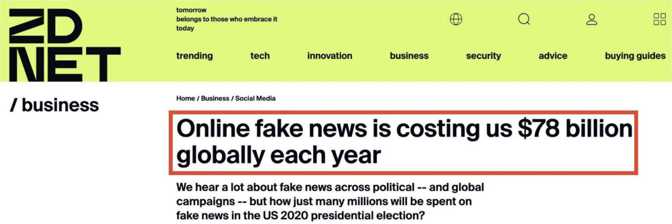
Attacking Fake News Detectors via Manipulating News Social Engagement

Haoran Wang[1], Yingtong Dou[2,4], Canyu Chen[1],
Lichao Sun[3], Philip S. Yu[2], Kai Shu[1]

[1]Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA
[2]Department of Computer Science, University of Illinois Chicago, Chicago, IL, USA
[3]Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA
[4]Visa Research, Palo Alto, CA, USA

Why Misinformation Research is Important?

- Research [1] has revealed that fake news is costing the global economy **$78 billion** each year.
- Social media is the **main source of news consumption** for younger generations [2].



[1] The Economic Cost of Bad Actors on the Internet https://s3.amazonaws.com/media.mediapost.com/uploads/EconomicCostOfFakeNews.pdf [2] The news consumption habits of 16- to 40-year-olds https://www.americanpressinstitute.org/publications/reports/survey-research/the-news-consumption-habits-of-16-to-40-year-olds/

Are Existing Fake News Detectors Robust?

Previous works [1,2] have shown text-based detectors are **vulnerable** to adversarial attacks.

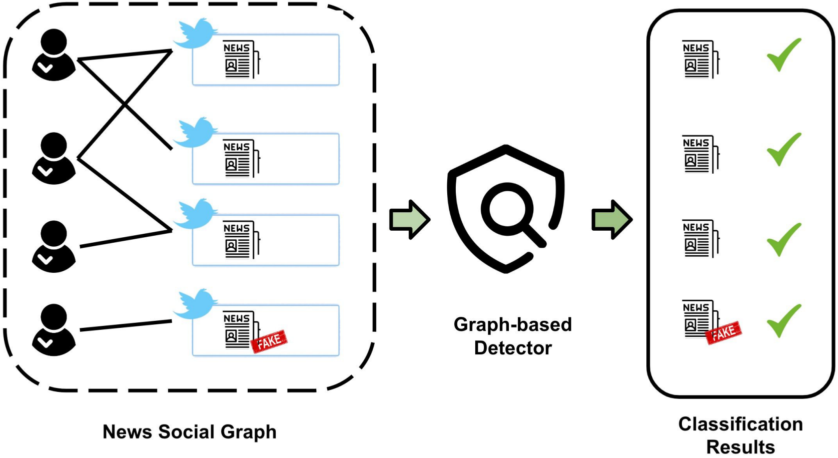
Text-based Detector

Social Context-based Detector

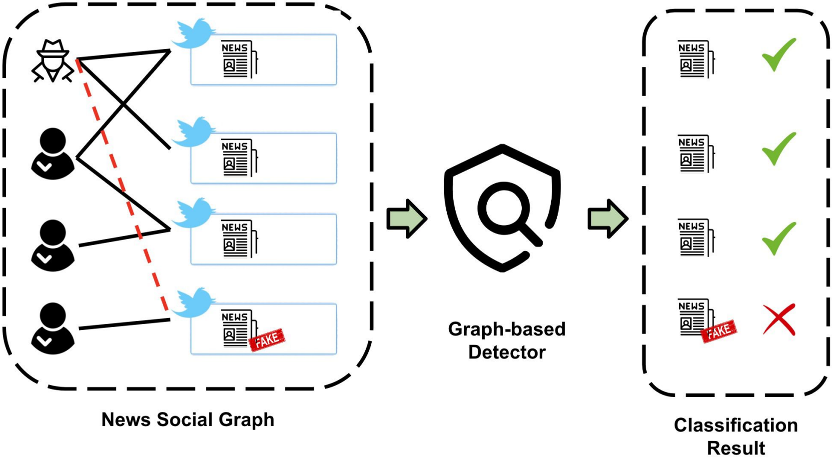
What about social-context-based detectors?

[1] He, Bing, Mustaque Ahamad, and Srijan Kumar. "Petgen: Personalized text generation attack on deep sequence embedding-based classification models." Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021. [2] Le, Thai, Suhang Wang, and Dongwon Lee. "Malcom: Generating malicious comments to attack neural fake news detection models." 2020 IEEE International Conference on Data Mining (ICDM). IEEE, 2020.

What are Social-Context-based Detectors?



Attack by Manipulating Social Engagement



Fake News Campaign: A Coordinated Effort

- Research [1] has shown that various coordinated groups are involved in spreading misinformation.
- We classify accounts into bot, cyborg, and crowd worker agents. Each type of agent has its own cost and influence.
- Attackers have a fixed budget for the number of agents they can use.

Social Bot **Cyborg** **Crowd Worker**

[1] Pacheco, Diogo, Alessandro Flammini, and Filippo Menczer. "Unveiling coordinated groups behind white helmets disinformation." Companion proceedings of the web conference 2020. 2020.

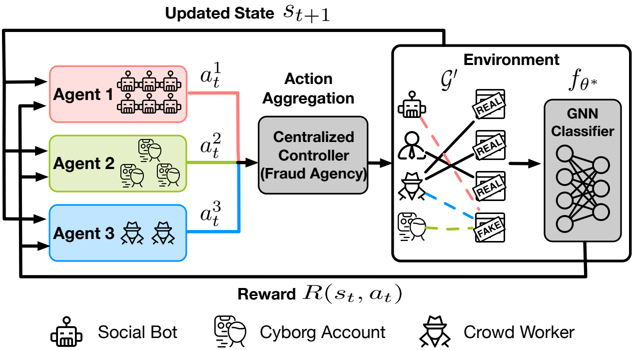
Malicious Actors

- Social Bot: registered and fully controlled by automated programs. Accounts with only one connection.
- Cyborg: registered by human and partially controlled by automated programs. Accounts with more than 10 connections.
- Crowd Worker: credible and rich social profiles. Accounts with more than 20 connections, where 100% of them connect to real news.

Agent	Cost	Influence	Budget
Bot	low	low	high
Cyborg	moderate	moderate	moderate
Crowd Worker	high	high	low

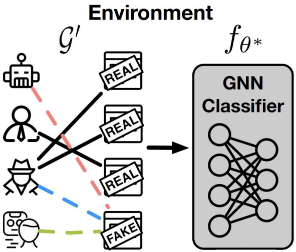
Proposed Framework

Multi-agent Reinforcement Learning Framework (MARL)



Cooperative Markov Game

- **Action:** Each controlled user can only add edges to target news. A centralized controller aggregates agent actions.
- **State:** All agents share the same state at each episode.
- **Reward:** At each episode, we reward each agent for successfully flipping the classification result of target news.



Experimental Settings

- Random-Edge: randomly selects controlled users and target news to add edges.
- Random-Node: randomly injects new user nodes and connects them with the target news.
- Single Agent RL: limits to a single type of agent.

Data	U	V	E
Politifact	276,277	581	1,074,890
Gossipcop	565,660	10,333	3,084,931

Models	Politifact		Gossipcop	
	Accuracy	F1	Accuracy	F1
GCN	0.8673	0.8632	0.8278	0.7864
GAT	0.8600	0.8543	0.8423	0.8010
SAGE	0.8034	0.7973	0.8824	0.8636

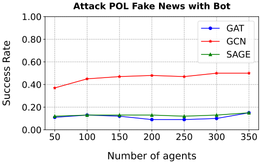
Key Findings

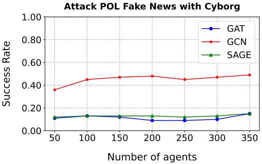
- MARL is effective at attacking fake news in both datasets.
- GCN is more vulnerable compared to GAT and GraphSAGE.
- MARL has better overall performance on Politifact than Gossipcop.

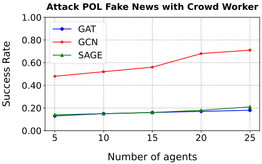
Method	Politifact						Gossipcop					
	Fake			Real			Fake			Real		
	GAT	GCN	SAGE	GAT	GCN	SAGE	GAT	GCN	SAGE	GAT	GCN	SAGE
RD-Edge	0.14	0.45	0.13	0.11	0.33	0.15	0.06	0.28	0.25	0.08	0.22	0.14
RD-Node	0.12	0.48	0.14	0.13	**0.38**	0.15	0.12	0.32	0.22	0.12	0.23	0.16
RL - A1	0.17	0.42	0.16	0.08	0.07	**0.21**	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**

Key Findings

- The overall attack performance increases with incremental attack budget for all three types of agents.
- Crowd worker agents have more influence than bot and cyborg agents.







Key Findings

- News with higher degrees is more robust than news with lower degrees.
- GAT has less performance drop on news with low and mid degrees compared to GCN and GraphSAGE.

News Degrees	Politifact			Gossipcop		
	GAT	GCN	SAGE	GAT	GCN	SAGE
Low	0.16	0.30	0.21	0.14	0.22	0.25
Mid	0.14	0.15	0.11	0.11	0.13	0.12
High	0.03	0.06	0.03	0.02	0.02	0.05

Conclusion

- We are **the first work** to study the robustness of social-context-based fake news detectors.
- MARL **mimics real-world misinformation campaign** by employing different types of agents.
- Existing social-context-based detectors are **vulnerable** to social engagement attacks.

Future Works

- We would like to **automate** the process of selecting optimal agents for action aggregation.
- We would like to find a way to **effectively reduce** the search space of Q-network.
- We would like to explore a more **complex** MARL framework and test on more robust GNNs.

Thank you! Any Questions?

- Paper: https://arxiv.org/abs/2302.07363
- Code: https://github.com/hwang219/AttackFakeNews