over k-hop neighborhood distributions and uses transport plans to guide sparse, plausible local edge edits for evasion.

1. Build a weighted neighborhood distribution for each node
2. Learn OT-based geometry between bot and human neighborhoods
3. Find boundary bot "cloaks" near the human region
4. Compute transport plans to match target bots to cloak templates
5. Decode high-mass transport matches into feasible edge edit
6. Apply sparse, realistic edits to evade GNN-based bot detectors



(a) Node injection

(b) Node editing

BOCLOAK pipeline: (a) Injection constructs a new bot with a human-like neighborhood (b) Editing transforms an existing bot's neighborhood to induce misclassification

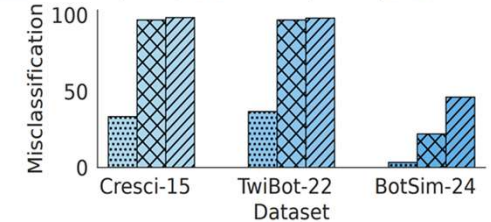
Experiments

Node Editing Attack ($\Delta = 1$): Misclassification rate (%) for flipping 50 correctly classified bots against GAT, BotRGCN, and S-HGN under adversarial defenses

Dataset	Attack	GAT				BotRGCN				S-HGN			
		Unconstrained	Vanilla	+GNNFeat	+GNNFeat + BotRGCN	Unconstrained	Vanilla	+GNNFeat	+GNNFeat + BotRGCN	Unconstrained	Vanilla	+GNNFeat	+GNNFeat + BotRGCN
Cresci-15	Nettack	96.46±0.0	8.17±0.0	6.59±0.0	2.28±0.0	7.28±0.0	96.47±0.0	4.32±0.0	3.78±0.0	2.62±0.0	94.11±0.0	6.04±0.0	4.81±0.0
	FGA	96.75±0.0	8.52±0.0	5.74±0.0	4.73±0.0	2.83±0.0	96.49±0.0	3.40±0.0	2.98±0.0	2.19±0.0	92.54±0.0	5.22±0.0	4.36±0.0
	PR-BCD	95.02±0.0	6.79±0.0	6.85±0.0	7.08±0.0	7.92±0.0	96.56±0.0	4.10±0.0	3.60±0.0	2.05±0.0	1.90±0.0	94.12±0.0	7.57±0.0
	GOMack	96.94±0.0	7.61±0.0	2.14±0.0	5.08±0.0	6.26±0.0	96.55±0.0	3.55±0.0	3.41±0.0	2.28±0.0	2.46±0.0	94.42±0.0	8.10±0.0
	BotCloak (ours)	95.25±0.0	93.18±0.0	96.29±0.0	95.45±0.0	95.48±0.0	95.36±0.0	96.34±0.0	96.05±0.0	96.75±0.0	96.88±0.0	95.48±0.0	95.48±0.0
Twibot-22	Nettack	81.22±0.0	13.51±0.0	11.92±0.0	8.59±0.0	12.01±0.0	79.77±0.0	9.51±0.0	8.67±0.0	4.67±0.0	83.51±0.0	5.00±0.0	3.26±0.0
	FGA	73.75±0.0	12.32±0.0	8.99±0.0	9.04±0.0	5.57±0.0	71.90±0.0	4.00±0.0	7.15±0.0	6.62±0.0	73.85±0.0	5.72±0.0	3.61±0.0
	PR-BCD	71.86±0.0	13.51±0.0	13.52±0.0	11.32±0.0	12.29±0.0	71.77±0.0	6.67±0.0	6.67±0.0	6.23±0.0	78.22±0.0	4.58±0.0	3.76±0.0
	GOMack	82.22±0.0	11.41±0.0	10.30±0.0	9.78±0.0	10.33±0.0	82.22±0.0	4.09±0.0	7.15±0.0	6.01±0.0	80.85±0.0	5.55±0.0	4.68±0.0
	BotCloak (ours)	95.25±0.0	96.48±0.0	95.73±0.0	95.98±0.0	95.93±0.0	96.47±0.0	96.47±0.0	96.47±0.0	95.83±0.0	97.42±0.0	96.48±0.0	95.11±0.0
BotSim-24	Nettack	53.77±0.0	5.52±0.0	3.74±0.0	4.73±0.0	4.93±0.0	56.57±0.0	0.00±0.0	0.00±0.0	0.00±0.0	55.54±0.0	4.71±0.0	3.64±0.0
	FGA	52.09±0.0	5.79±0.0	4.83±0.0	4.48±0.0	4.92±0.0	55.67±0.0	0.00±0.0	0.00±0.0	0.00±0.0	56.41±0.0	5.52±0.0	4.58±0.0
	PR-BCD	51.26±0.0	7.32±0.0	6.59±0.0	6.65±0.0	7.24±0.0	50.62±0.0	2.00±0.0	5.53±0.0	5.02±0.0	52.11±0.0	5.70±0.0	4.43±0.0
	GOMack	59.98±0.0	7.08±0.0	6.20±0.0	6.62±0.0	5.87±0.0	57.18±0.0	0.00±0.0	0.00±0.0	0.00±0.0	58.33±0.0	4.62±0.0	3.69±0.0
	BotCloak (ours)	95.25±0.0	96.74±0.0	96.60±0.0	95.96±0.0	96.40±0.0	96.23±0.0	96.23±0.0	96.40±0.0	96.40±0.0	96.75±0.0	96.36±0.0	96.96±0.0

BOCLOAK incurs **99.80% less** GPU footprint and runs up to **20 × faster** than SOTA attacks.

Random($\Delta \leq 5$) Random($\Delta = 20$) BOCLOAK($\Delta \leq 5$)



BOCLOAK achieves **injection success** with very small budgets

Ablation Study: Misclassification rate for node editing attack

Dataset	$-\lambda_{\text{BCE}}$	$-\lambda_{\text{sp}}$	$-\lambda_{\text{pl}}$	$-\alpha_{\text{deg}}$	$-\alpha_{\text{age}}$	BOCLOAK
Cresci-17	38.21±5.12	96.88±5.89	57.99±3.38	73.23±0.23	87.23±1.78	99.67±0.20
TwiBot-22	22.45±6.44	92.12±7.65	55.23±6.77	70.98±4.29	83.88±6.63	94.00±2.00
BotSim-24	32.45±5.64	96.55±0.66	60.34±5.09	69.22±2.15	85.24±1.38	99.33±1.15