2024 YINCHUAN CHINA MM

基于社交机器人的多媒体传播安全可控

王岚君

2024.7

报告内容

- 研究背景及意义
- 恶意机器人干扰虚假信息检测
- 可控机器人干预控制传播范围
- 机器人检测及鲁棒性评估优化
- 总结与展望

研究背景及意义

■ 网络媒体的发展变迁

专业生成内容 PGC
Professionally Generated Content

用户生成内容 UGC
User Generated Content

人工智能生成内容 AIGC
Artificial Intelligence Generated Content

研究背景及意义

■ 网络媒体内容的传播力、引导力、影响力巨大

国际层面	国家层面	商业层面
		
中印边界冲突 网络媒体内容影响国家政治地位和外交环境	**香港暴乱** 网络媒体内容影响社会管理和公共安全	**空姐滴滴遇害** 网络媒体内容影响产品口碑和客户流量

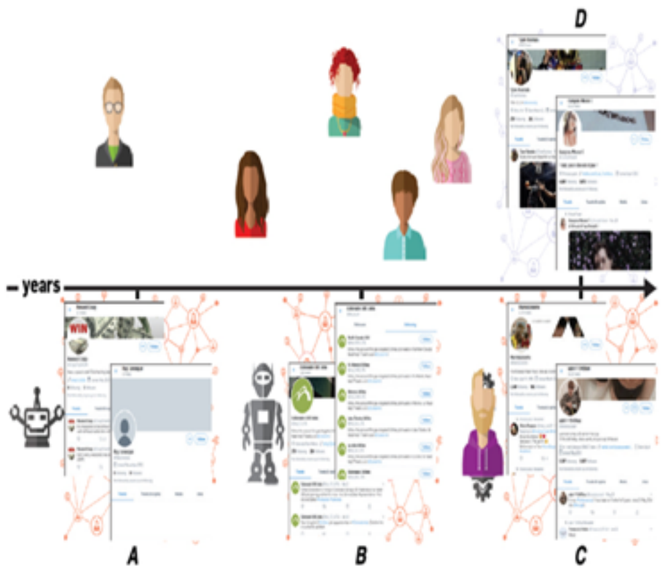
网络媒体传播安全可控是国家重大需求

研究背景及意义

■ 社交机器人（Social bots），是指一种运行于社交平台上、自动生成内容并参与人类社交互动的、无物质实体的自动程序型智能体[1]。

人类账号

拟人化机器人账号



影响大，难管控

2016美国大选中，支持特朗普和希拉里的推文中分别有36.1%，23.5%由社交网络机器人驱动。

俄乌战争，认知域战争

网络媒体平台呈现“人+社交机器人”共生的状态

[1] 中国新媒体发展报告 No.11(2020)

研究背景及意义

■ 社交机器人广泛存在于社交网络，对传播安全造成的影响未知

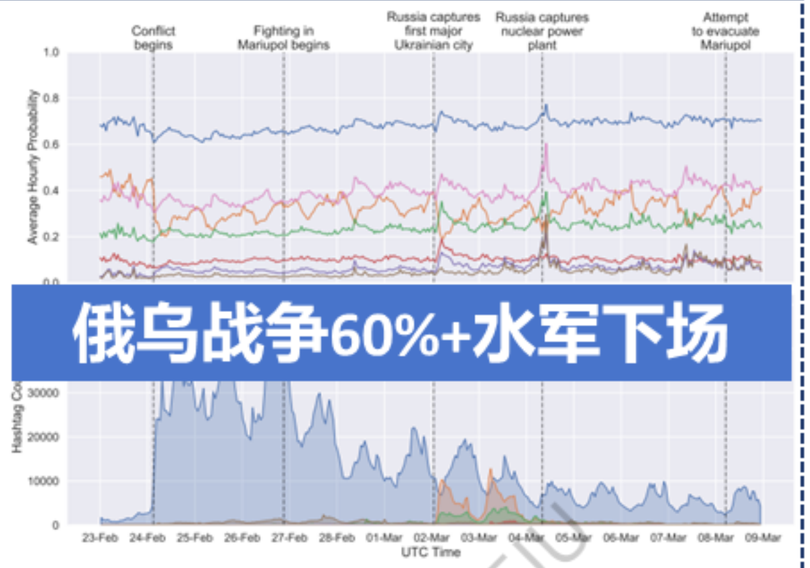
传播内容

机器人传播ISIS极端信息

传播路径

机器人账户灌注虚假流量

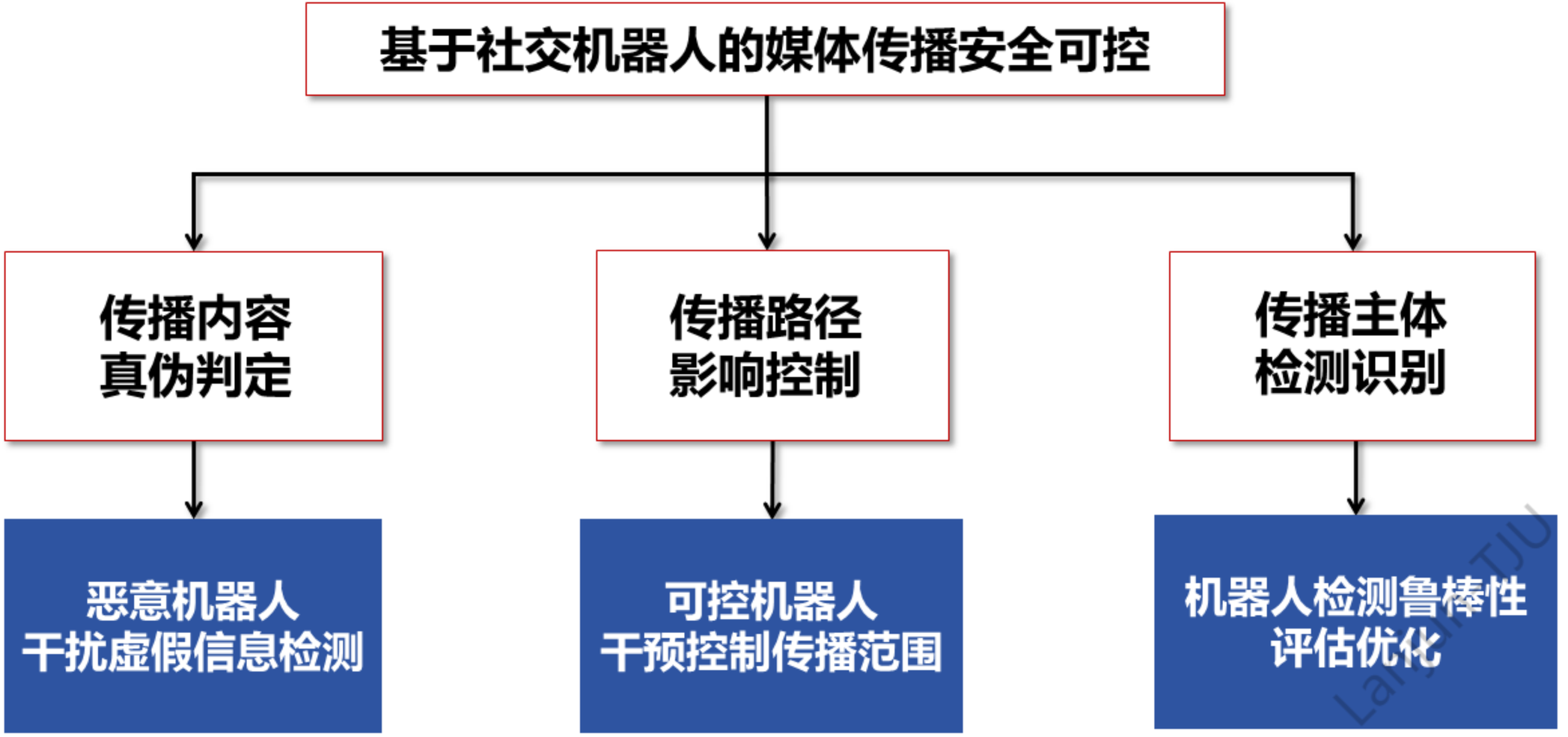
传播主体



机器人伪造用户诱导民意

亟需分析传播内容、控制传播路径、识别传播主体
以确保人机共生下的传播安全

研究背景及意义



报告内容

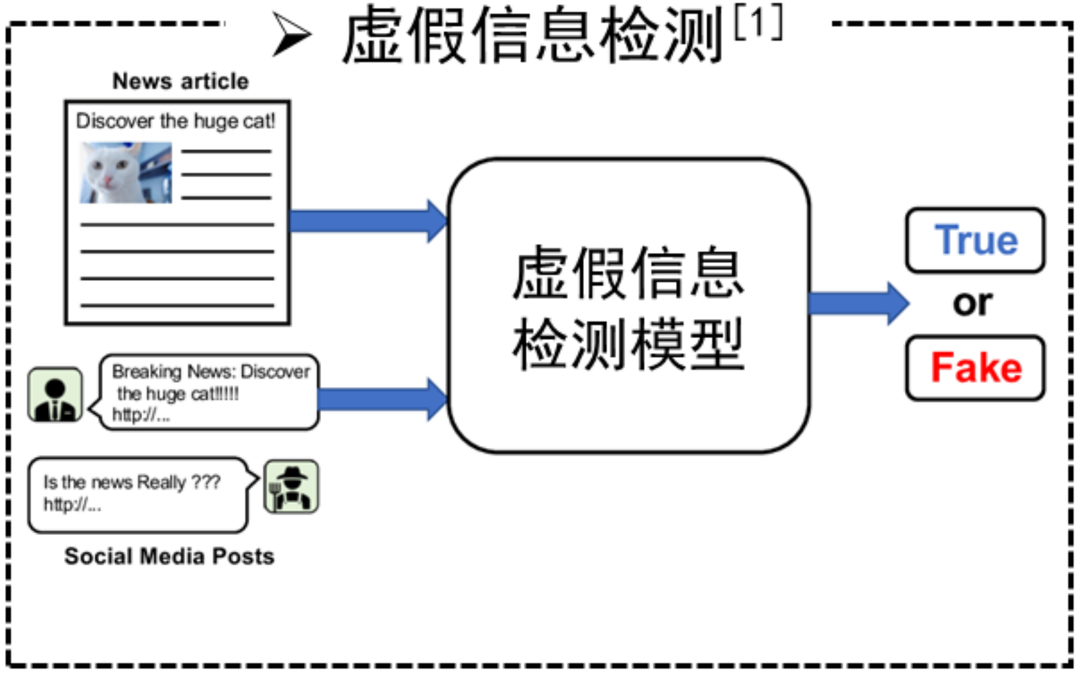
- 研究背景及意义
- 恶意机器人干扰虚假信息检测
- 可控机器人干预控制传播范围
- 机器人检测及鲁棒性评估优化
- 总结与展望

恶意机器人干扰虚假信息检测

■ 研究背景

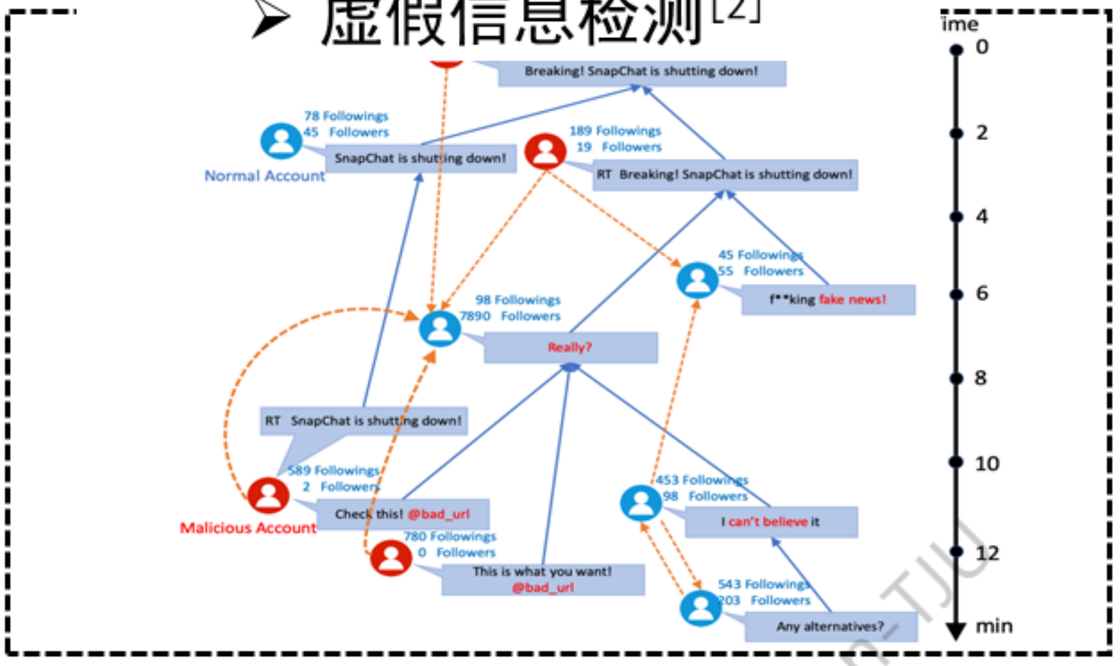
- 信息真实性检测：文本、图象.....
- 挑战：主题/内容动态变化

➢ 虚假信息检测[1]



- 信息真实性检测：传播过程中的信息与用户的交互结构

➢ 虚假信息检测[2]



[1] Analysis and Detection of Political Fake News Using Deep Learning with High-Performance Hybrid Model. ICIS 2023.
[2] Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media. WWW 2022.

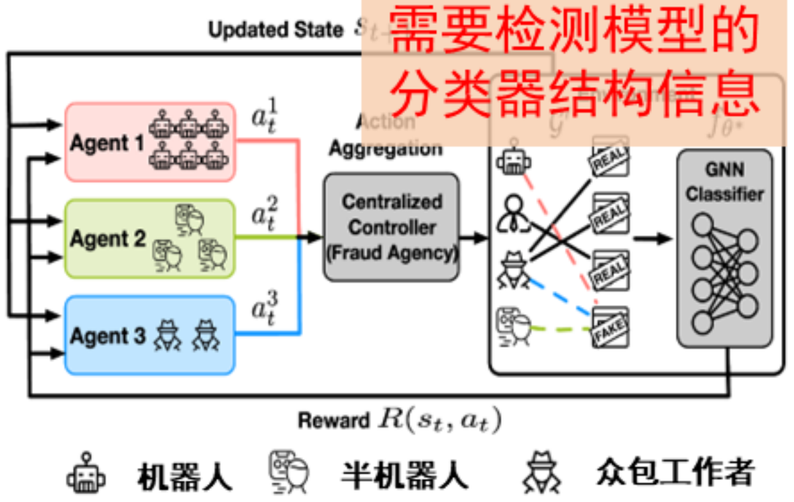
利用交互结构进行虚假信息检测存在脆弱性

恶意机器人干扰虚假信息检测

■ 研究现状

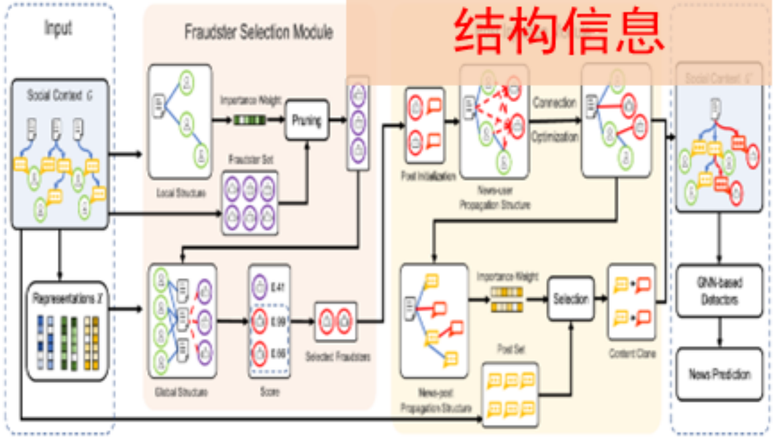
□ 灰盒方法[1]

需要检测模型的分类器结构信息



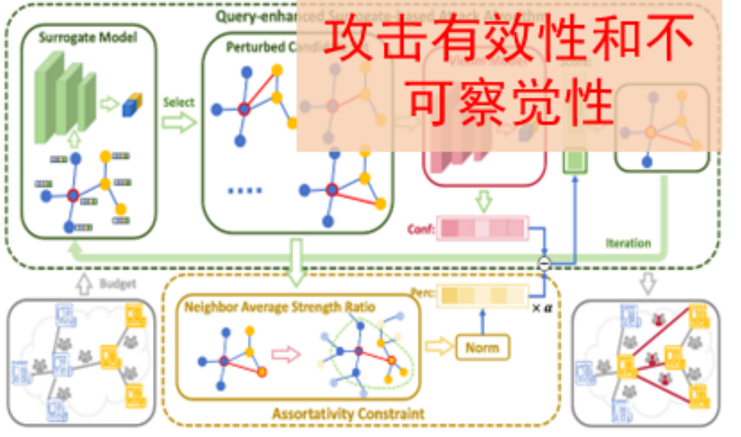
□ 黑盒方法[2]

不需要检测模型结构信息



□ 黑盒方法[3]

均衡黑盒方法的攻击有效性和不可察觉性



[1] Attacking Fake News Detectors via Manipulating News Social Engagement. WWW 2023.
[2] A General Black-box Adversarial Attack on Graph-based Fake News Detectors. IJCAI 2024.
[3] Bots Shield Fake News: Adversarial Attack on User Engagement based Fake News Detection. CIKM 2024.

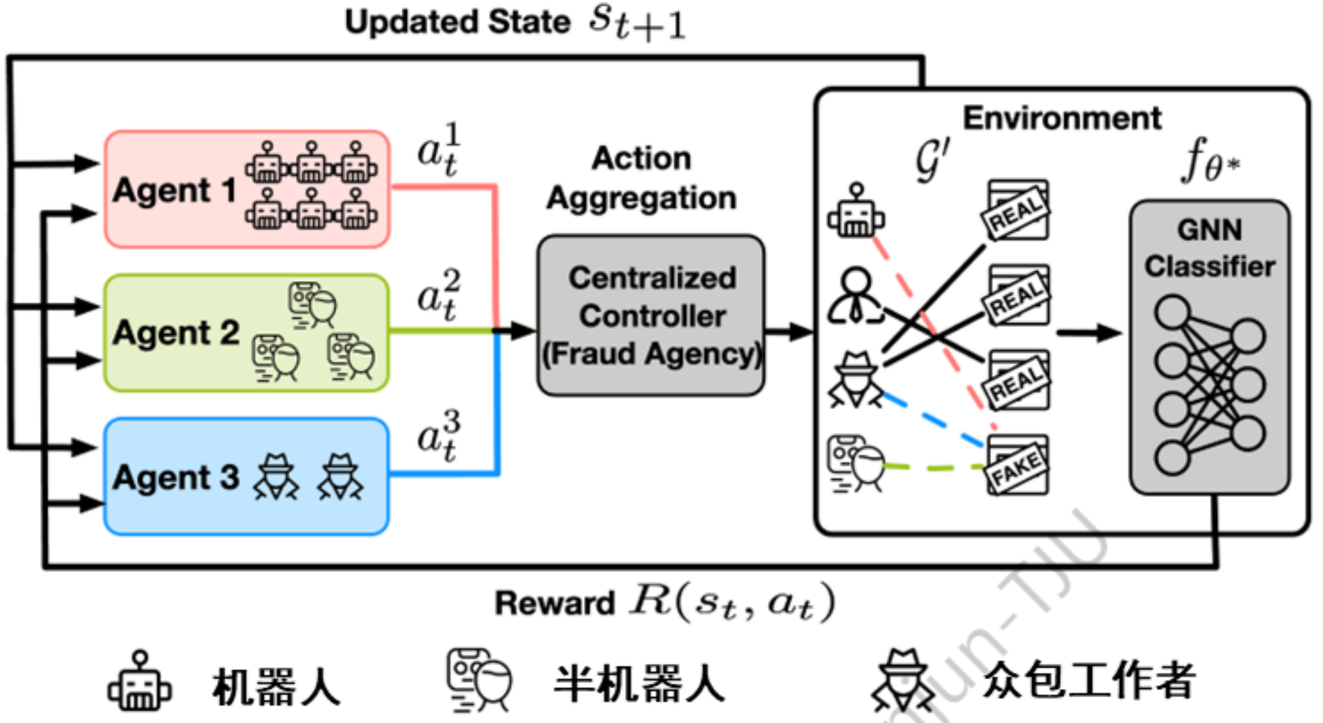
降低对检测模型结构信息的需求和可察觉性，使干扰易在现实场景中实现

恶意机器人干扰虚假信息检测

- 基于灰盒的干扰方法MARL
  - 构建代理模型：利用检测模型结构信息构建代理模型
  - 搜索干扰策略：使用深度Q-learning搜索多种类型账号（包含机器人）之间的最佳组合以提升干扰效果

Agent	User	Cost	Influence	Budget
1	bot	low	low	high
2	cyborg	moderate	moderate	moderate
3	crowd worker	high	high	low

不同类型的账号具有不同的
培养成本、影响力和扰动代价



[1] Attacking Fake News Detectors via Manipulating News Social Engagement. WWW 2023.

恶意机器人干扰虚假信息检测

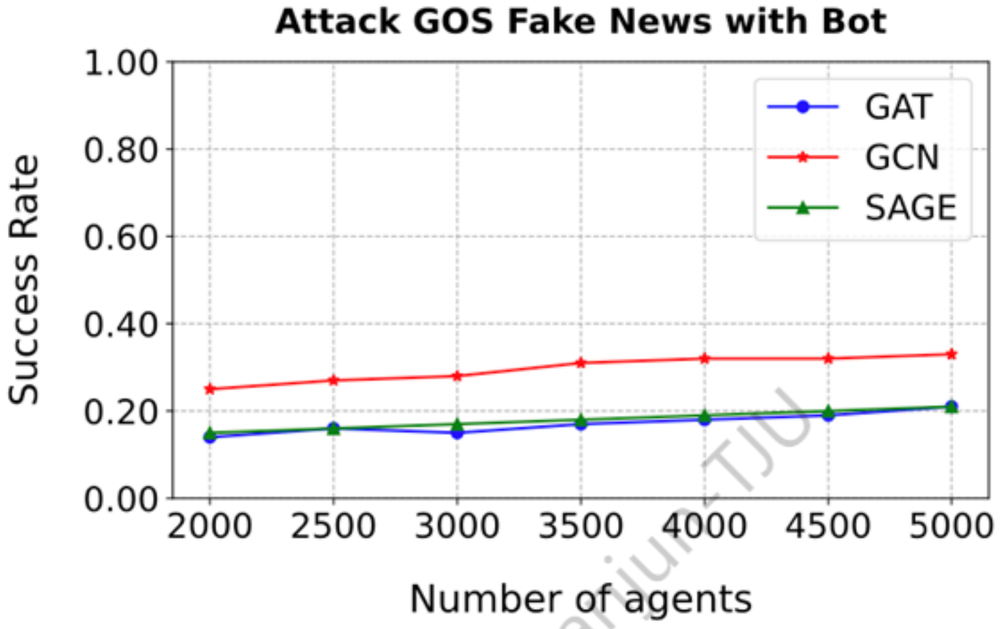
- 基于灰盒的干扰方法MARL
  - 实验结果：
    - 利用机器人账号和其他类型账号的联合干扰效果要好于只使用机器人
    - 增加机器人账号的数量可以提升干扰效果

Table 4: Results of using MARL to perform *indirect* targeted attacks comparing to several baselines. Experiments are repeated five times, and the average success rate is reported.

Method	Politifact						Gossipcop					
	Fake			Real			Fake			Real		
	GAT	GCN	SAGE	GAT	GCN	SAGE	GAT	GCN	SAGE	GAT	GCN	SAGE
RD-Edge	0.14	0.45	0.13	0.11	0.33	0.15	0.06	0.28	0.25	0.08	0.22	0.14
RD-Node	0.12	0.48	0.14	0.13	**0.38**	0.15	0.12	0.32	0.22	0.12	0.23	0.16
RL - A1	0.17	0.42	0.16	0.08	0.07	**0.21**	0.14	0.45	0.23	0.08	0.80	0.16
RL - A2	0.15	0.38	0.16	0.08	0.13	0.18	0.18	0.52	0.32	0.06	0.83	0.24
RL - A3	0.18	0.64	0.19	0.08	0.13	0.18	0.19	0.51	0.31	0.12	0.85	0.22
MARL	**0.33**	**0.92**	**0.28**	**0.22**	0.31	0.19	**0.21**	**0.64**	**0.36**	**0.18**	**0.89**	**0.28**

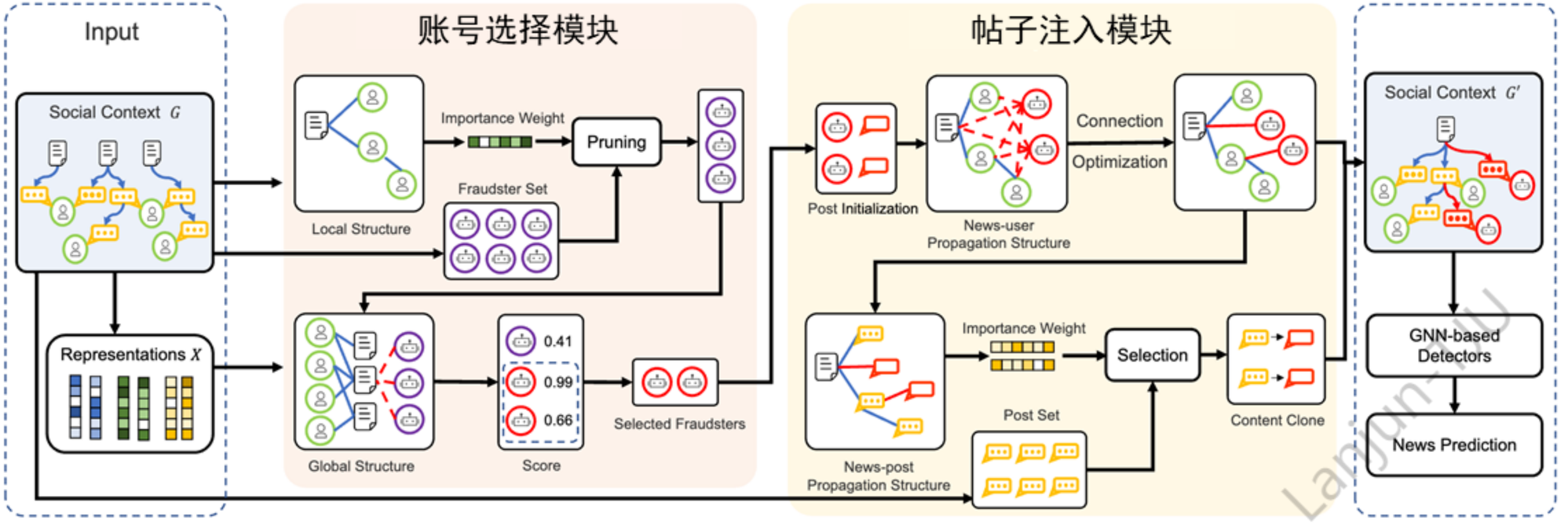
平均提升54.1%

平均提升140.87%



恶意机器人干扰虚假信息检测

- 基于黑盒的干扰方法GAFSI
  - 优化MARL需要检测模型结构信息的问题
  - 基本思想：
    - 账号选择模块：结合局部和全局结构信息选择具有最高影响力的账号
    - 帖子注入模块：控制被选择账号发送新帖子进行干扰

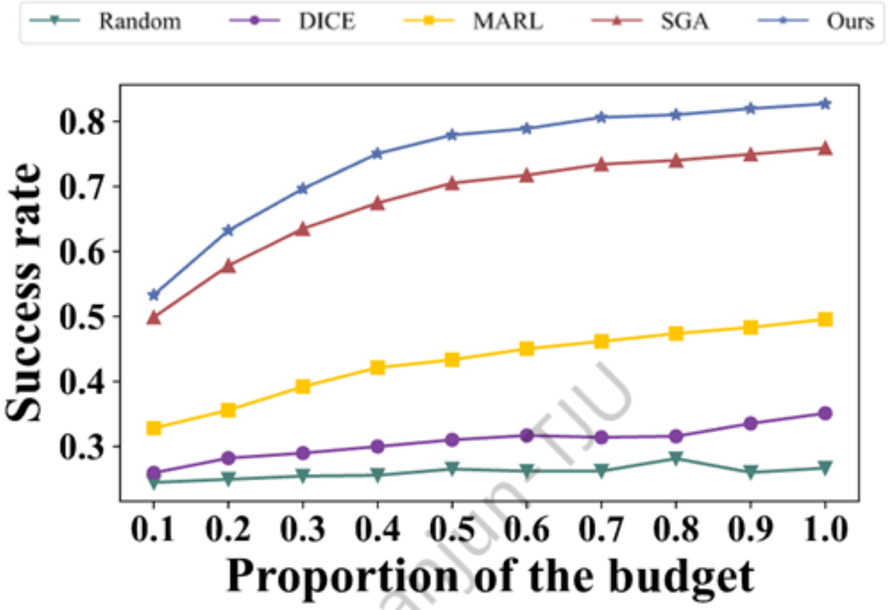


[1] A General Black-box Adversarial Attack on Graph-based Fake News Detectors. IJCAI 2024.

恶意机器人干扰虚假信息检测

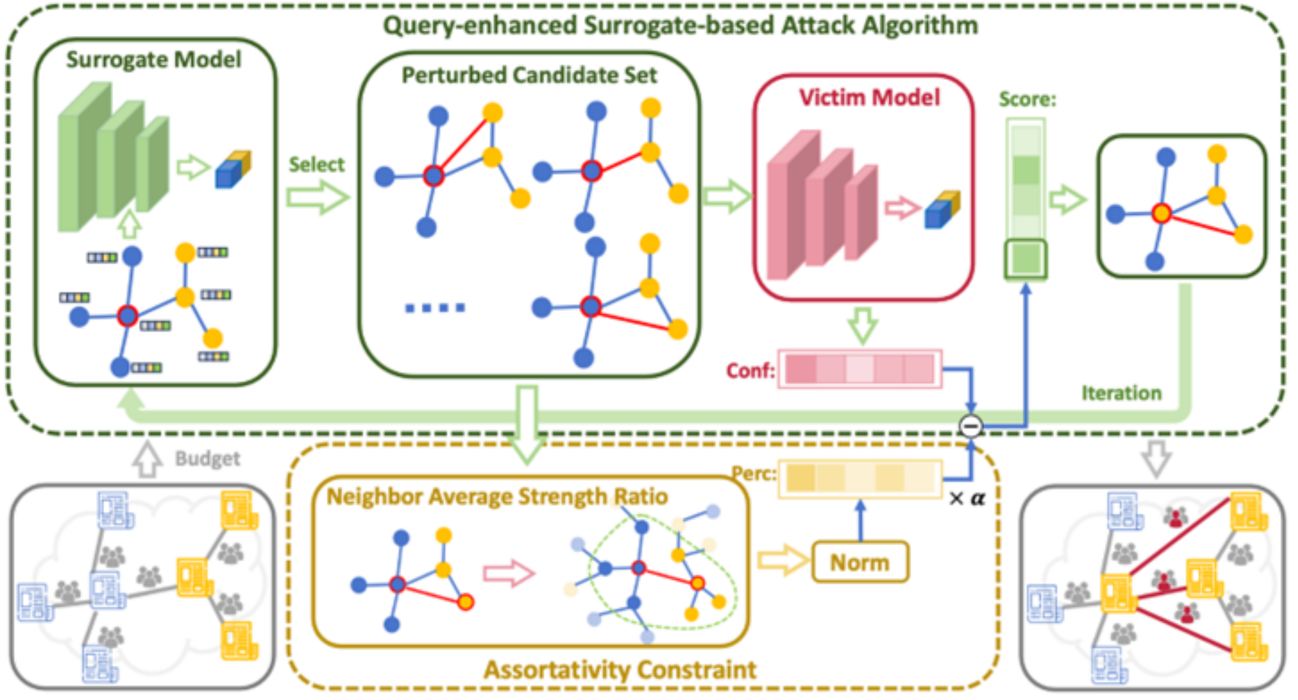
- 基于黑盒的干扰方法GAFSI
  - 实验结果：
    - 对于不同结构的检测模型，GAFSI在大多数情况下具有更好的干扰效果
    - 随着被选择账号数量的增加，GAFSI在干扰效果上具有更明显的提升

Dataset	Graph	Model	Fake News						Real News					
			-	Random	DICE	MARL	SGA	GAFSI	-	Random	DICE	MARL	SGA	GAFSI
Politifact	G1	GCN	0.20	0.16	0.29	0.97	**1.00**	**1.00**	0.04	0.18	0.30	0.81	**0.95**	**0.95**
		SAGE	0.27	0.30	0.37	0.56	0.84	**0.85**	0.12	0.12	0.15	0.32	0.65	**0.70**
		GAT	0.21	0.24	0.34	0.43	0.68	**0.70**	0.10	0.12	0.18	0.39	**0.67**	0.55
	G2	DECOR-GCN	0.18	0.31	0.45	0.57	0.86	**0.92**	0.10	0.11	0.21	0.36	0.32	**0.37**
	G3	UPFD-GCN	0.26	0.29	0.31	0.39	0.55	**0.98**	0.10	0.11	0.11	0.13	0.15	**0.72**
		UPFD-SAGE	0.26	0.35	0.52	0.56	0.89	**0.96**	0.13	0.16	0.26	0.32	0.73	**0.86**
		UPFD-GAT	0.27	0.26	0.28	0.31	0.45	**0.48**	0.14	0.16	0.17	0.25	0.41	**0.52**
	G4	BiGCN	0.21	0.22	0.23	0.29	0.31	**0.91**	0.10	0.12	0.15	0.13	0.18	**0.92**
Gossipcop	G1	GCN	0.02	0.01	0.04	0.63	**0.77**	0.57	0.08	0.15	0.18	0.75	**1.00**	**1.00**
		SAGE	0.10	0.09	0.10	0.31	0.37	**0.92**	0.02	0.12	0.13	0.35	0.85	**0.91**
		GAT	0.04	0.05	0.08	0.21	**0.40**	0.23	0.03	0.08	0.11	0.22	**0.48**	**0.48**
	G2	DECOR-GCN	0.07	0.06	0.09	0.23	0.64	**0.78**	0.04	0.13	0.15	0.08	0.20	**0.24**
	G3	UPFD-GCN	0.02	0.02	0.04	0.05	0.06	**0.85**	0.05	0.06	0.08	0.08	0.12	**0.84**
		UPFD-SAGE	0.02	0.03	0.06	0.04	0.21	**0.68**	0.05	0.05	0.07	0.18	0.52	**0.68**
		UPFD-GAT	0.02	0.03	0.06	0.04	0.20	**0.75**	0.06	0.06	0.08	0.16	0.44	**0.64**
	G4	BiGCN	0.04	0.05	0.07	0.05	0.08	**0.63**	0.05	0.07	0.08	0.09	0.10	**0.78**



恶意机器人干扰虚假信息检测

- 基于不可察觉性均衡的干扰方法QSA-AC
  - 优化了GAFSI等方法干扰容易被察觉的问题，均衡干扰成功率和不可察觉性的
  - 基本思想：
    - 创建机器人：基于代理模型和检测模型信息创建具有相应交互行为的机器人
    - 不可察觉性约束：限制干扰对社交网络中信息结构属性的扰动幅度



[1] Bots Shield Fake News: Adversarial Attack on User Engagement based Fake News Detection, CIKM 2024.

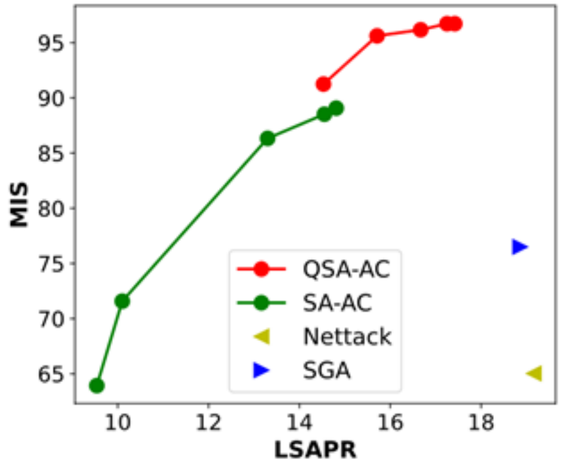
恶意机器人干扰虚假信息检测

■ 基于不可察觉性均衡的干扰方法QSA-AC

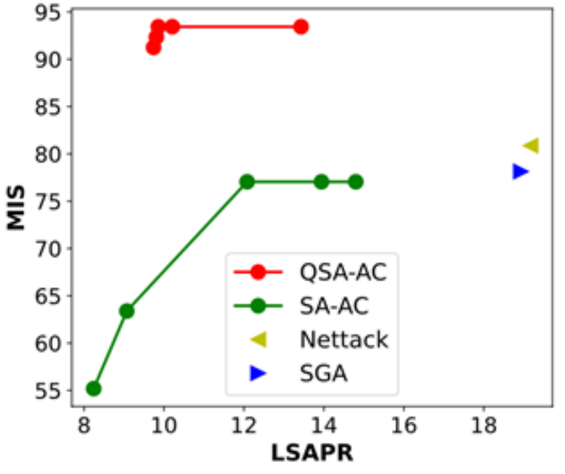
□实验结果：

□对于不同分类器结构的检测模型，QSA-AC在大多数情况下具有更好的干扰效果

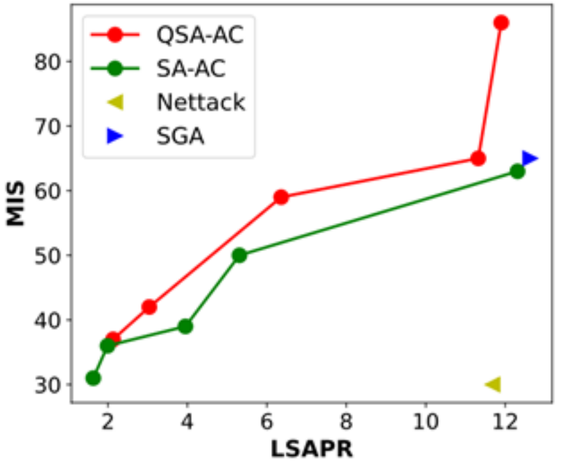
□通过调整参数，QSA-AC可以在干扰的有效性和不可察觉性之间达到一个均衡



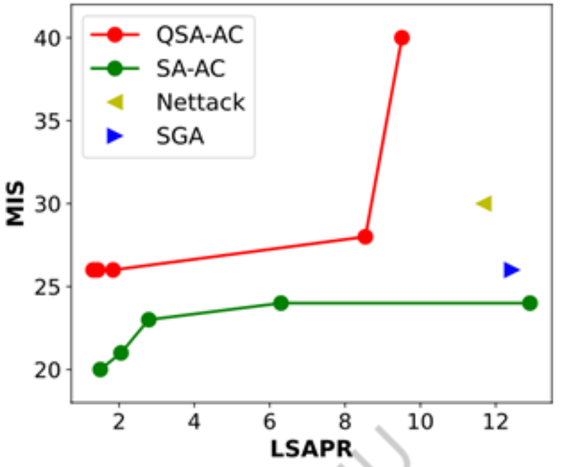
(a) Politifact GCN



(b) Politifact GAT



(e) Gossipcop1 GCN



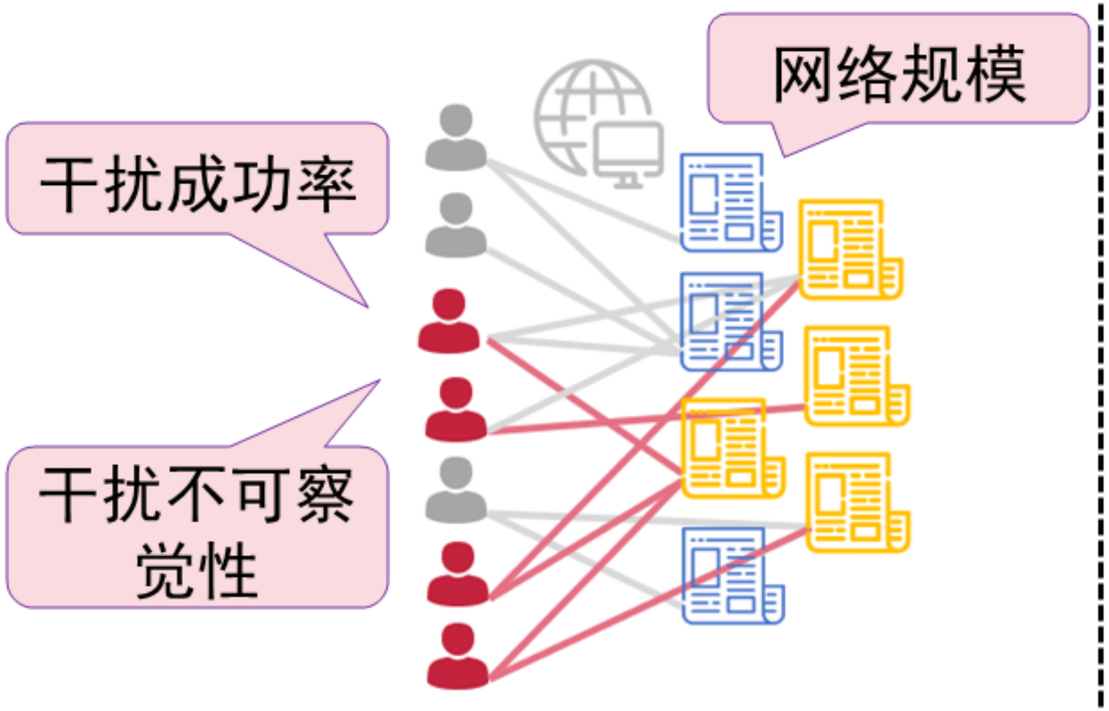
(f) Gossipcop1 GAT

- **MIS**：信息错误分类率，分值越高代表干扰成功率越高。
- **LSAPR**：局部同配性扰动比例，分值越高代表干扰越容易被察觉。

小结

■ 恶意机器人干扰虚假信息检测

➢ 领域现状



- 检测模型的结构是影响干扰成功率的重要因素
- 不可察觉性是衡量干扰的重要标准
- 网络规模的增加会使得干扰成功率下降

报告内容

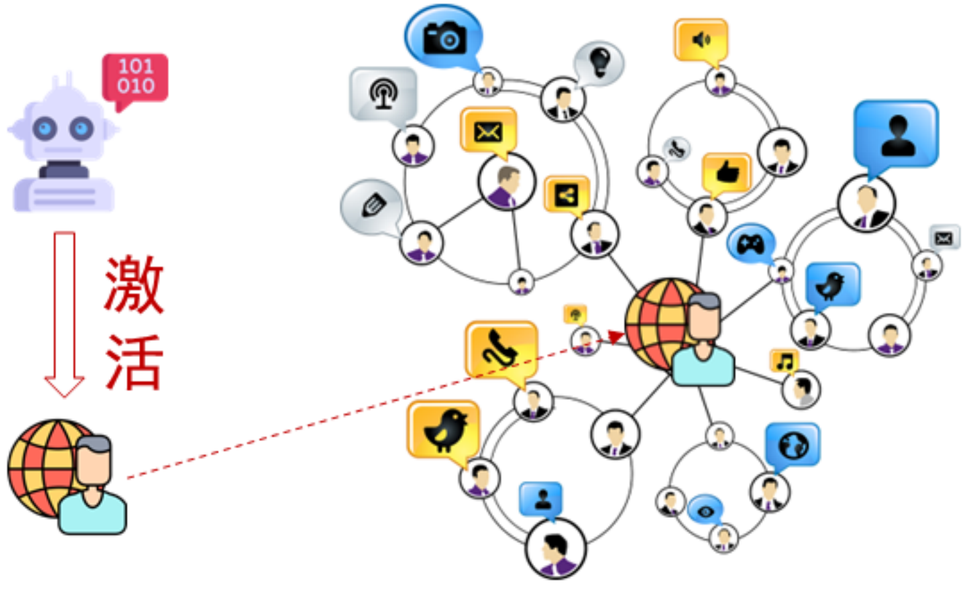
- 研究背景及意义
- 恶意机器人干扰虚假信息检测
- 可控机器人干预控制传播范围
- 机器人检测及鲁棒性评估优化
- 总结与展望

可控机器人干预控制传播范围

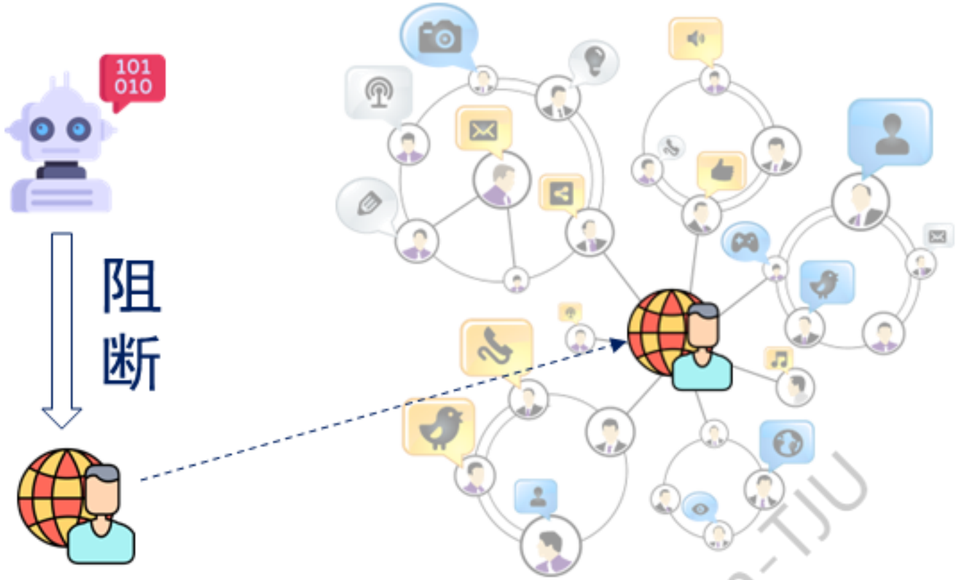
■ 研究背景

□社交网络上混杂信息传播，亟需不同的控制策略

➢ 正面信息传播范围最大化



➢ 负面信息传播范围最小化



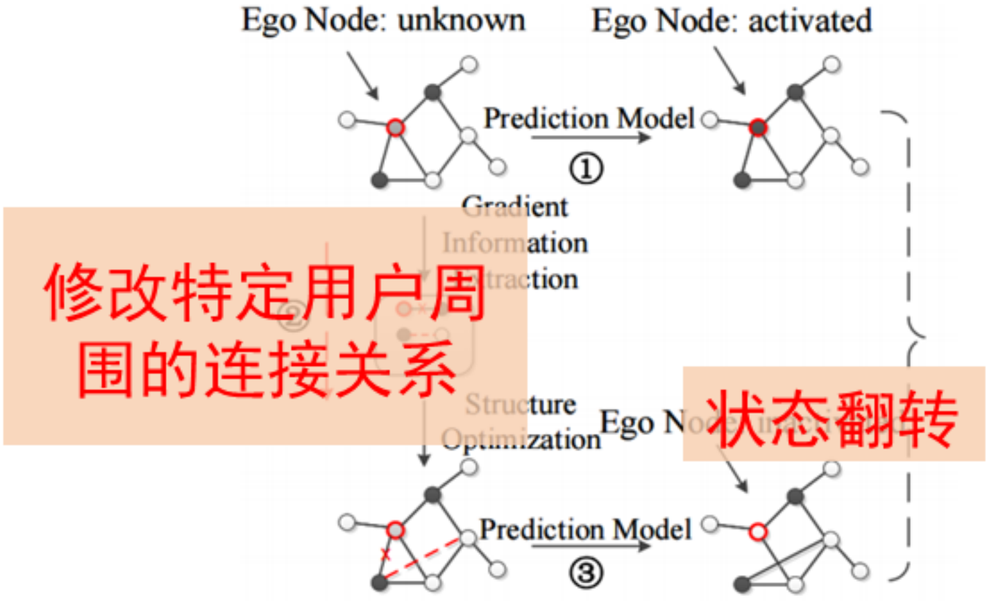
可控机器人通过改变特定用户的传播状态控制不同信息的传播范围

可控机器人干预控制传播范围

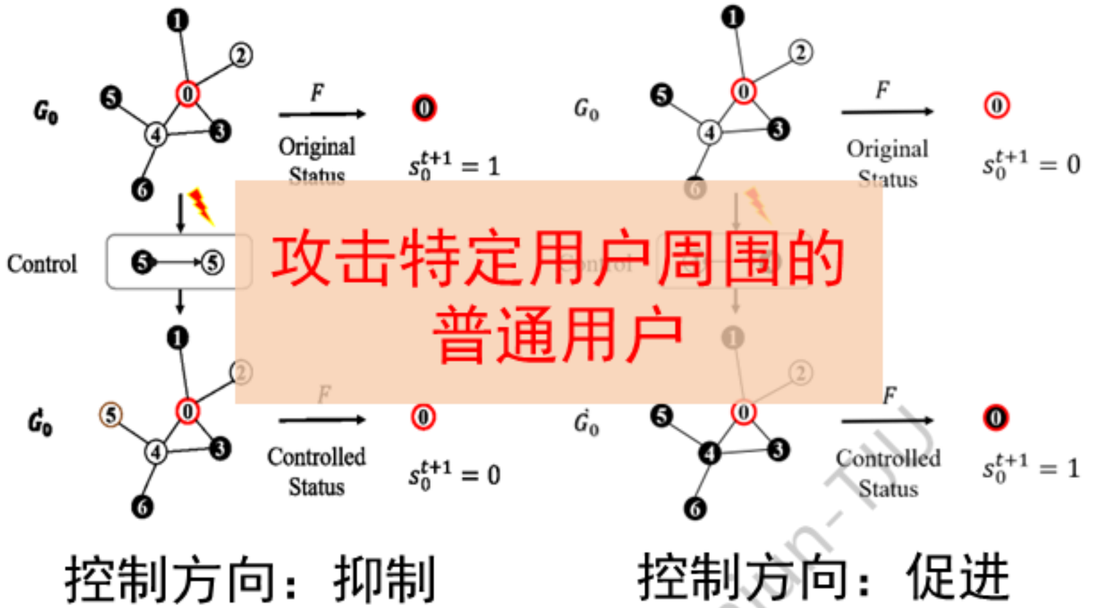
■ 研究挑战

- □直接控制关键节点（如，网红）不可实现
- □控制要有指向性：促进or抑制

➢ 基于边的传播状态翻转[1]



➢ 基于点的传播定向控制



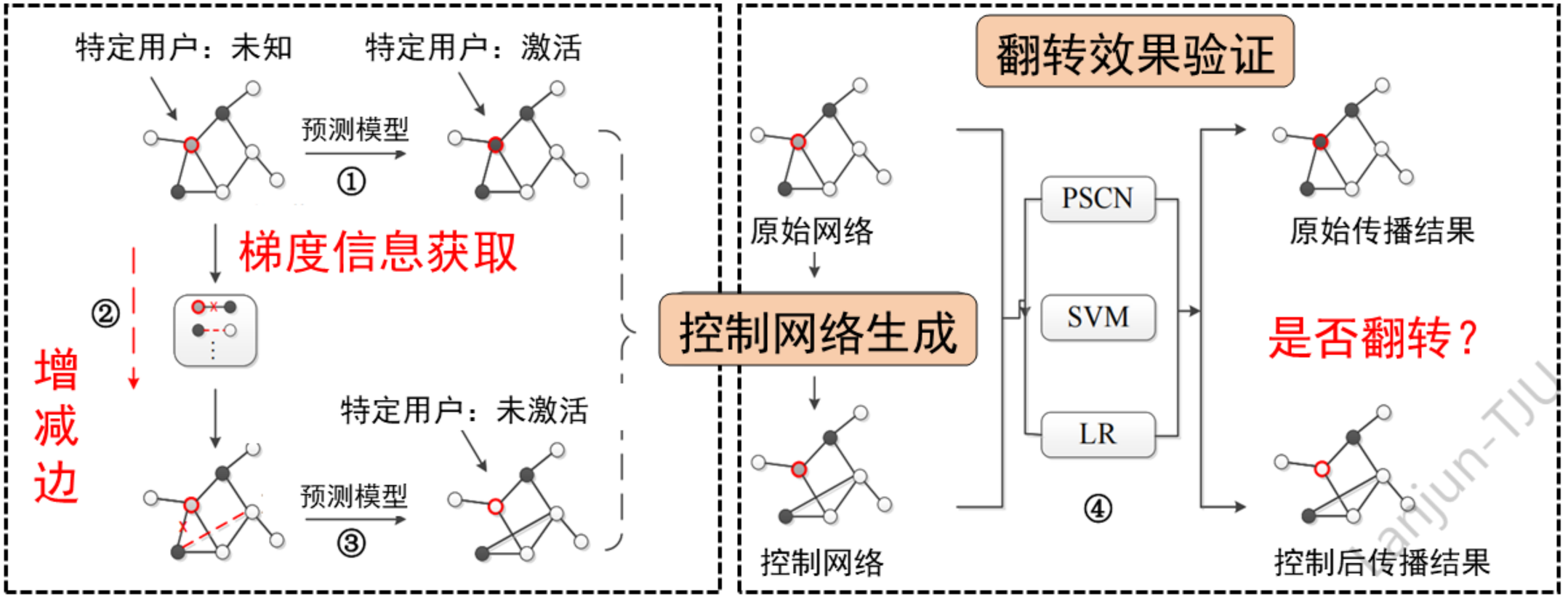
[1] Ctl-diff: Control information diffusion in social network by structure optimization. TCSS 2022.

可控机器人干预控制传播范围

■ 基于边的传播状态翻转

□控制网络生成：基于预测模型获取不同边的梯度信息，从而对网络结构进行调整

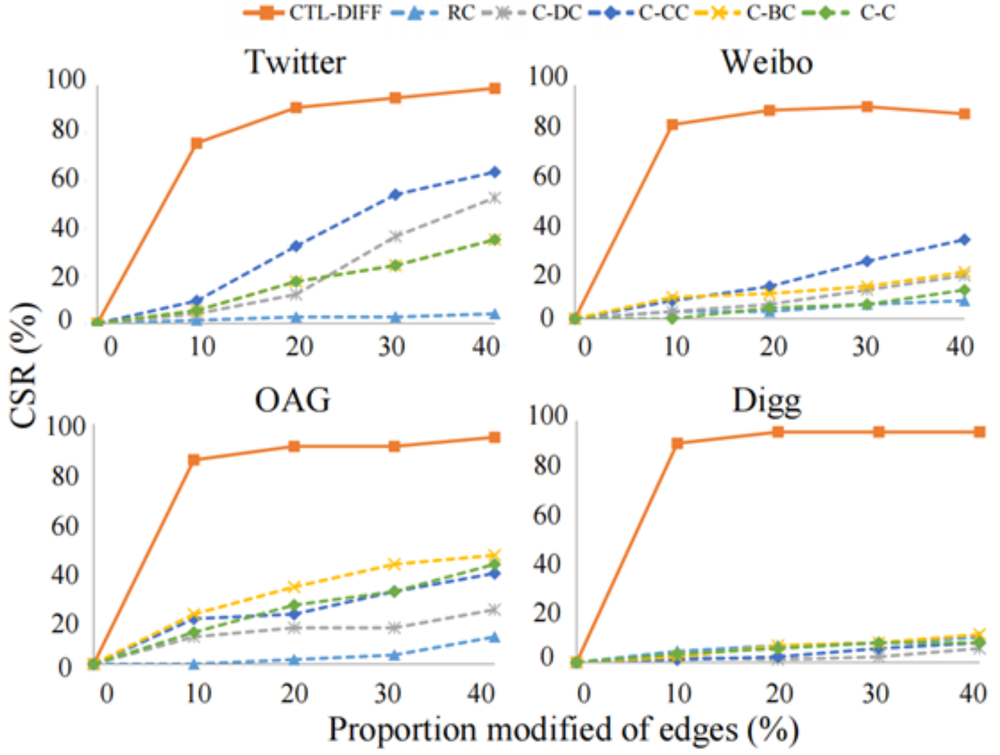
□翻转效果验证：在不同预测模型下对控制网络进行翻转效果的评估



[1] Ctl-diff: Control information diffusion in social network by structure optimization. TCSS 2022.

可控机器人干预控制传播范围

- 基于边的传播状态翻转
  - 在四个公开数据集上实现传播状态的翻转，在具有一定的迁移能力
  - 要达到较好的控制效果，需要修改40%的边关系



Datasets	Verification Methods	CSR(%)			
		10%	20%	30%	40%
Twitter	LR	65.00	82.50	92.50	95.00
	SVM	76.54	91.36	98.77	98.77
	PSCN	38.67	38.67	38.67	38.67
Weibo	LR	98.36	98.36	98.36	98.36
	SVM	98.31	98.31	98.31	98.31
	PSCN	45.45	45.45	45.45	45.45
OAG	LR	87.30	96.83	98.41	98.41
	SVM	87.30	98.41	98.41	98.41
	PSCN	56.36	56.36	56.36	56.36
Digg	LR	66.67	72.62	64.29	65.48
	SVM	65.88	72.94	64.70	65.88
	PSCN	47.50	52.50	52.50	50.00

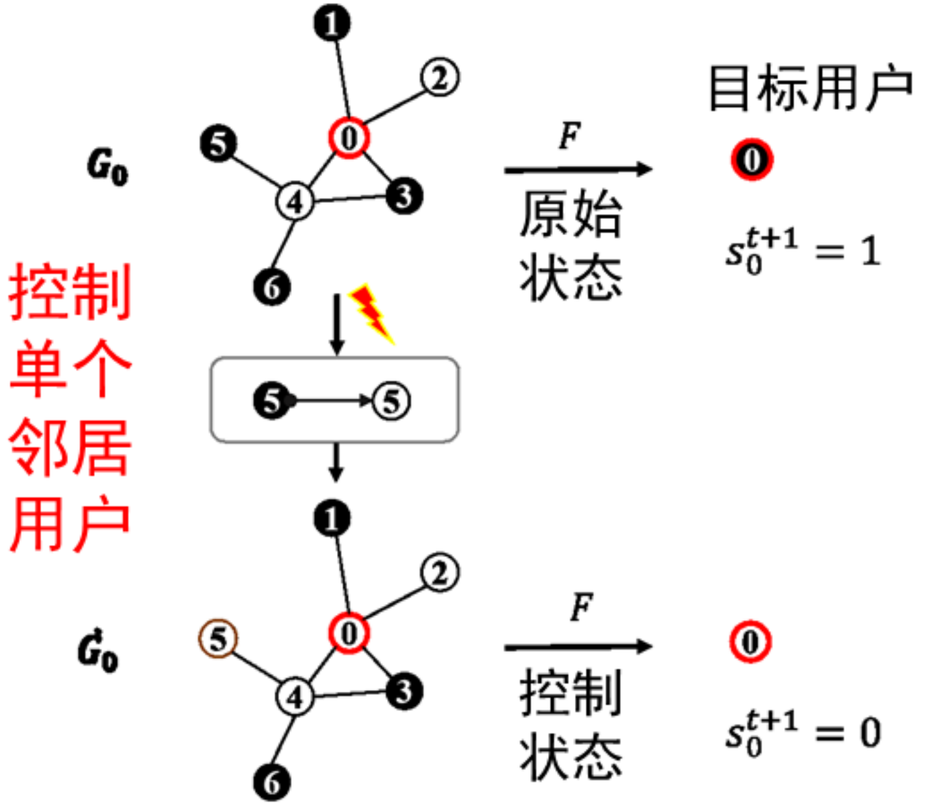
[1] Ctl-diff: Control information diffusion in social network by structure optimization. TCSS 2022.

可控机器人干预控制传播范围

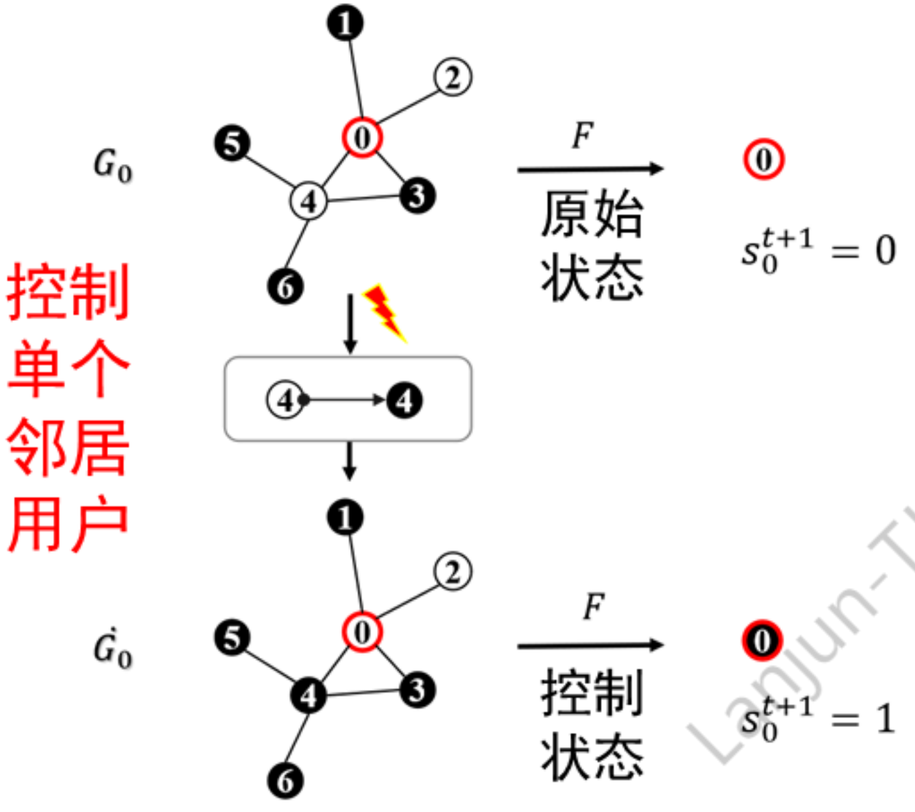
■ 基于点的传播定向控制

□确定控制方向：依据信息的性质，确定控制的方向，而非盲目翻转

□约束控制节点个数：确保不可察觉性 ->单点



控制方向：抑制



控制方向：促进

可控机器人干预控制传播范围

- 基于点的传播定向控制
  - 卧底邻居控制
    - 查询：通过查询控制成功分数找到特定用户周围的最佳卧底邻居用户
    - 控制：通过控制最佳卧底邻居间接控制特定用户
  - 控制效果验证
    - 定向控制成功率计算

$$CSR_{inhibtion} = \frac{N_{inhibtion\_success}}{N_{inhibtion\_total}}$$

$$CSR_{promotion} = \frac{N_{promotion\_success}}{N_{promotion\_total}}$$

Algorithm 1 SNC for inhibition control

Input: The maximum number of iterations T; The maximum number of control candidates N; Diffusion prediction model F.

Output: The number of target users N_{total}; The number of success-controlled target users $N_{success}$.

```
1:  N_total ← 0, N_success ← 0
2:  for n = 1 → N do
3:      Initialize ω
        查询
4:      if F(v|G_v, S_v^t) = 1 then
5:          Count the target users N_total ← N_total + 1
6:          for t = 1 → T do                    ▷ The query stage
7:              Sample an neighbor u via Eq. (2)
8:              Undercover the neighbor by s_u = 0
9:              Query the control score l_u via Eq. (3)
10:             Update the neighbor weight ω̂_u ← l_u
11:         end for
12:     end if
13: end for
        控制
14: for n = 1 → N_total do
15:     Obtain the best neighbor û via Eq. (4)  ▷ The control stage
16:     Undercover the neighbor by s_û = 0
17:     if F(v|G_v, Ŝ_v^t) = 0 then             ▷ If control success
18:         Count the success candidates N_success ← N_success + 1
19:     end if
20: end for
```

可控机器人干预控制传播范围

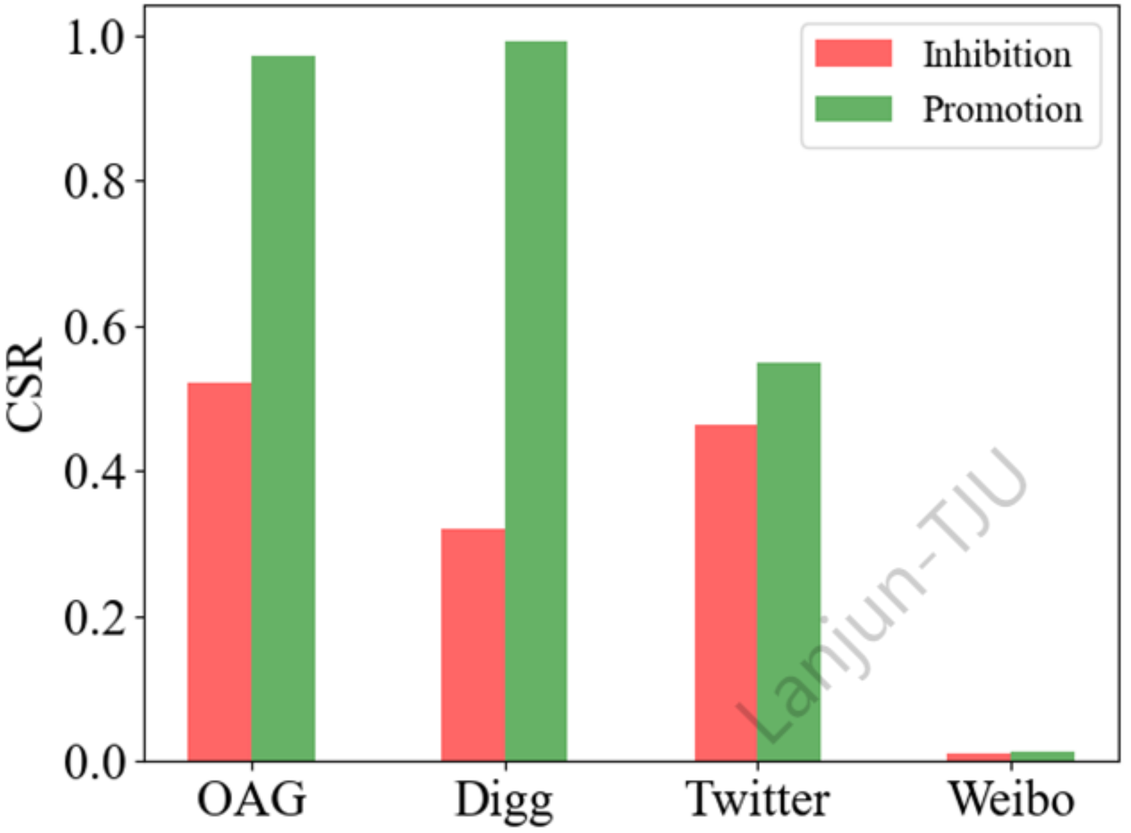
- 基于点的传播定向控制
  - 在3个公开数据集上能够实现更好的控制效果
  - 促进比抑制的控制成功率更高

Table 6: Overall performance in the inhibition setting.

Datasets	OAG	Digg	Twitter	Weibo
CTL-DIFF [5]	0.1253	0.0672	0.1238	**0.1588**
Random	0.0791	0.0502	0.0263	0.0020
SNC(ours)	**0.5222**	**0.3200**	**0.4625**	0.0106

Table 7: Overall performance in the promotion setting.

Datasets	OAG	Digg	Twitter	Weibo
CTL-DIFF [5]	0.1176	0.3477	0.0732	**0.0601**
Random	0.0979	0.0123	0.0187	0.0011
SNC(ours)	**0.9705**	**0.9925**	**0.5490**	0.0144



可控机器人干预控制传播范围

- 基于点的传播定向控制
  - 网络特征：具有连接不那么紧密、节点间距离较远的ego-network更难以控制。
  - 目标用户：邻居较少且影响力较低的目标用户更容易控制。

Table 1: Network characteristics of the public datasets. *#Samples* denotes the number of ego networks.

Datasets	OAG	Digg	Twitter	Weibo
# Samples	257,273	20,361	228,798	528,571
Ave. Edges	152	229	152	167
Ave. Clustering	0.5657	0.4202	0.3240	**0.2803**
Ave. Diameter	5.1178	4.6311	5.2391	**5.5059**
Ave. Degree assortativity	-0.2683	-0.3067	-0.2203	-0.2502

Table 2: Average clustering and diameter of the success controlled v.s. failed controlled ego networks on Weibo.

Control Result	Ave. Clustering	Ave. Diameter
Failed	0.2784	5.5083
Success	**0.4216**	**5.3297**

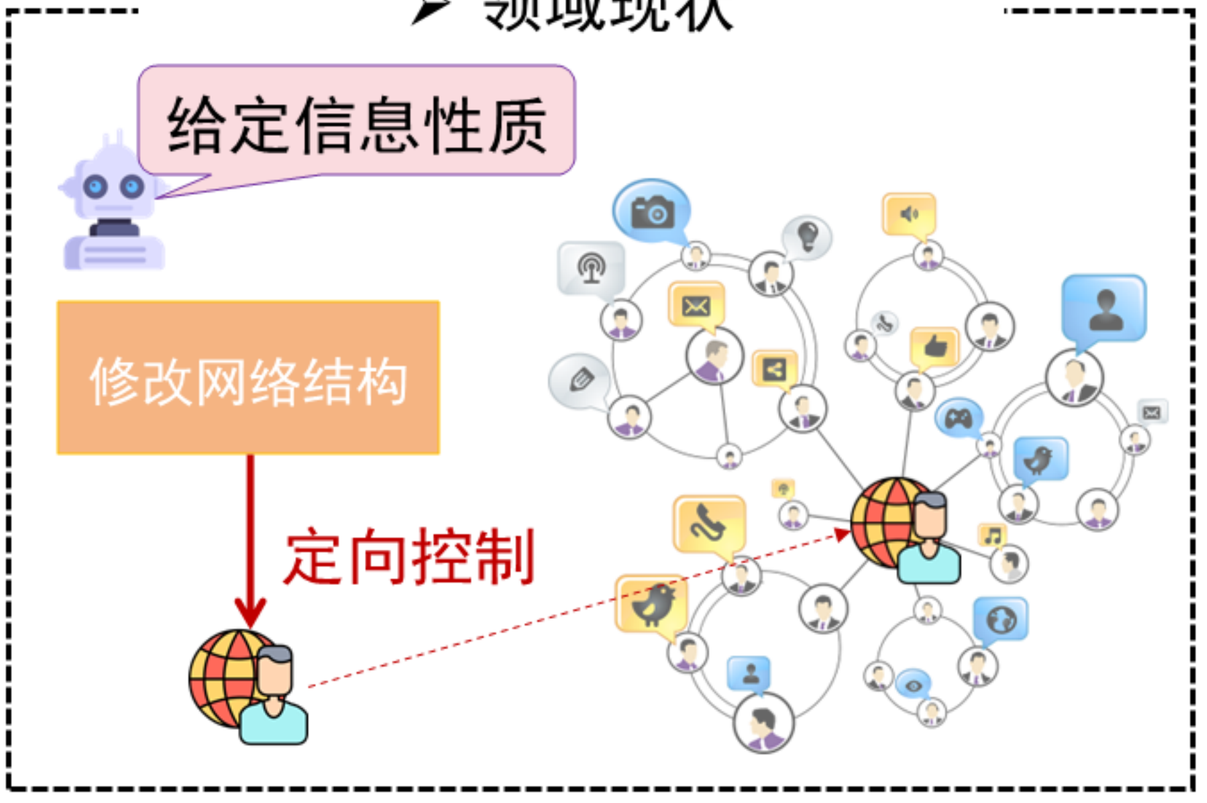
Table 4: Characteristics of target users controlled by SNC.

Datasets	Control Result	Ave. Degree	Ave. DC	Ave. CC	Ave. BC	Ave. EC
OAG	Failed	11.9435	0.2438	0.5489	0.1463	0.2094
	Success	15.4998	0.3164	0.5719	0.2147	0.2760
Digg	Failed	10.0152	0.2044	0.5312	0.1424	0.1561
	Success	**9.7929**	**0.1999**	0.5343	**0.0730**	**0.1234**
Twitter	Failed	25.5927	0.5223	0.6761	0.4053	0.3595
	Success	**18.6625**	**0.3809**	**0.5974**	**0.3113**	**0.3124**
Weibo	Failed	16.3463	0.3336	0.5711	0.1898	0.2686
	Success	18.6760	0.3811	0.5884	**0.0845**	**0.2308**

小结

■ 可控机器人干预控制传播范围

➢ 领域现状



- 基于边的传播状态翻转
  - 引入间接控制的思想，解决直接控制不可实现的问题
- 基于点的传播定向控制
  - 引入定向控制的思想，解决控制方向有指向性的问题

报告内容

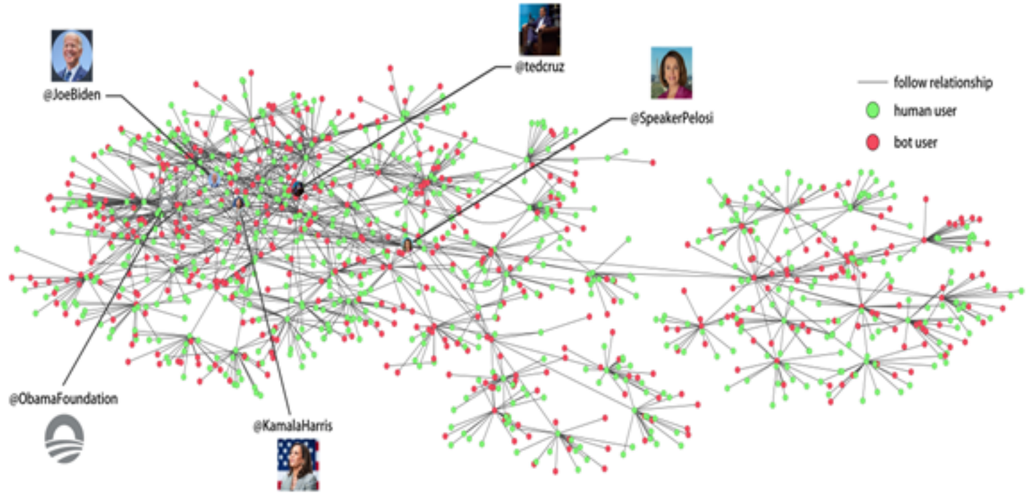
- 研究背景及意义
- 恶意机器人干扰虚假信息检测
- 可控机器人干预控制传播范围
- 机器人检测及鲁棒性评估优化
- 总结与展望

机器人检测及鲁棒性评估优化

■ 研究背景

□社交机器人能够自动产生、传播、回应内容，影响公众意见和信息流动

➢ 高度拟人，难以区分



思维认知、行为模式、语言习惯

➢ 广泛存在，潜在威胁

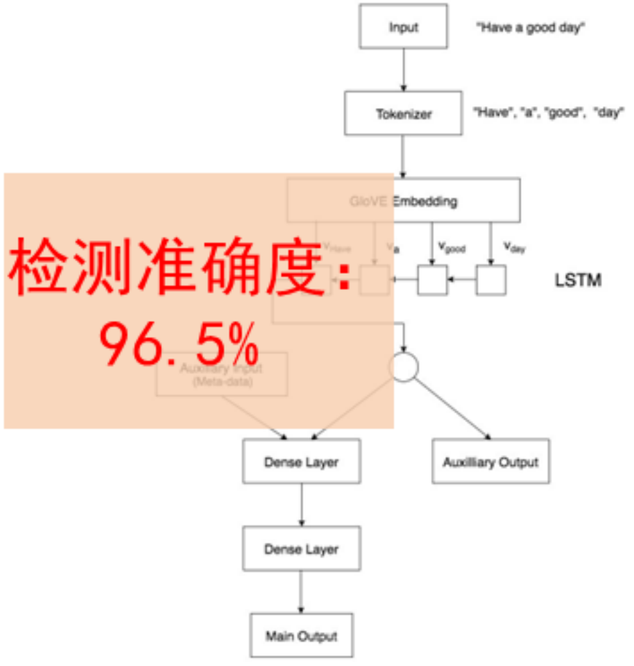
散播谣言、操纵舆论

亟需**社交机器人检测关键技术**，区分真实用户和社交机器人用户

机器人检测及鲁棒性评估优化

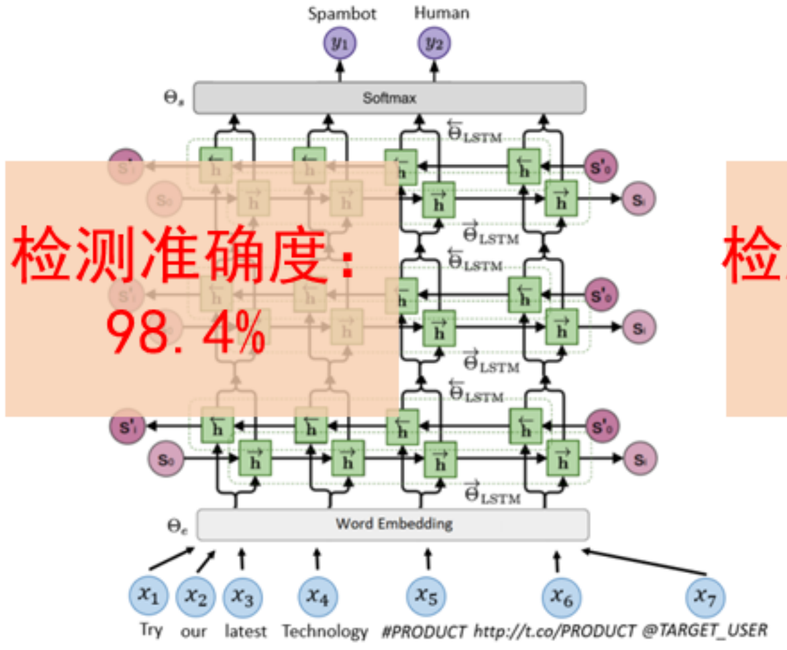
■ 研究现状

□ 基于用户属性[1]



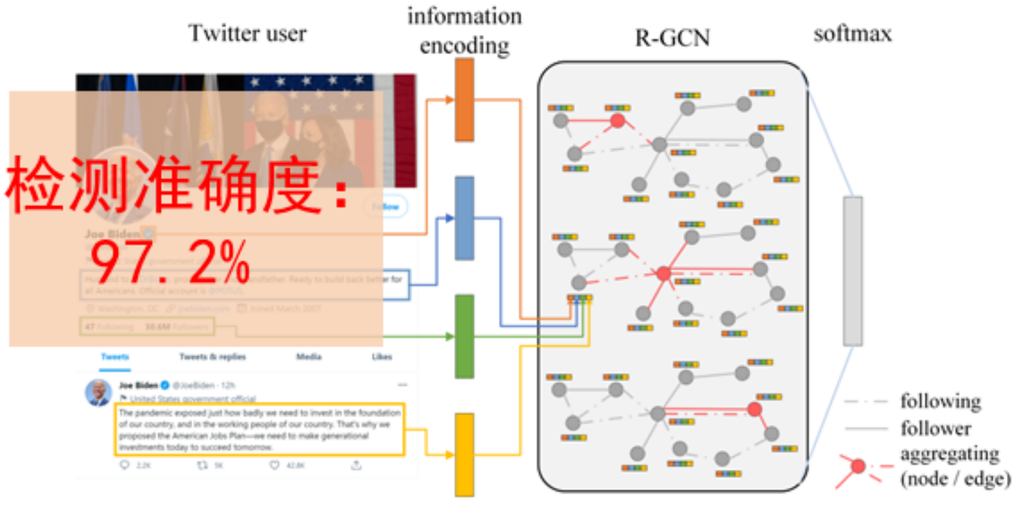
检测准确度：96.5%

□ 基于文本内容[2]



检测准确度：98.4%

□ 基于网络结构[3]



检测准确度：97.2%

[1] Deep neural networks for bot detection. Information Sciences 2018

[2] Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. TPS-ISA 2019

[3] BotRGCN: Twitter bot detection with relational graph convolutional networks. ASONAM 2021

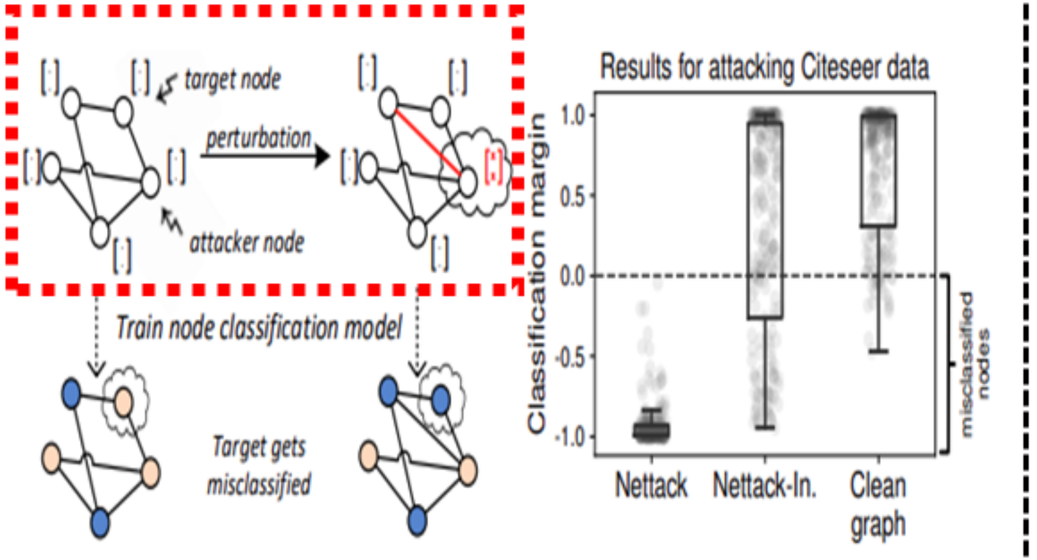
社交机器人检测算法性能提升，但社交网络中仍存在大量社交机器人

机器人检测及鲁棒性评估优化

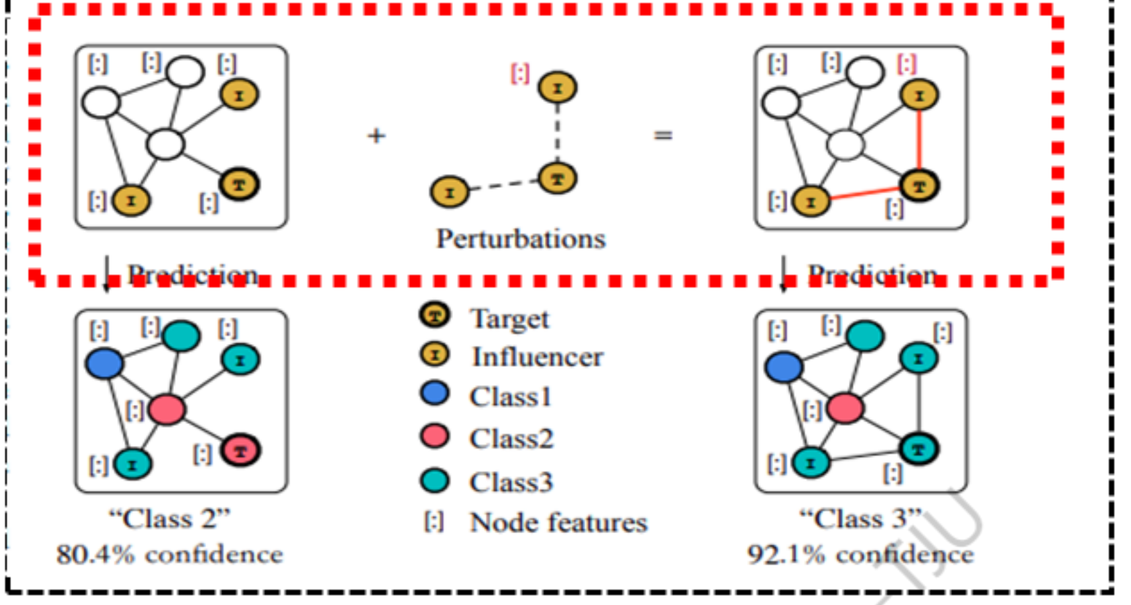
■ 现存问题

□图神经网络具有脆弱性、容易遭受攻击导致错误分类。

➢ 图神经网络脆弱[1]



➢ 易被攻击致误分类[2]



[1] Adversarial Attacks on Neural Networks for Graph Data. IJCAI 2019

[2] A Survey of Adversarial Learning on Graphs. CoRR 2020

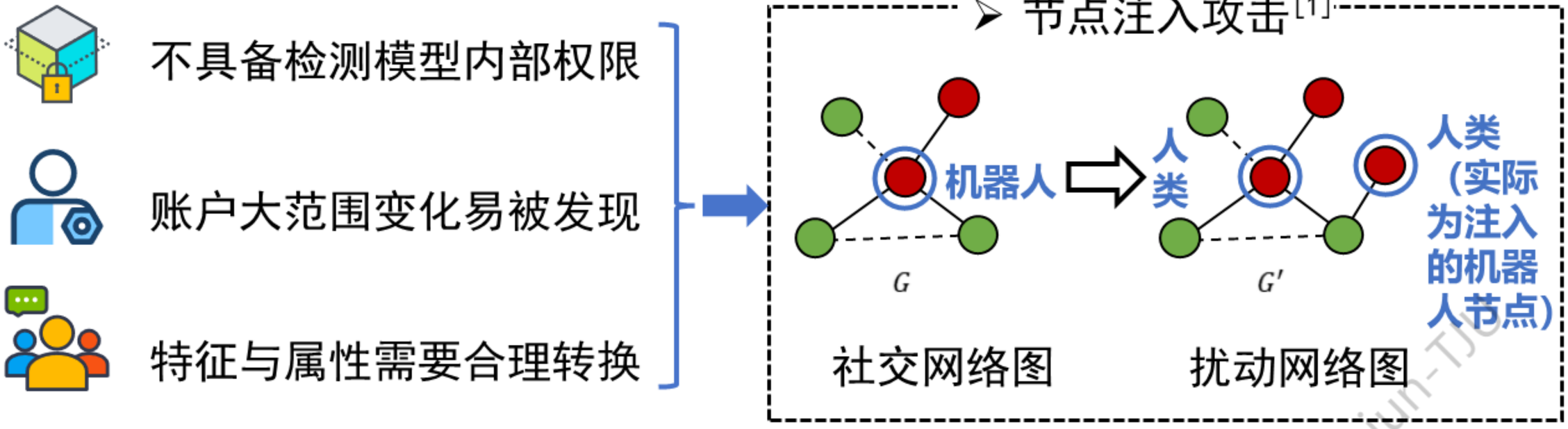
社交机器人检测算法存在短板，需要对模型进行鲁棒性评估和增强

机器人检测及鲁棒性评估优化

■ 研究挑战

□评估：**对抗性攻击 -> 检测模型漏洞**

□挑战：①攻击方法限制为黑盒；②扰动确保不可察觉；③节点需具备社交属性；

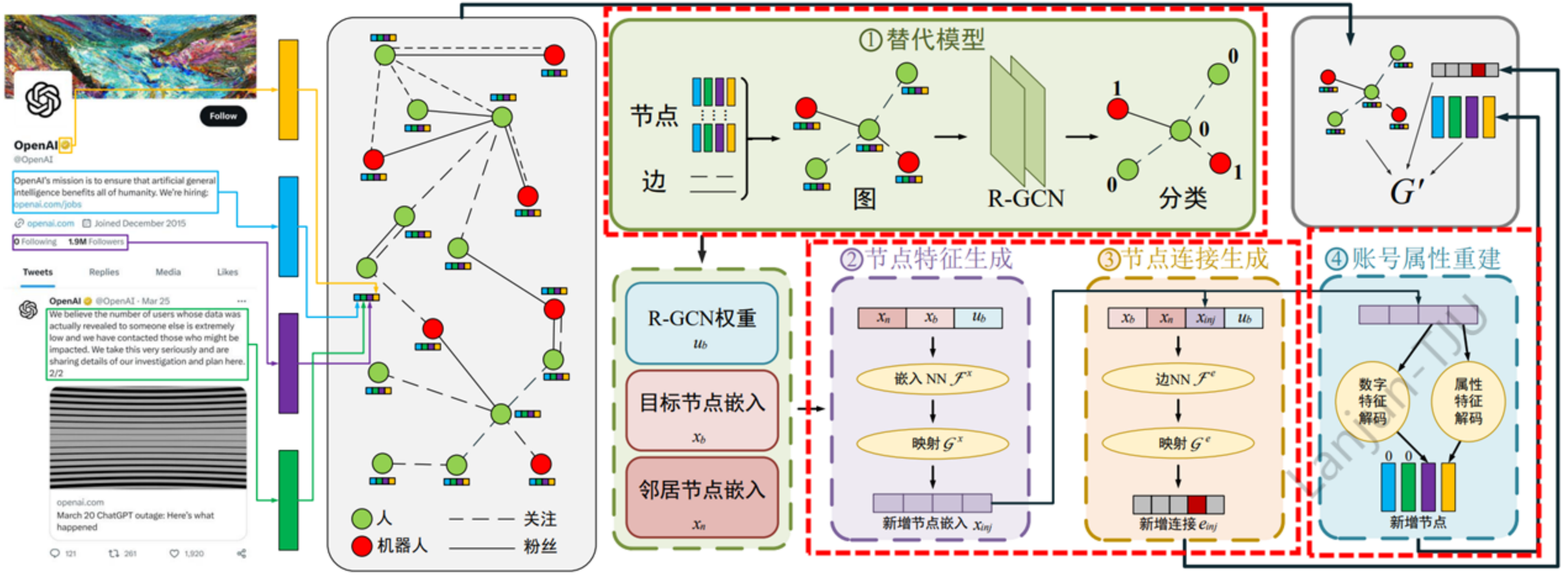


[1] My Brother Helps Me: Node Injection Based Adversarial Attack on Social Bot Detection. ACM MM 2023

机器人检测及鲁棒性评估优化

■ 评估：节点注入攻击

- 黑盒攻击：构建**替代模型**，提供检测模型的梯度信息；
- 不可察觉：单节点注入攻击，扰动限制为**“单点单边”**；
- 属性重建：构建**属性恢复**模块，重建新增节点用户属性。



机器人检测及鲁棒性评估优化

■ 评估：所提出评估方法在六种不同类型的检测模型上进行实验

攻击效果：**Cresci-2015**：**93.02%-95.74%**，　**TwiBot-22**：**56.36%-97.15%**

Dataset	Threat Model	ASR ↑	NNBB ↓	R-ASR(ATRM) ↓	R-NNBB(ATRM) ↑
Cresci-2015	GCN	95.68 ± 1.44	0.00 ± 0.00	3.61 ± 2.98	96.39 ± 3.01
	HGT	94.79 ± 1.18	0.06 ± 0.12	0.00 ± 0.00	100.00 ± 0.00
	Simple-HGN	95.74 ± 1.25	0.00 ± 0.00	2.37 ± 0.42	100.00 ± 0.00
	R-GCN	95.74 ± 1.50	0.06 ± 0.12	0.00 ± 0.00	100.00 ± 0.00
	GAT	93.49 ± 1.11	1.78 ± 0.00	0.00 ± 0.00	100.00 ± 0.00
	RGT	93.02 ± 1.37	0.00 ± 0.00	4.02 ± 0.68	93.49 ± 2.29
TwiBot-22	GCN	78.36 ± 22.39	7.24 ± 17.33	22.16 ± 1.94	99.83 ± 0.17
	HGT	97.15 ± 2.31	5.89 ± 9.01	42.35 ± 14.38	99.94 ± 0.09
	Simple-HGN	56.36 ± 5.34	9.64 ± 10.88	45.84 ± 6.51	82.23 ± 3.93
	R-GCN	67.69 ± 13.99	9.31 ± 14.40	36.04 ± 7.90	99.50 ± 0.73
	GAT	70.09 ± 21.72	0.00 ± 0.00	27.09 ± 4.47	100.00 ± 0.00
	RGT	63.19 ± 2.21	0.00 ± 0.00	27.04 ± 5.09	98.06 ± 2.28

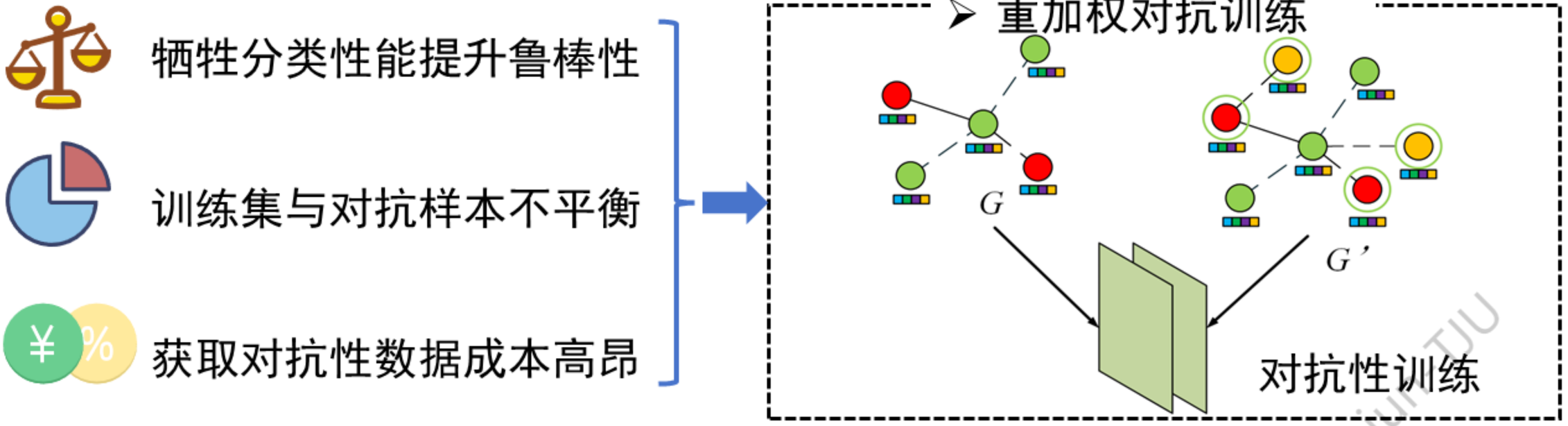
- **ASR**：攻击成功率，度量模型的鲁棒性，分值越高，代表模型鲁棒性越低。
- **NNBB**：新节点暴露率，度量新注入节点被社交机器人检测出来的概率。

不同类型的机器人检测模型均具有**网络结构攻击鲁棒性脆弱**

机器人检测及鲁棒性评估优化

■ 研究挑战

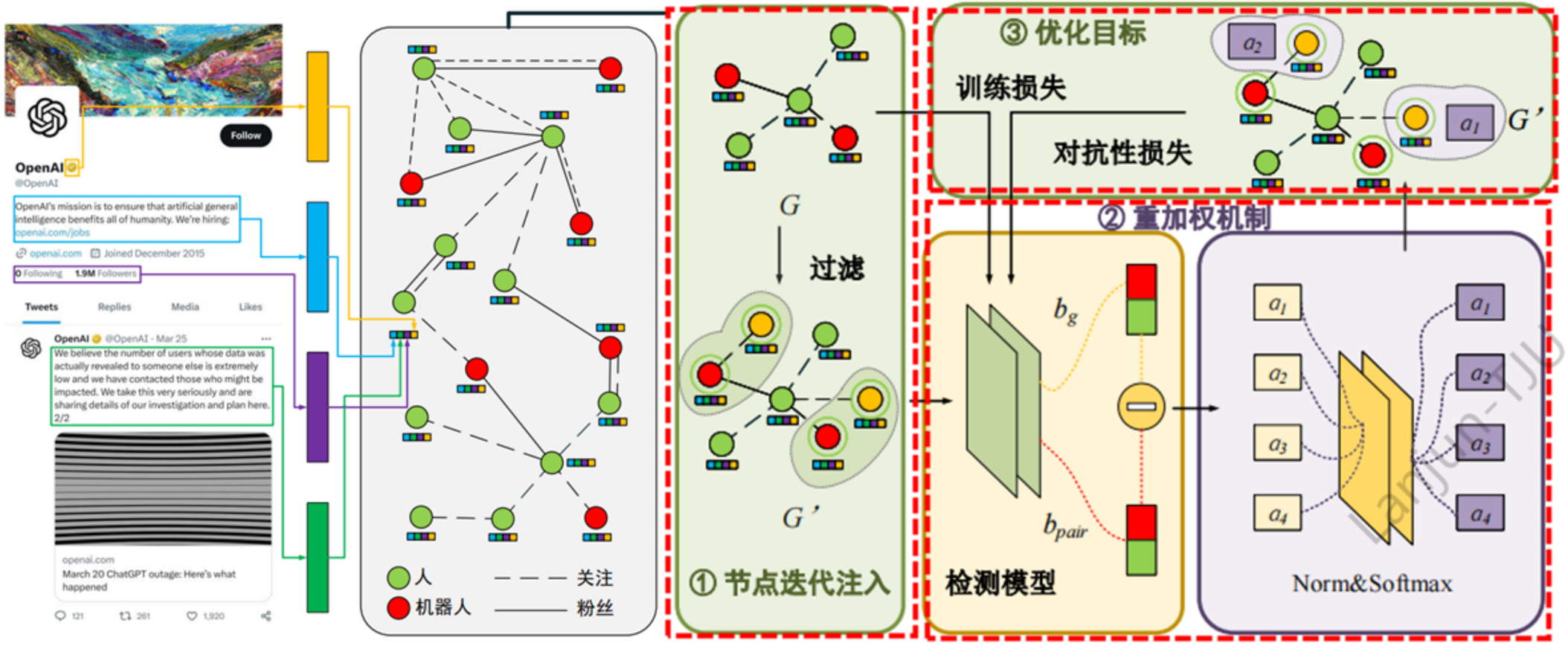
- □优化：**对抗性训练 -> 增强模型鲁棒性**
- □挑战：①检测性能与鲁棒性平衡；②样本数量不平衡；③计算成本高昂；



机器人检测及鲁棒性评估优化

■ 优化：重加权对抗性训练

- □性能平衡：采用**对抗性训练**，模型训练过程引入对抗性信息；
- □降低成本：**节点迭代注入**策略，一次性获取所有对抗性样本。
- □提高质量：控制对抗性样本数量，**重加权机制**放大有效信息。



机器人检测及鲁棒性评估优化

■ 优化：所提出评估方法在六种不同类型的检测模型上进行实验

防御效果：**Cresci-2015**：**89.00%-95.74%↑**，　**TwiBot-22**：**10.52%-56.20% ↑**

Dataset	Threat Model	ASR ↑	NNBB ↓	R-ASR(ATRM) ↓	R-NNBB(ATRM) ↑
Cresci-2015	GCN	95.68 ± 1.44	0.00 ± 0.00	3.61 ± 2.98	96.39 ± 3.01
	HGT	94.79 ± 1.18	0.06 ± 0.12	0.00 ± 0.00	100.00 ± 0.00
	Simple-HGN	95.74 ± 1.25	0.00 ± 0.00	2.37 ± 0.42	100.00 ± 0.00
	R-GCN	95.74 ± 1.50	0.06 ± 0.12	0.00 ± 0.00	100.00 ± 0.00
	GAT	93.49 ± 1.11	1.78 ± 0.00	0.00 ± 0.00	100.00 ± 0.00
	RGT	93.02 ± 1.37	0.00 ± 0.00	4.02 ± 0.68	93.49 ± 2.29
TwiBot-22	GCN	78.36 ± 22.39	7.24 ± 17.33	22.16 ± 1.94	99.83 ± 0.17
	HGT	97.15 ± 2.31	5.89 ± 9.01	42.35 ± 14.38	99.94 ± 0.09
	Simple-HGN	56.36 ± 5.34	9.64 ± 10.88	45.84 ± 6.51	82.23 ± 3.93
	R-GCN	67.69 ± 13.99	9.31 ± 14.40	36.04 ± 7.90	99.50 ± 0.73
	GAT	70.09 ± 21.72	0.00 ± 0.00	27.09 ± 4.47	100.00 ± 0.00
	RGT	63.19 ± 2.21	0.00 ± 0.00	27.04 ± 5.09	98.06 ± 2.28

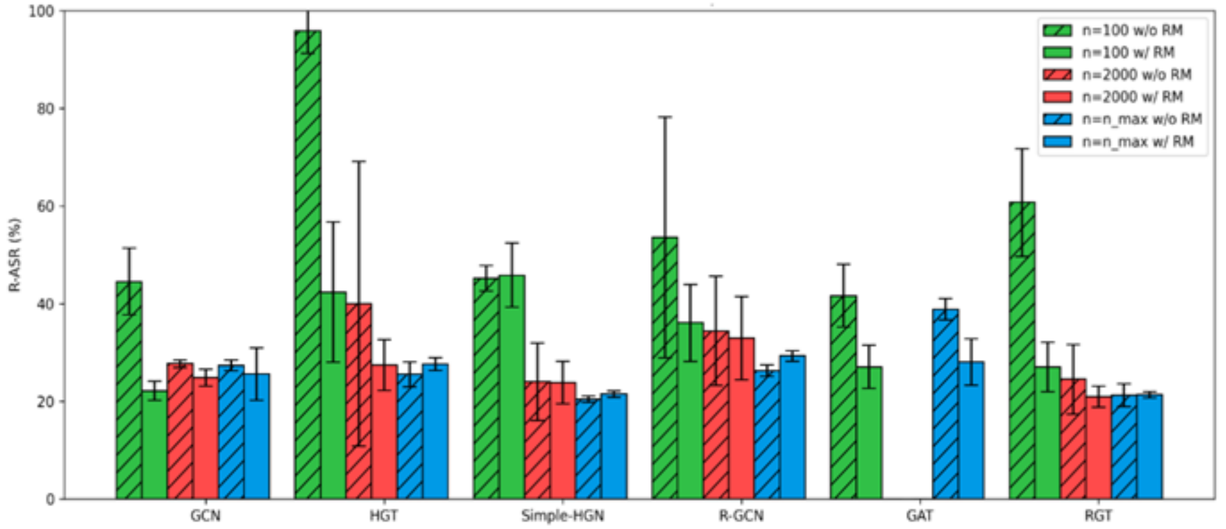
- **R-ASR**：再攻击成功率，分值越低，代表模型鲁棒性越高。
- **R-NNBB**：再攻击新节点暴露率，分值越高，代表模型鲁棒性越高。

基于重加权的对抗训练方法可以**有效提升检测模型的鲁棒性**。

机器人检测及鲁棒性评估优化

■ 优化：参数分析及性能验证

1. 对抗性样本数对于实验性能的影响



2. 检测性能与鲁棒性的平衡分析

受害者模型	BAC (Cresci-2015)		BAC (TwiBot-22)	
	T	ATRM	T	ATRM
GCN	94.05 ± 1.00	93.46 ± 1.19	76.99 ± 0.21	76.69 ± 0.29
HGT	93.61 ± 1.77	92.48 ± 3.09	75.04 ± 1.95	76.74 ± 0.53
Simple-HGN	93.57 ± 1.39	94.54 ± 0.67	77.97 ± 0.29	78.17 ± 0.12
R-GCN	93.64 ± 1.61	94.95 ± 0.55	75.31 ± 2.52	76.77 ± 0.42
GAT	93.88 ± 1.15	93.91 ± 1.08	77.07 ± 0.13	76.19 ± 1.78
RGT	94.24 ± 1.02	90.09 ± 5.06	78.37 ± 0.33	78.19 ± 0.27

➢ **BAC**：检测模型的分类性能，分值越高，代表模型分类能力越好。

结论：基于重加权的对抗训练方法

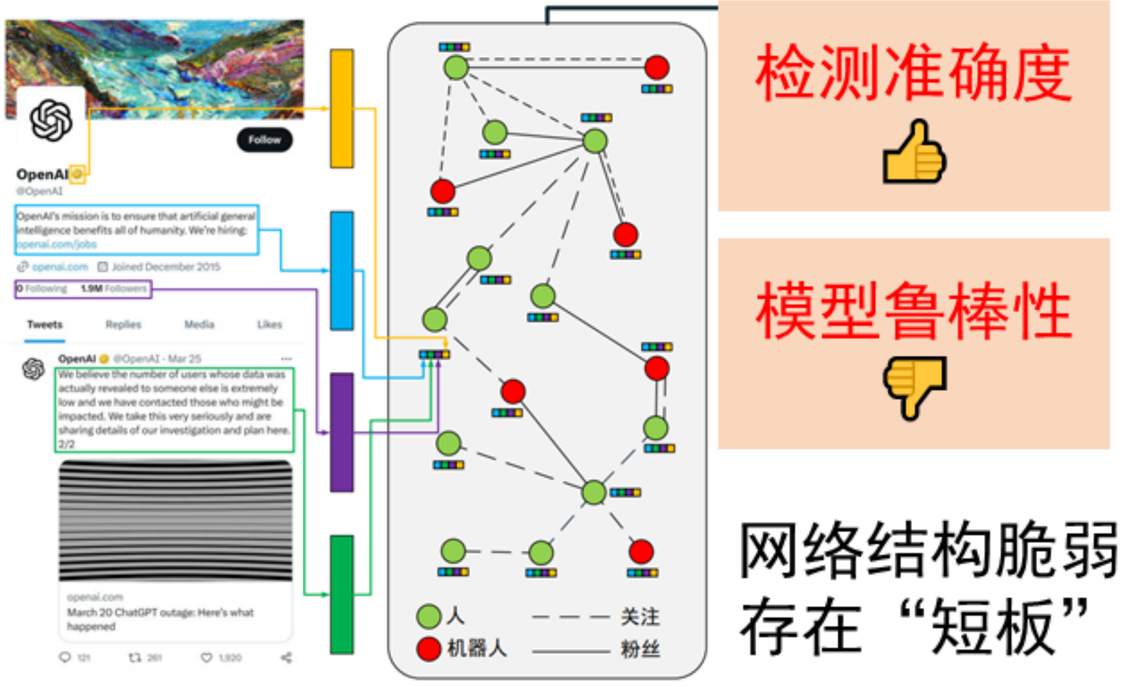
1. 在对抗样本数稀少的情况下实现近乎一致的性能。

2. 在提升检测模型的鲁棒性和保持检测性能之间取得了良好的平衡。

小结

■ 机器人检测及鲁棒性评估优化

➢ 内容总结



检测准确度

模型鲁棒性

网络结构脆弱
存在“短板”

- 机器人检测模型准确度
  - 充分研究，取得优越性能
- 机器人检测模型鲁棒性
  - 关注较少，仍有潜在漏洞

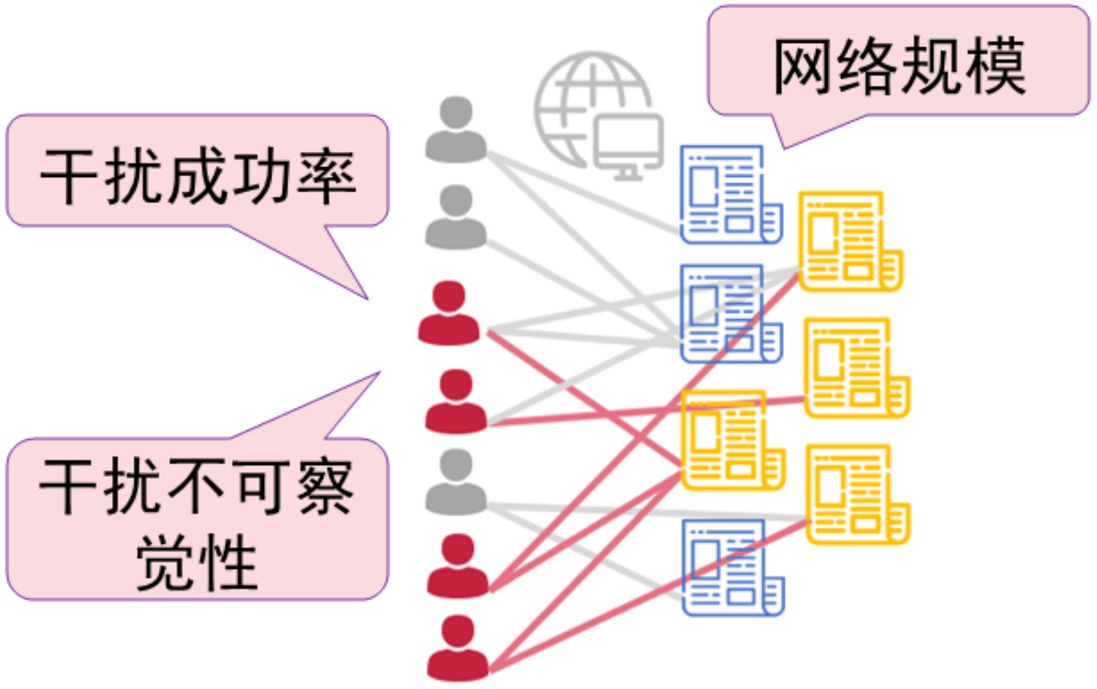
报告内容

- 研究背景及意义
- 恶意机器人干扰虚假信息检测
- 可控机器人干预控制传播范围
- 机器人检测及鲁棒性评估优化
- 总结与展望

总结与展望

■ 恶意机器人干扰虚假信息检测

➢ 领域现状



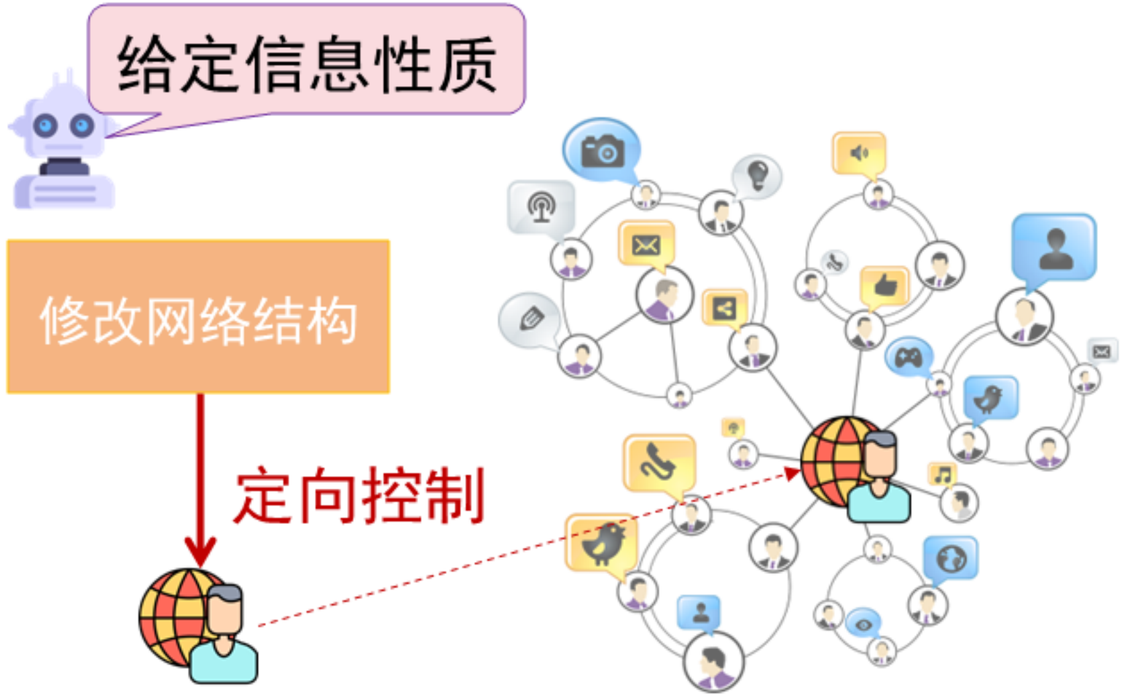
➢ 未来展望

- □ 干扰成功率和不可察觉性的平衡
- □ 不可察觉性
  - ◆ 计算复杂度优化
  - ◆ 指标定义待统一
- □ 网络规模扩展性
  - ◆ 扩展到十万用户规模以上

总结与展望

■ 可控机器人干预控制传播范围

➢ 领域现状



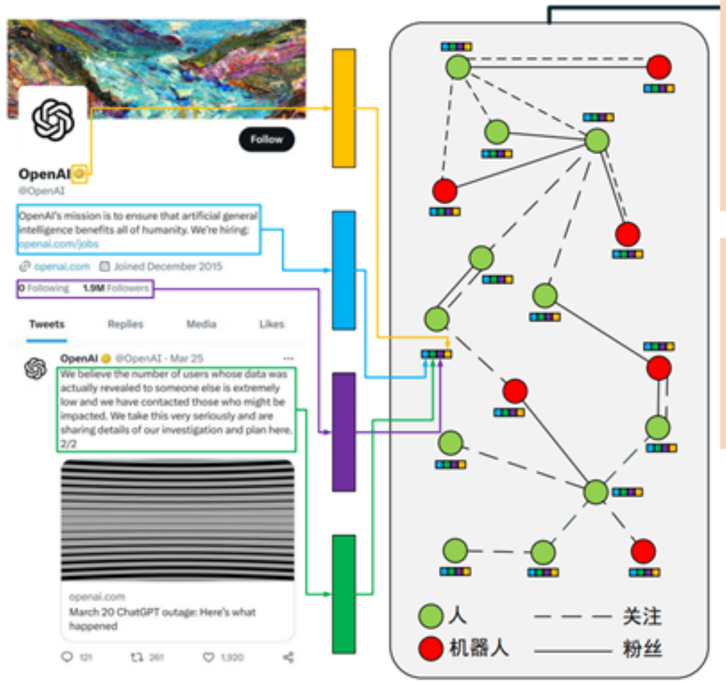
➢ 未来展望

- 多目标控制
  - ◆ 单目标控制的传播范围有限
- 混合控制
  - ◆ 单节点/边控制的传播效果有限
- 自适应控制
  - ◆ 基于传播环境自适应选择目标和控制策略

总结与展望

■ 机器人检测及鲁棒性评估优化

➢ 领域现状



检测准确度

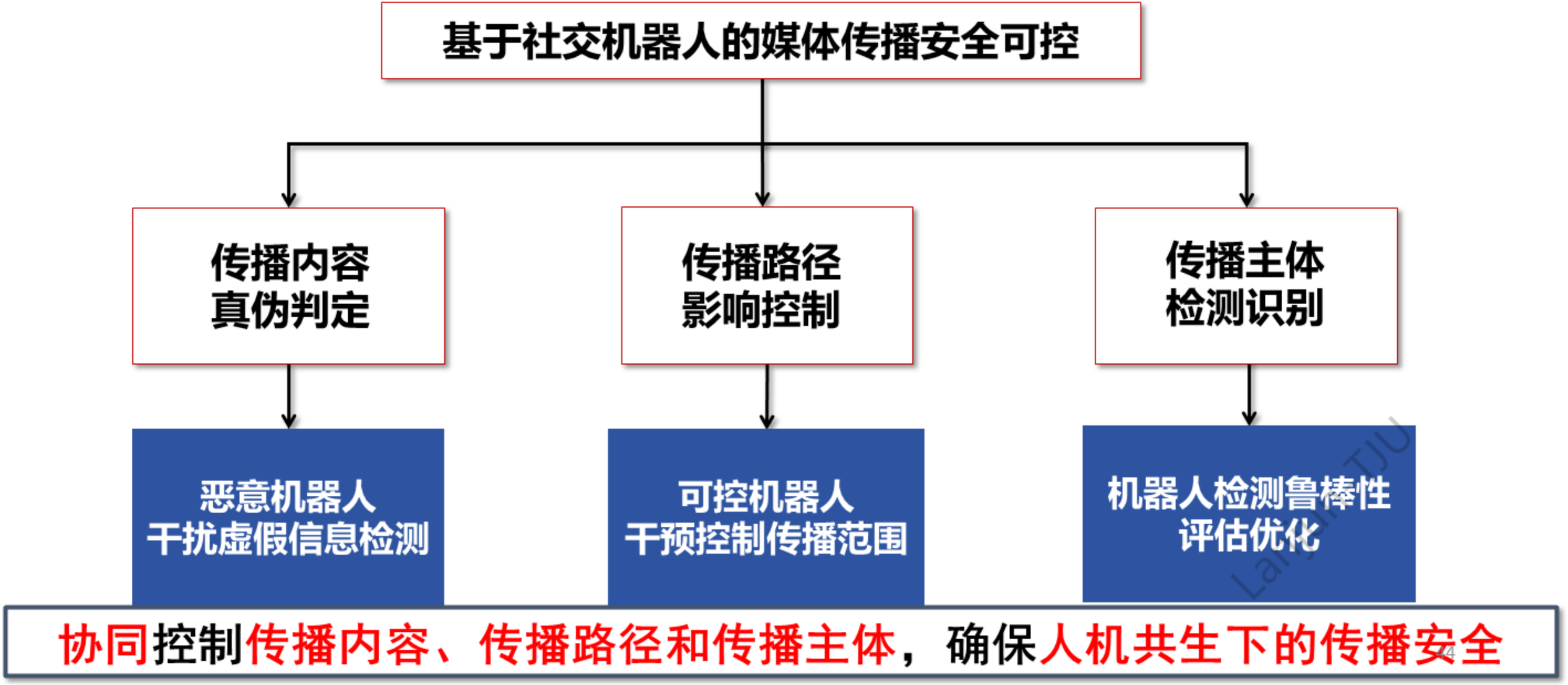
模型鲁棒性

“短板”需要继续挖掘

➢ 未来展望

- 群组性攻击
  - 使用群组注入攻击全网络节点
- 生成式内容
  - 使用AIGC技术生成攻击文本内容

总结与展望



总结与展望

■ 数据：传播内容真伪判定、传播路径安全可控、传播主体检测识别任务的数据集调研[1]

□ 从传播内容、路径和主体三个角度收集了对应**10种分支任务**的**共54个数据集**

传播内容

虚假信息检测 |23

Rumor Det.	9
Fake News Det.	14

传播路径

宏观信息扩散预测|12

Cascade Size Pred.	8
Popularity Pred.	5
User Attitudes Pred.	1

微观信息扩散预测|14

Next User Pred.	6
Social Influence Pred.	6

传播主体

社交机器人检测 |8

Feature-based Bot Det.	3
Text-based Bot Det.	2
Graph-based Bot Det.	3

□ 比较了数据集来自传播内容、路径和主体三个角度的**六大属性**

传播内容 真实性标签

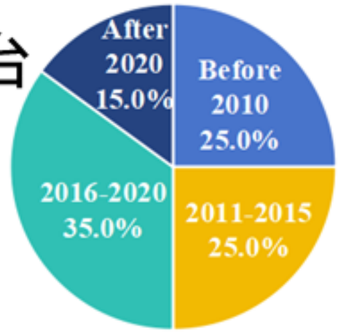
社交网络 传播网络

用户属性 机器人标签

[1] A Survey of Datasets for Information Diffusion Task. arxiv 2024

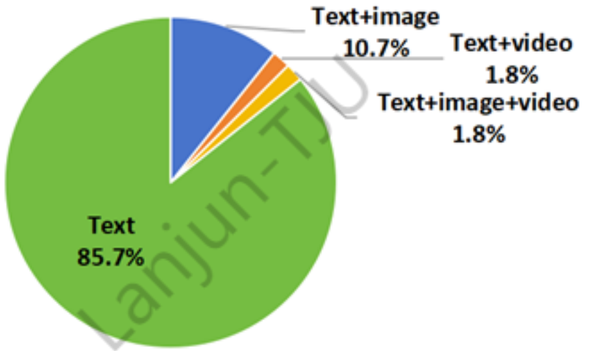
总结与展望

- 数据：传播内容真伪判定、传播路径安全可控、传播主体检测识别任务的数据集调研[1]
  - 结论：缺乏合适的传播数据集
    - 属性：数据集的属性（用户属性、社交网络、机器人标签、传播内容、传播网络和真实性标签）不完整，**不能满足三类任务同时进行**
    - 时间范围较陈旧：2020年之前的数据集占75%
    - 语言单一：英语数据集占近80%
    - 模态单一：文本模态数据集占85%
    - 来源平台单一：Twitter平台



时间范围统计

语言统计

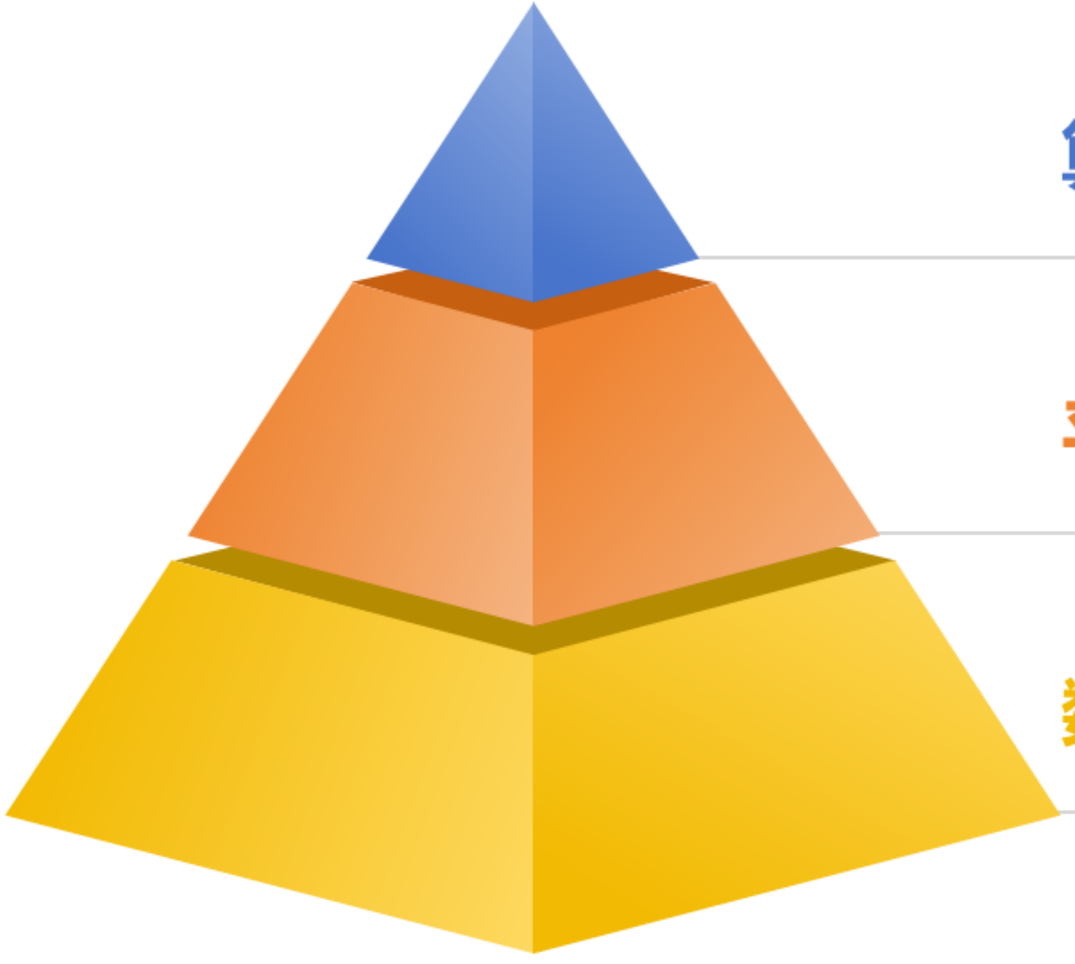


模态统计

[1] A Survey of Datasets for Information Diffusion Task. arxiv 2024

总结与展望

■ 协同实现多媒体传播安全可控



算法 **保障**社交机器人内容，路径及主体传播安全

平台 **缺乏**真实/虚假舆论场的传播验证平台

数据 **缺乏**包含内容，路径及主体的协同数据集

感谢各位专家
敬请批评指正

https://wanglanjun-academic.github.io/