



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, platform integrity, and trust in online information.
 - Effectiveness of GNN-based bot detectors under **realistic attack settings is still unclear**.
 - Existing graph attacks assume **unrealistic attacker power**, such as arbitrary rewiring or full graph access.
 - Real attackers **face hard constraints**: limited observability, no forced human follow-backs, temporal plausibility, and only local edits around the bot.
 - Need attack framework **that reflects these real-world constraints** while remaining effective.
 - **Optimal transport** provides way to compare neighborhood structures and guide sparse, plausible edits.
- 

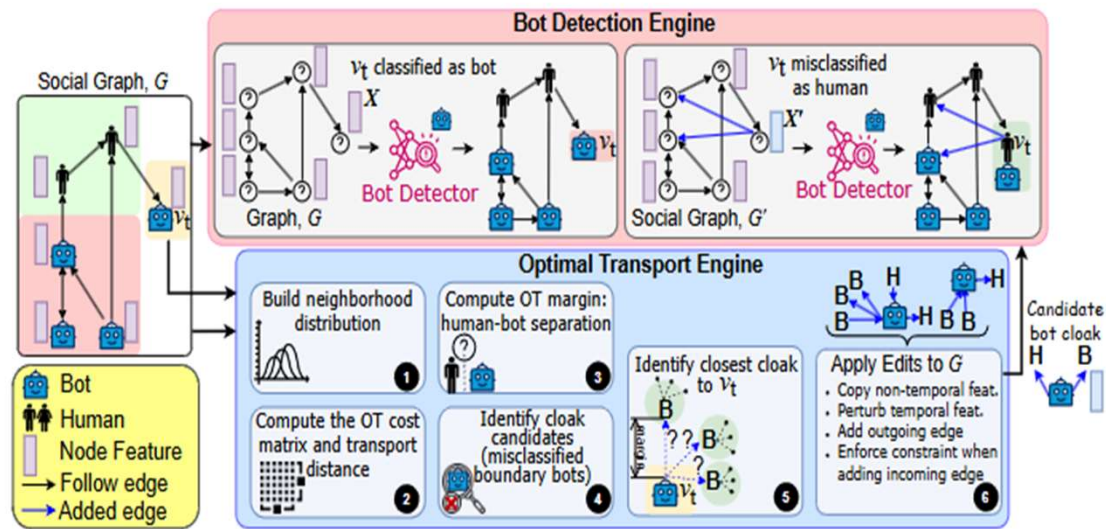
PROBLEM SETUP

- Consider a graph-based bot detection problem with node features and node labels.
- Detector performs binary classification for each node.
- Attacker operates at test time by injecting a new bot or editing an existing bot.
- Attack is performed in a black-box setting.

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

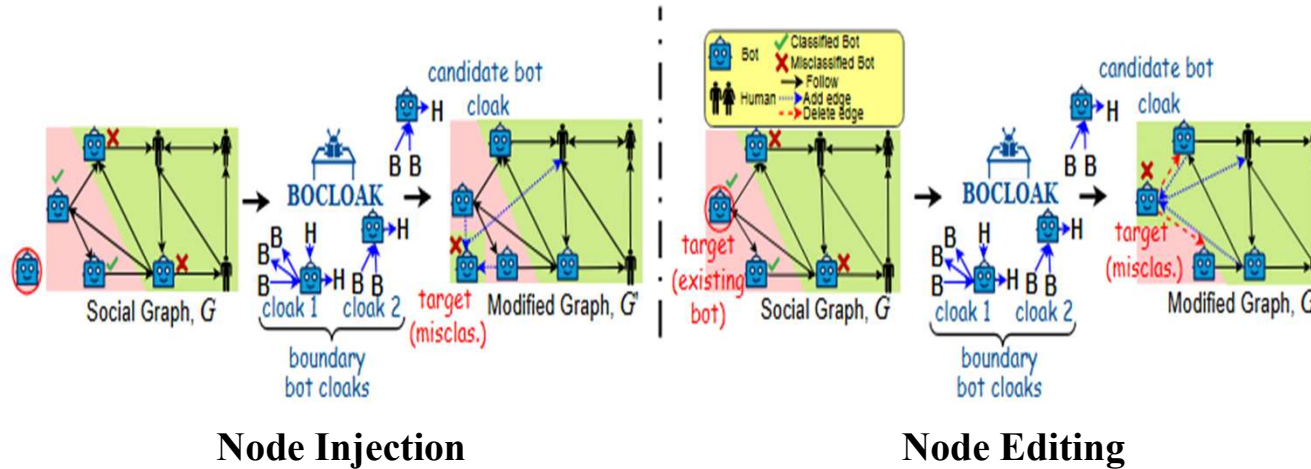
- Available information: training graph, node labels, and node features.
- Unknown information: victim model architecture, parameters, gradients, and logit outputs..

BOCLOAK FRAMEWORK



- Learns an **optimal transport geometry** over k-hop neighborhood distributions.
- Compares **bot and human neighborhood structures** using transport plans.
- Identifies boundary bot “**cloaks**” near the human region.
- Matches **target bots to cloak templates** through transport-based alignment.
- Decodes **high-mass transport matches into feasible edge edits**.
- Applies **sparse, realistic local edits** to evade GNN-based bot detectors.

ATTACK MODES



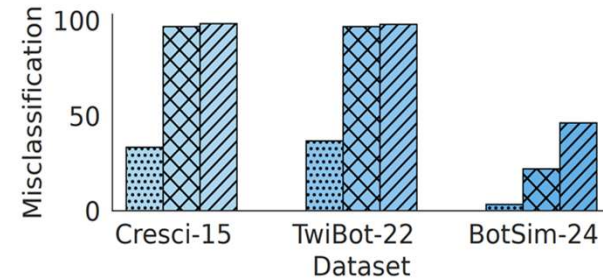
- **Node injection:** construct a new bot with a human-like neighborhood.
- **Node editing:** transform an existing bot's neighborhood to induce misclassification.

EXPERIMENTS

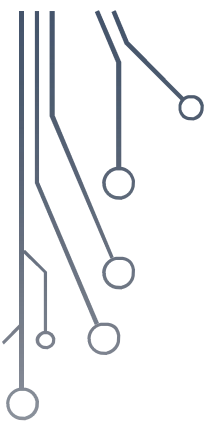
Dataset	Attack	GAT					BotRGCN					S-HGN				
		Unconstrained	Constrained				Unconstrained	Constrained				Unconstrained	Constrained			
			Vanilla	+GNNGuard	+GRAND	+RobustGCN		Vanilla	+GNNGuard	+GRAND	+RobustGCN		Vanilla	+GNNGuard	+GRAND	+RobustGCN
Cresci-15	Nettack	95.45 $\pm$ 3.22	8.17 $\pm$ 3.23	6.59 $\pm$ 3.40	7.54 $\pm$ 2.07	7.28 $\pm$ 3.19	94.67 $\pm$ 6.55	4.32 $\pm$ 2.10	3.78 $\pm$ 1.95	3.21 $\pm$ 1.88	2.67 $\pm$ 1.72	94.11 $\pm$ 2.45	6.04 $\pm$ 3.00	4.81 $\pm$ 3.18	4.51 $\pm$ 3.60	4.88 $\pm$ 2.71
	FGA	94.75 $\pm$ 1.50	8.52 $\pm$ 1.92	5.74 $\pm$ 2.39	4.73 $\pm$ 3.67	7.93 $\pm$ 2.91	96.89 $\pm$ 4.81	3.95 $\pm$ 2.44	3.40 $\pm$ 2.12	2.98 $\pm$ 1.76	2.15 $\pm$ 1.60	92.54 $\pm$ 1.23	5.22 $\pm$ 3.17	4.96 $\pm$ 3.09	4.31 $\pm$ 2.07	3.44 $\pm$ 2.01
	PR-BCD	95.02 $\pm$ 1.38	8.79 $\pm$ 1.91	6.85 $\pm$ 2.35	7.48 $\pm$ 2.37	7.92 $\pm$ 3.24	96.36 $\pm$ 8.36	4.10 $\pm$ 3.05	3.60 $\pm$ 2.80	2.85 $\pm$ 2.20	1.90 $\pm$ 1.45	94.12 $\pm$ 1.23	7.57 $\pm$ 3.84	6.91 $\pm$ 2.43	6.70 $\pm$ 2.84	5.58 $\pm$ 2.45
	Gottack	94.98 $\pm$ 5.11	7.61 $\pm$ 2.91	7.14 $\pm$ 3.65	5.08 $\pm$ 3.06	6.20 $\pm$ 1.94	98.54 $\pm$ 12.40	3.55 $\pm$ 2.34	3.41 $\pm$ 2.52	2.28 $\pm$ 1.22	2.66 $\pm$ 1.67	94.12 $\pm$ 1.21	8.10 $\pm$ 2.83	7.57 $\pm$ 2.04	7.92 $\pm$ 1.87	6.39 $\pm$ 2.12
	BOCLOAK (ours)	95.25 $\pm$ 1.11	93.10 $\pm$ 2.34	94.20 $\pm$ 1.89	92.45 $\pm$ 3.11	91.88 $\pm$ 4.56	99.34 $\pm$ 0.29	99.34 $\pm$ 0.29	98.91 $\pm$ 0.54	99.06 $\pm$ 0.41	99.72 $\pm$ 0.18	95.80 $\pm$ 2.50	92.18 $\pm$ 4.73	91.60 $\pm$ 2.15	91.92 $\pm$ 3.08	91.40 $\pm$ 1.67
Twibot-22	Nettack	81.22 $\pm$ 1.23	13.51 $\pm$ 8.15	11.92 $\pm$ 9.38	8.59 $\pm$ 7.73	12.91 $\pm$ 8.91	79.77 $\pm$ 6.77	9.33 $\pm$ 3.06	8.67 $\pm$ 4.62	4.67 $\pm$ 5.03	8.00 $\pm$ 5.29	83.51 $\pm$ 1.22	5.03 $\pm$ 4.72	3.24 $\pm$ 9.20	3.72 $\pm$ 10.81	4.73 $\pm$ 8.85
	FGA	73.75 $\pm$ 1.22	12.32 $\pm$ 8.10	8.59 $\pm$ 4.52	9.94 $\pm$ 6.41	5.57 $\pm$ 10.78	71.90 $\pm$ 4.23	4.00 $\pm$ 2.00	7.33 $\pm$ 2.31	6.67 $\pm$ 4.16	1.33 $\pm$ 6.43	75.85 $\pm$ 2.54	5.79 $\pm$ 6.52	4.57 $\pm$ 4.96	3.41 $\pm$ 6.87	3.19 $\pm$ 9.89
	PR-BCD	73.86 $\pm$ 1.50	13.51 $\pm$ 4.52	12.52 $\pm$ 9.38	11.32 $\pm$ 6.41	12.24 $\pm$ 4.52	75.77 $\pm$ 6.05	8.67 $\pm$ 2.31	8.67 $\pm$ 5.03	6.23 $\pm$ 4.16	7.33 $\pm$ 2.31	78.12 $\pm$ 1.98	4.58 $\pm$ 5.29	3.74 $\pm$ 3.34	3.61 $\pm$ 7.20	3.20 $\pm$ 9.92
	Gottack	82.12 $\pm$ 2.56	11.41 $\pm$ 7.59	10.50 $\pm$ 6.38	9.78 $\pm$ 3.93	10.33 $\pm$ 5.34	84.29 $\pm$ 11.63	6.89 $\pm$ 3.33	7.15 $\pm$ 4.25	6.01 $\pm$ 2.48	5.32 $\pm$ 2.58	80.85 $\pm$ 2.30	5.55 $\pm$ 7.98	4.04 $\pm$ 8.93	4.44 $\pm$ 7.82	4.95 $\pm$ 8.39
	BOCLOAK (ours)	95.25 $\pm$ 2.50	94.38 $\pm$ 3.41	93.72 $\pm$ 1.66	93.90 $\pm$ 4.28	95.12 $\pm$ 3.33	86.67 $\pm$ 2.31	86.67 $\pm$ 2.31	84.67 $\pm$ 11.02	83.67 $\pm$ 3.06	87.33 $\pm$ 1.15	98.12 $\pm$ 1.75	97.42 $\pm$ 1.39	96.15 $\pm$ 3.74	95.90 $\pm$ 2.06	95.11 $\pm$ 4.28
BotSim-24	Nettack	53.77 $\pm$ 4.12	5.52 $\pm$ 1.92	3.74 $\pm$ 2.39	4.73 $\pm$ 3.67	4.93 $\pm$ 8.10	56.57 $\pm$ 6.63	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	3.00 $\pm$ 2.00	55.54 $\pm$ 5.55	4.71 $\pm$ 2.19	3.44 $\pm$ 3.09	3.35 $\pm$ 2.67	3.65 $\pm$ 7.44
	FGA	52.09 $\pm$ 1.59	5.79 $\pm$ 1.91	4.85 $\pm$ 2.35	4.48 $\pm$ 2.37	4.92 $\pm$ 3.24	55.67 $\pm$ 6.96	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	54.98 $\pm$ 1.33	0.00 $\pm$ 0.00	50.41 $\pm$ 2.50	5.72 $\pm$ 2.56	4.63 $\pm$ 2.12	5.06 $\pm$ 3.15	4.54 $\pm$ 3.73
	PR-BCD	51.26 $\pm$ 4.11	7.32 $\pm$ 8.10	6.59 $\pm$ 4.52	6.65 $\pm$ 7.59	7.14 $\pm$ 3.65	50.62 $\pm$ 8.04	2.00 $\pm$ 2.00	5.33 $\pm$ 2.31	0.67 $\pm$ 1.15	0.00 $\pm$ 0.00	52.11 $\pm$ 2.50	5.20 $\pm$ 8.04	4.24 $\pm$ 4.63	4.43 $\pm$ 4.99	4.26 $\pm$ 1.88
	Gottack	91.50 $\pm$ 1.50	7.08 $\pm$ 3.06	6.20 $\pm$ 1.94	6.67 $\pm$ 3.09	5.87 $\pm$ 2.27	57.18 $\pm$ 2.32	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	88.33 $\pm$ 1.33	4.42 $\pm$ 1.83	3.25 $\pm$ 3.38	3.65 $\pm$ 3.12	3.21 $\pm$ 2.45
	BOCLOAK (ours)	90.19 $\pm$ 1.50	88.74 $\pm$ 3.12	90.05 $\pm$ 1.47	89.30 $\pm$ 4.66	88.40 $\pm$ 2.29	58.22 $\pm$ 7.20	58.22 $\pm$ 7.20	52.68 $\pm$ 1.15	54.98 $\pm$ 2.45	55.10 $\pm$ 3.98	90.75 $\pm$ 2.33	89.16 $\pm$ 4.25	90.44 $\pm$ 2.11	89.78 $\pm$ 1.63	88.90 $\pm$ 3.74

Attack	BotSim-24		
	Time (sec) ($\downarrow$)	RAM (MB) ($\downarrow$)	GPU (MB) ($\downarrow$)
Nettack	19.7 $\pm$ 1.65	2973 $\pm$ 628	156 $\pm$ 40
FGA	0.3 $\pm$ 1.57	1720 $\pm$ 13	286 $\pm$ 2
PR-BCD	2.1 $\pm$ 0.75	1857 $\pm$ 21	286 $\pm$ 1
Gottack	7.5 $\pm$ 1.75	2069 $\pm$ 623	240 $\pm$ 2
BOCLOAK	0.1 $\pm$ 0.13	1450 $\pm$ 2	11 $\pm$ 1

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



- Evaluate misclassification rate under node editing attacks.
- Achieves injection success with very small budgets
- Uses 99.80% less GPU footprint and runs up to 20 $\times$ faster than prior attacks.



BoCloak

Email: mkunal@vt.edu

<https://github.com/kunmukh/bocloak>

