Enhancing Robustness of Graph Neural Networks on Social Media with Explainable Inverse Reinforcement Learning

Yuefei Lyu[1], Chaozhuo Li[1*], Sihong Xie[2], Xi Zhang[1*]

[1]Key Laboratory of Trustworthy Distributed Computing and Service (BUPT)
Ministry of Education, Beijing University of Posts and Telecommunications, *Beijing, China*
[2]Artificial Intelligence Thrust,
The Hong Kong University of Science and Technology (Guangzhou), *China*
*Corresponding Authors: lichaozhuo@bupt.edu.cn, zhangx@bupt.edu.cn

Background

Social Network

Graph Neural Networks (GNNs)

Social Network Tasks

- Rumor Detection
- Spam Detection
- Stance Detection
- ……

Graph Adversarial Attack

Analyze

Adversarial Samples

Defend

Motivation

- Adversarial Training 👉 Enhance the robustness
  - Augment model generalization by introducing perturbed samples into training data
  - Depend on effective attack methods to generate adversarial samples
- Numerous attackers with diverse goals and styles
  - Insufficient defense
- The aim is to reconstruct the attack policy
  - Simulate various attackers
  - Make use of the adversarial samples captured by social media

Motivation

- Sequential attack samples 👉 Inverse Reinforcement Learning
  - Deduce the unknown reward function with expert demonstrations
  - Provide explanations with linear reward functions and interpretable features

Challenges

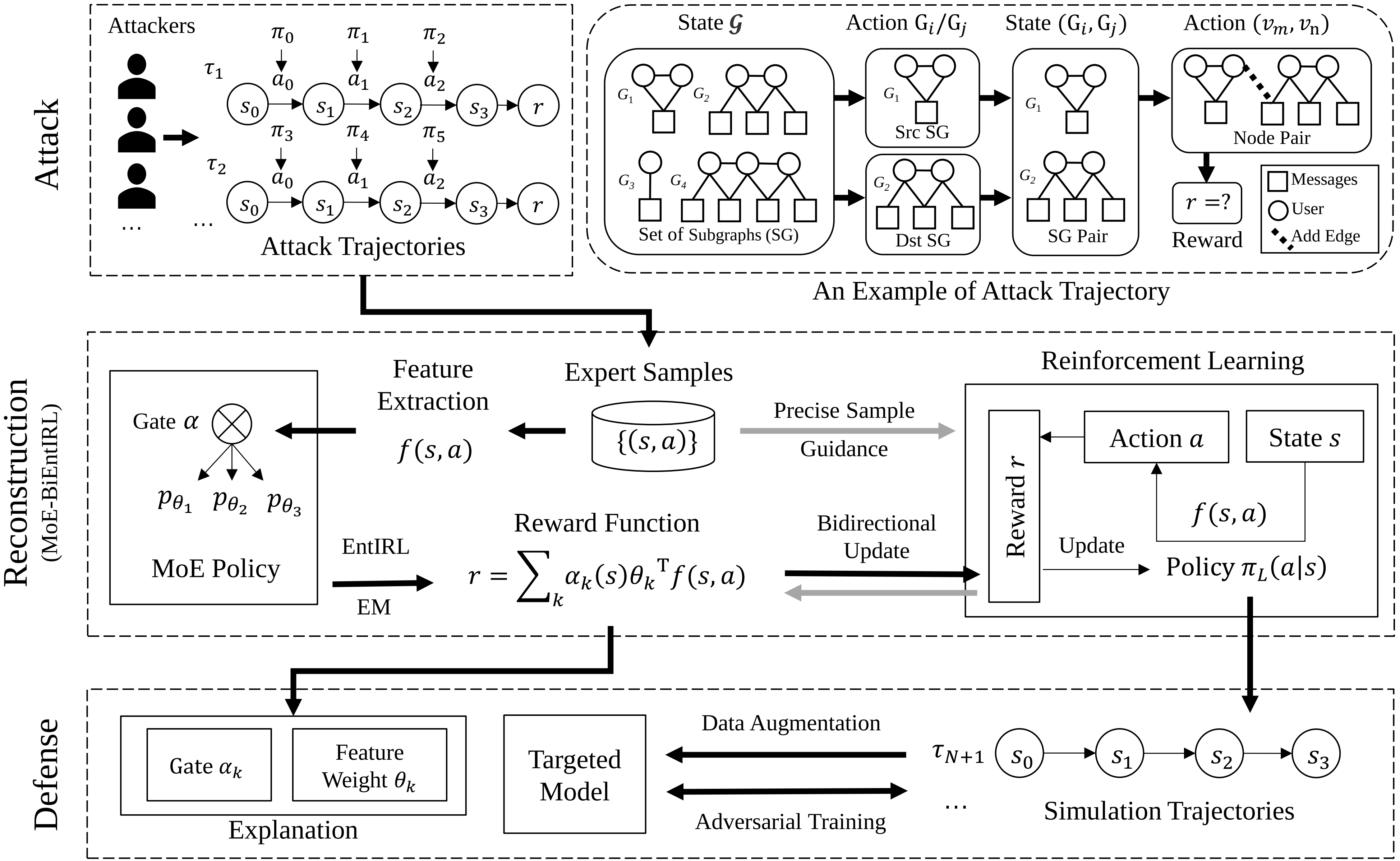
- **Reconstructing interpretable attack policies in social networks using inverse reinforcement learning (IRL)**
- **C1: Expert demonstrations from diverse attackers**
  - **Improve IRL to integrate various attack policies**
- **C2: Imprecise feature representation**
  - **Similar sample features → disparate ground true rewards**

Solution

- We purpose MoE-BiEntIRL
  - Improve maximum entropy inverse reinforcement learning (EntIRL)
  - Make use of mixture-of-experts (MoE) model (for C1)
  - Introduce precise sample guidance and bidirectional update mechanism (for C2)

- Contributions
  - Novel problem: reconstructing the attack policy with collected adversarial samples on social media
  - Enhance IRL to handle the attack samples in social graphs
  - Validate the policy reconstruction effectiveness and robustness enhancement

Method


Attack
Attackers
π_0 π_1 π_2
τ_1
a_0 a_1 a_2
s_0 s_1 s_2 s_3 r
π_3 π_4 π_5
τ_2
a_0 a_1 a_2
s_0 s_1 s_2 s_3 r
… …
Attack Trajectories
State $\mathcal{G}$
Action G_i/G_j
State (G_i, G_j)
Action (v_m, v_n)
G_1 G_2 G_3 G_4
Set of Subgraphs (SG)
Src SG
Dst SG
SG Pair
Node Pair
r =?
Reward
Messages
User
Add Edge
An Example of Attack Trajectory
Reconstruction
(MoE-BiEntIRL)
Gate α
p_{θ_1} p_{θ_2} p_{θ_3}
MoE Policy
Feature Extraction
$f(s,a)$
Expert Samples
$\{(s,a)\}$
Precise Sample Guidance
Reinforcement Learning
Action a
State s
Reward r
$f(s,a)$
Update
Policy $\pi_L(a|s)$
EntIRL
EM
Reward Function
$r = \sum_k \alpha_k(s)\theta_k{}^{\mathrm{T}} f(s,a)$
Bidirectional Update
Defense
Gate α_k
Feature Weight θ_k
Explanation
Targeted Model
Data Augmentation
Adversarial Training
τ_{N+1} s_0 s_1 s_2 s_3
…
Simulation Trajectories

Method

- **EntIRL[1] (locally optimal example[2,3])**
  - **Action probability**
    $$p(a|s) = \frac{1}{\boxed{Z}}\exp(r_\theta(s,a)),$$
    Partition factor
  - **Linear reward function**
    $$r_\theta(s,a) = \theta^\top \boxed{f(s,a)}$$
    Feature extraction function
  - **Loss function**
    $$L(\theta) = \sum_{a\in\boxed{\mathcal{A}_s}} \log p(a|s),$$
    Action space for state s
- **MoE policy**

$$p(a^{(t)}|s^{(t)},\theta) = \sum_{k=1}^{K} \underbrace{\alpha_k(s^{(t)},\varphi)}_{\text{Gate}}\underbrace{p(a^{(t)}|s^{(t)},\theta_k)}_{\text{Expert}},$$

$$p(a^{(t)}|s^{(t)},\theta_k) = \frac{\exp(\theta_k^\top f(s^{(t)},a^{(t)}))}{\boxed{\sum_{a\in\boxed{\mathcal{A}_{s,t}}}\exp(\theta_k^\top f(s^{(t)},a))}},$$

Estimated by sampling

Action space for state $s^{(t)}$

Method

- **EM algorithm**
  - **The likelihood function of complete data**

$$P(\tau, \gamma|\theta) = \prod_{j=1}^{N} P(\tau_j, \gamma_{j,1,0}, \gamma_{j,2,0}, \dots, \gamma_{jKT})$$

$$\gamma_{\text{ikt}} = \begin{cases} 1, & \text{if the } t\text{-th pair of } \tau_j \text{ is decided by the } k\text{-th expert} \\ 0, & \text{otherwise} \end{cases}$$

  - **E-Step & M-Step**

$$\mathcal{Q}(\theta, \theta^{(i)}) = \mathbb{E}\left[\log P(\tau, \gamma|\theta)|a_j^{(t)}, s_j^{(t)}, \theta^{(i)}\right] \qquad \theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^{(i)}).$$

  - **Gradient ascent**

$$L_{gate}(\varphi) = \sum_{t=0}^{T-1} \sum_{k=1}^{K} \sum_{j=1}^{N} \hat{\gamma}_{jkt} \log \alpha_k(s_j^{(t)}),$$

$$\boxed{L_{ex}(\theta_k) = \sum_{t=0}^{T-1} \sum_{j=1}^{N} \hat{\gamma}_{jkt} \log p(a_j^{(t)}|s_j^{(t)}, \theta_k).}$$

Match the EntIRL loss

$$\nabla L_{ex}(\theta_k) = \tilde{f}_k - \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{j=1}^{N} \hat{\gamma}_{jkt} \boxed{\sum_{a \in \mathcal{A}_{s_j,t}}} p(a|s_j^{(t)}, \theta_k) f(s_j^{(t)}, a).$$

Involving sampling

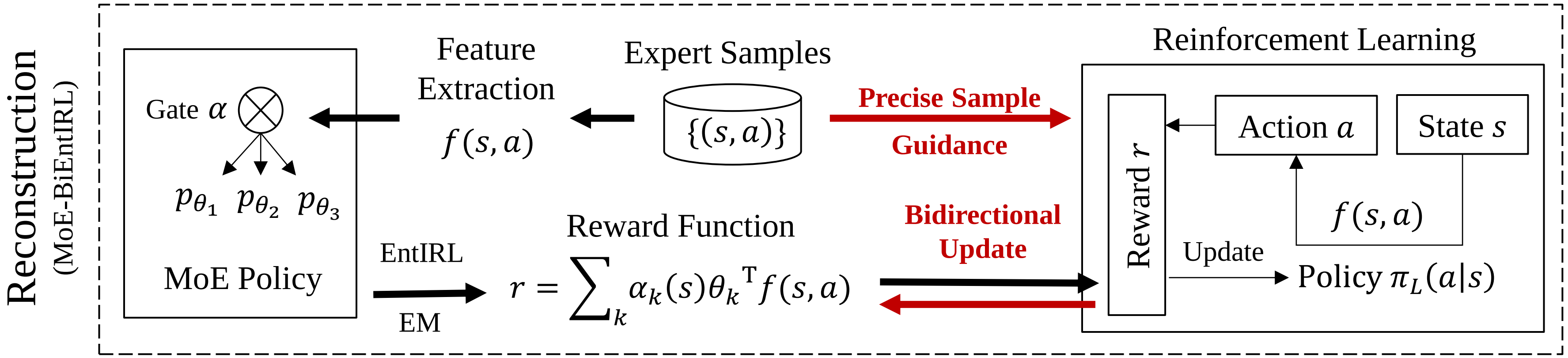
$$\tilde{f}_k = \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{j=1}^{N} \hat{\gamma}_{jkt} f(s_j^{(t)}, a_j^{(t)}).$$

- **Reward Function**

$$r_\theta(s, a) = \sum_{k=1}^{K} \alpha_k(s) \theta_k^\top f(s, a).$$

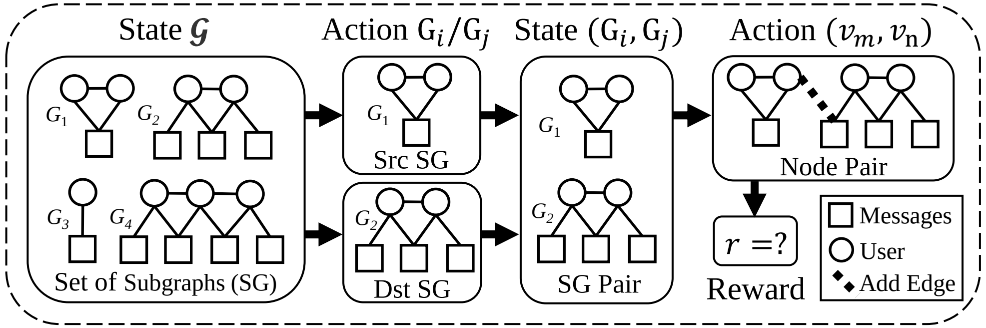
Method

- **Precise sample guidance**
  - **Introduce expert structural perturbations directly during the policy learning early process**
  - **Avoid the accumulated deviations in imprecise feature representation**
- **Bidirectional update mechanism**
  - **Provide feedback opposite to the output of the reward function**
  - **Ensure synchronized learning of the learner policy and the reward function**



Method

- **Hierarchical reinforcement learning**
  - **the source subgraph, the destination subgraph, and the node pair**



- **Defense with adversarial samples**
  - **data augmentation or adversarial training**

Ground truth Prediction on clean graph / perturbed graph

$$L_D = \sum_i L_{\sigma}(y_i, y_i') + \beta \sum_i L_{\sigma,\omega}(y_i, \tilde{y_i}').$$

Target model parameters   Learner policy parameters

Experiments

- **Datasets**
  - **For rumor detection task**

Table 1: Dataset statistics.

	Weibo	Pheme
Nodes	10,280	2,708
Edges	16,412	4,401
Rumors	1,538	284
Non-rumors	1,849	859
Users	2,440	1,008
Comments	4,453	557

- **Target model：**
  - **GCN rumor detector**

- **Metric：the reduction in the attack loss**

$$\Delta L_A = \boxed{L_A(0)} - \boxed{L_A(T)}.$$

The attack loss on clean graph / after T-step attacks

  - **Attack loss**

$$L_A = \sum_{v_i \in \boxed{\mathcal{O}}} (\boxed{g(v_i)} - \boxed{y_i})$$

Rumor probability of node v_i

Target node set Ground truth

Experiments

- **The Performance of Policy Reconstruction**
  - **Attack methods**
    - **PageRank, GC-RWCS[4], PR-BCD[5], AdRumor-RL[6]**
  - **Baselines**
    - **Apprenticeship Learning[7], EntIRL[1]**

		High-Cost Attack			Low-Cost Attack		
		PRBCD	*AdRumor*	Mixture	*PageRank*	*GC-RWCS*	Mixture
Weibo T=5	Expert	4.865	4.877	-	3.000	3.000	-
	Apprenticeship	1.275	0.788	0.704	0.850	0.763	1.071
	EntIRL	4.650	4.770	4.550	**5.000**	**4.950**	**4.950**
	MoE-BiEntIRL	**4.989**	**4.990**	**4.929**	4.860	4.900	4.900
Weibo T=20	Expert	19.521	19.854	-	5.449	5.160	-
	Apprenticeship	1.142	3.066	3.945	0.030	0.040	0.020
	EntIRL	19.030	19.749	19.199	19.830	**20.000**	**20.000**
	MoE-BiEntIRL	**19.876**	**19.936**	**19.979**	**19.970**	19.700	18.749
Pheme T=5	Expert	4.804	5.947	-	2.991	3.990	-
	Apprenticeship	1.788	3.387	2.619	0.000	0.000	0.000
	EntIRL	0.000	0.018	0.010	0.000	0.062	0.000
	MoE-BiEntIRL	**2.205**	**4.965**	**4.277**	**1.488**	**2.105**	**1.549**

Experiments

- **The improvement of robustness with generated samples**
  - **No defense (w/o Def.)**
  - **Data augmentation with expert samples (EDA)**
  - **Data augmentation with generated samples (DA)**
  - **Adversarial training (AT)**

		w/o Att.	*PageRank*	*GC-RWCS*	*PR-BCD*
	w/o Def.	70.4031	-0.4042	-0.4406	-0.1966
EDA	*PageRank*	70.5998	**-0.1821**	**-0.2440**	**0.0000**
EDA	*GC-RWCS*	70.7965	-0.4043	-0.4407	-0.1967
EDA	*PR-BCD*	70.3048	-0.2185	**-0.2440**	**0.0000**
EDA	*AdRumor-RL*	70.7965	∗-0.2076	**-0.2440**	**0.0000**
EDA	All above	70.7965	-0.2805	-0.3424	-0.0984
DA	*PageRank*	70.6981	-0.5025	-0.5390	-0.2950
DA	*GC-RWCS*	70.5015	-0.3059	-0.2440	0.0000
DA	*PR-BCD*	70.4031	**-0.1092**	**-0.1456**	**0.0984**
DA	*AdRumor-RL*	70.6981	-0.3059	-0.3423	-0.0983
DA	*MoE-BiEntIRL*	70.6981	**-0.1092**	**-0.1456**	**0.0984**
AT	*PageRank*	71.0914	**-0.2075**	**-0.2440**	**0.0000**
AT	*GC-RWCS*	70.2065	-0.4042	-0.4407	-0.1967
AT	*PR-BCD*	70.4031	-0.3059	-0.3423	-0.0983
AT	*AdRumor-RL*	70.6981	-0.3059	-0.3423	-0.0983
AT	*MoE-BiEntIRL*	72.0747	∗-0.2731	∗-0.2589	**0.0000**

[1] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al Maximum entropy inverse reinforcement learning. AAAI 2008.
[2] Sergey Levine and Vladlen Koltun. Continuous inverse optimal control with locally optimal examples. ICML 2012.
[3] Nathan D. Ratliff, Brian D. Ziebart, Kevin M. Peterson, J. Andrew Bagnell, Martial Hebert, Anind K. Dey, and Siddhartha S. Srinivasa. Inverse optimal heuristic control for imitation learning. AISTATS 2009.
[4] Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. Towards more practical adversarial attacks on graph neural networks. Advances in neural information processing systems, 2020.
[5] Simon Geisler, Tobias Schmidt, Hakan Sirin, Daniel Zügner, Aleksandar Bojchevski, and Stephan Günnemann. Robustness of graph neural networks at scale. NeurIPS 2021.
[6] Y uefei Lyu, Xiaoyu Yang, Jiaxin Liu, Sihong Xie, Philip S. Y u, and Xi Zhang. Interpretable and effective reinforcement learning for attacking against graph-based rumor detection. IJCNN 2023.
[7] Pieter Abbeel and Andrew Y . Ng. Apprenticeship learning via inverse reinforcement learning. ICML 2004.

Conclusion

- **MoE-BiEntIRL: a threat model to recover the graph adversarial attack policy against GNN model on social media**
  - ➢ **IRL techniques and MoE mindset**
  - ➢ **feature-level explanations**
  - ➢ **precise sample guidance and bidirectional update mechanism**
- **Enhance the robustness of the target model with samples produced by the reconstructed policy**

ACKNOWLEDGMENT
Yuefei Lyu, Chaozhuo Li and Xi Zhang were supported by the Natural Science Foundation of China (No. 62372057). Sihong Xie was supported in part by the National Key R&D Program of China (Grant No. YFF0725001), the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0008), and Education Bureau of Guangzhou Municipality. This material is based upon work supported by the National Science Foundation under Grant Number 2008155 & 1931042.

Thanks