UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Defending Network Intrusion Detection Systems Based on Graph Neural Networks Against Structural Adversarial Attacks

Dimitri Galli, Andrea Venturi, Dario Stabili, Mauro Andreolini, Mirco Marchetti

dimitri.galli@unimore.it, aventuri@vicomtech.org, dario.stabili@unimore.it, mauro.andreolini@unimore.it, mirco.marchetti@unimore.it

NCA 2025

November 5-7, 2025 - Lisbon, Portugal

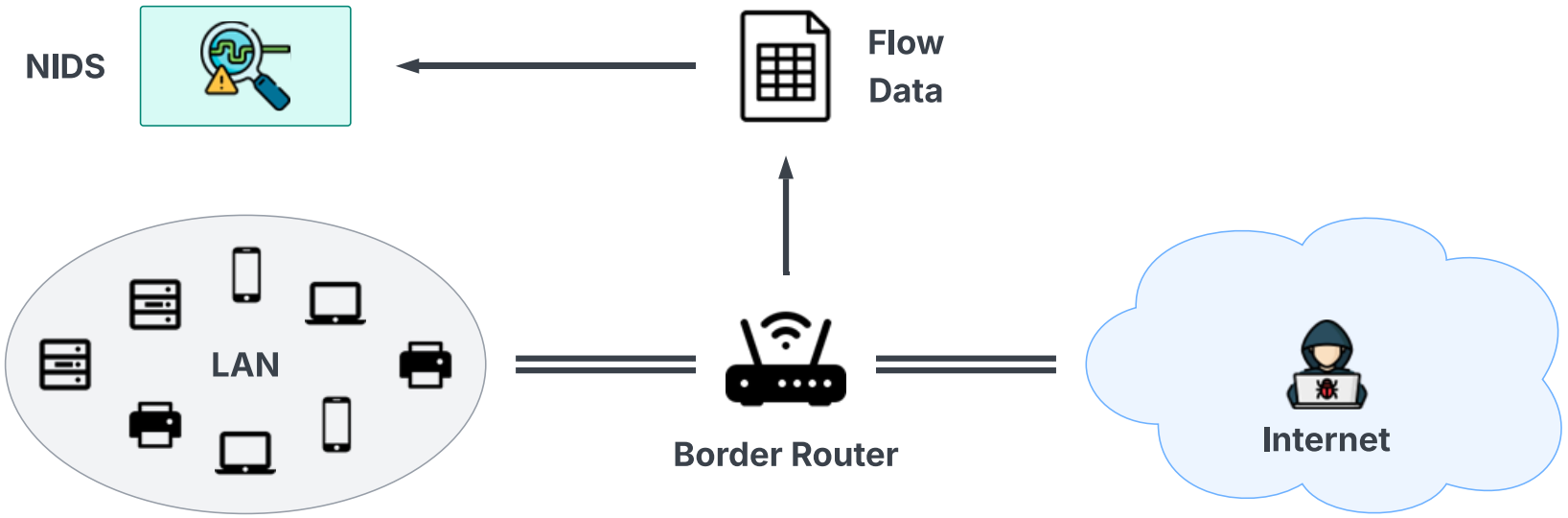
ML-NIDS

Digital threats are continuously evolving

- ML solutions improve adaptability and effectiveness of NIDS

ML-NIDS are vulnerable to adversarial manipulation

- Ensuring robustness against such attacks is essential for real-world deployment



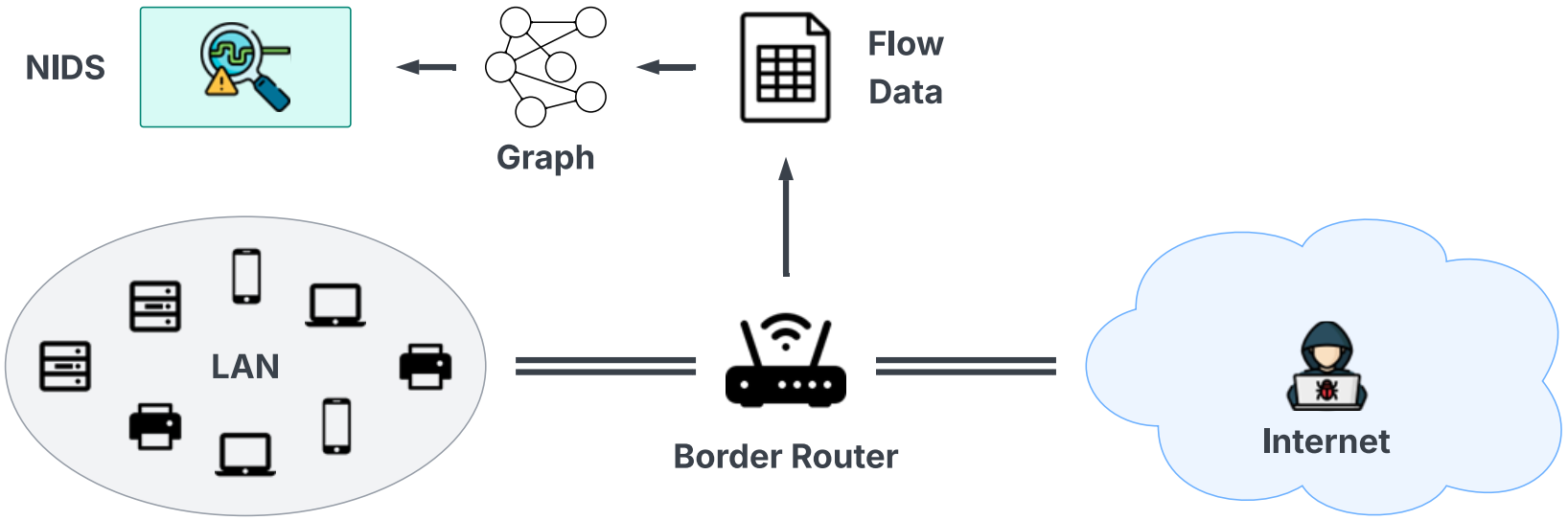
GNN-NIDS

GNNs define network traffic as a graph

- Hosts as nodes, communications as edges

GNN-NIDS enable context-aware intrusion detection

- Their reliance on graph structure introduces new vulnerabilities



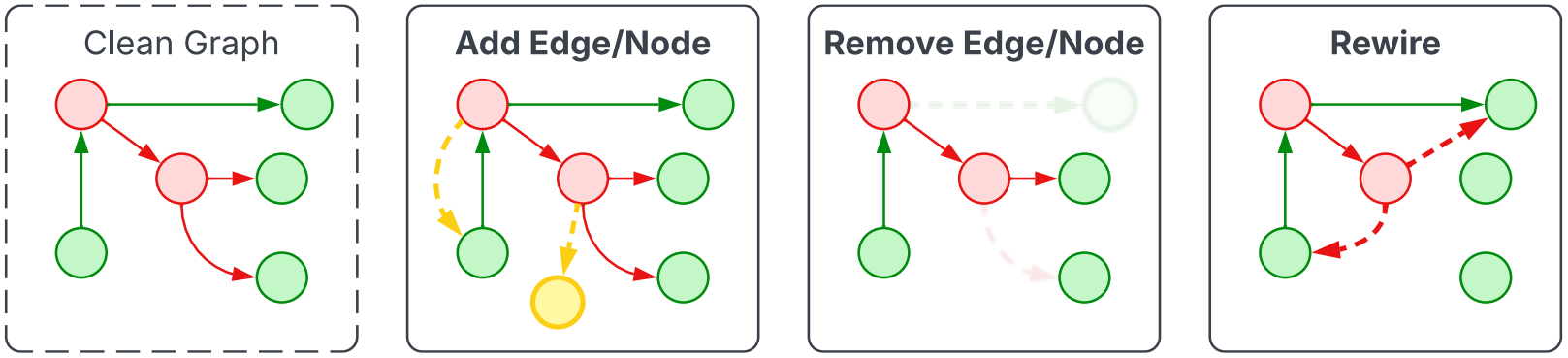
Structural Adversarial Attacks

Adversaries can alter the graph topology

- Inject, remove, or rewire edges/nodes to deceive the GNN
- E.g., establish benign-looking communications to hide malicious patterns

Consequences

- Structural perturbations can cause misclassifications
- Attackers can modify the input network graph to evade detection



Existing Adversarial Defenses

Existing approaches

- Assume knowledge of specific attack strategies *(attack-specific)*
- Impose high computational cost at detection time *(heavy)*
- Overlook real-world network constraints *(unrealistic)*

Our method

- Modifies structural features of legitimate samples *(attack-agnostic)*
- Operates on the netflows in the training set *(lightweight)*
- Preserves network traffic semantics *(practical)*

Contributions

Adversarial defense for GNN-NIDS

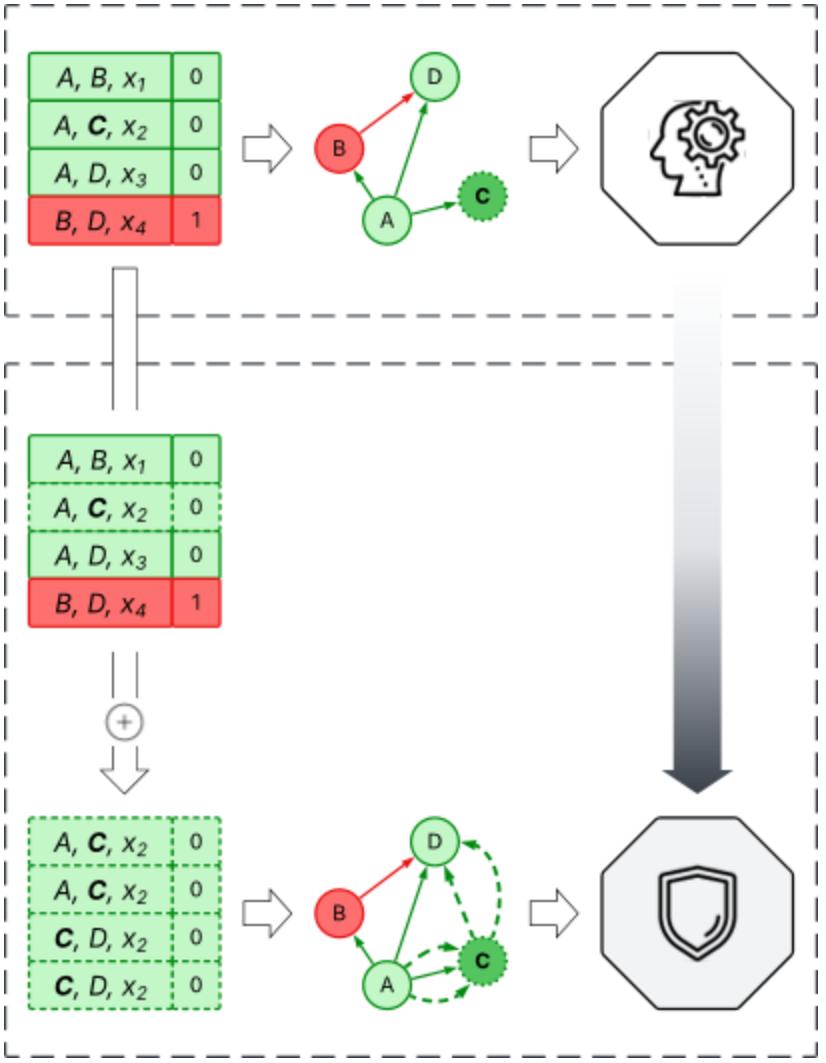
- Generate adversarial flows by modifying endpoints of benign netflows

Two-phase learning framework

- *Warm-up* on clean records
- *Hardening* on perturbed samples

Experimental testbed

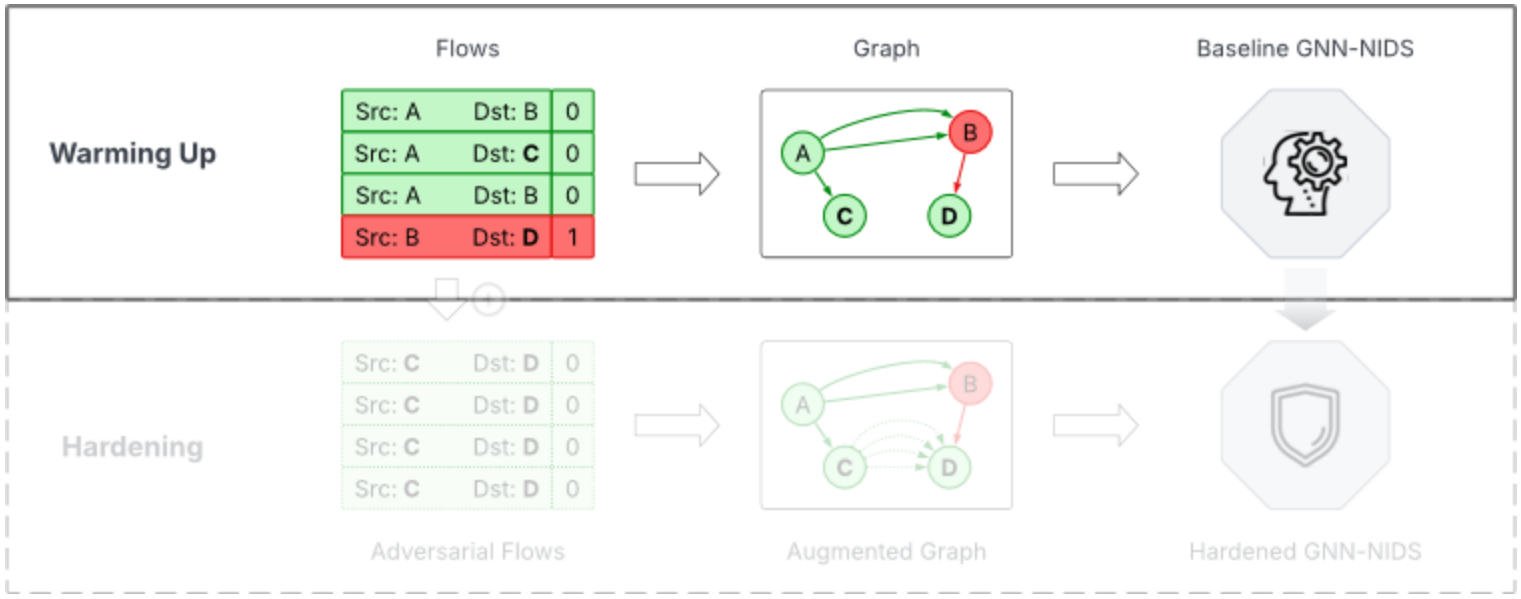
- 2 public real-world datasets
- 13 attack-specific GNN detectors
- 2 different evaluation scenarios



Defense Framework: Warming Up

Goal: build decision boundaries on clean examples

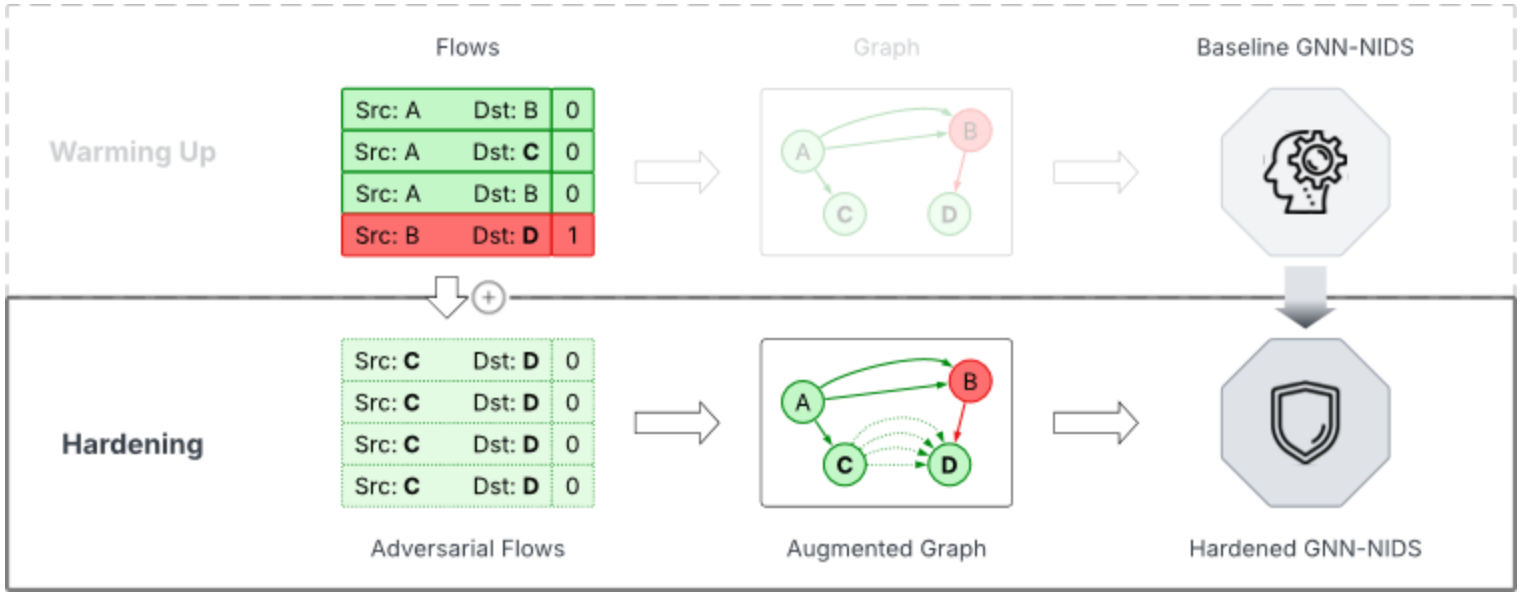
- Train GNN on unperturbed flow graphs only
- Learn benign and malicious structural patterns
- Capture communications between network hosts
- Avoid instability from early exposure to perturbations



Defense Framework: Hardening

Goal: improve performance against structural perturbations

- Generate adversarial data by replacing endpoints
- Leverage low-degree nodes as new hosts of benign flows
- Gradually increase the perturbation intensity during training
- Retrain on augmented graphs (clean + perturbed flow samples)



Testbed

Datasets

- *CTU-13* (enterprise) + *TON-IoT* (industrial)
- Labeled flows (10:1 benign-to-malicious ratio)

Detector

- *E-GraphSAGE* (edge-level GNN architecture)
- One binary classifier per threat type (13 in total)

Implementation

- *Warm-up:* standard training on clean graphs for 100 epochs
- *Hardening:* adversarial training with perturbed legitimate netflows

Evaluation

- *Clean traffic:* assess baseline detection performance
- *Perturbed traffic:* test robustness under structural attacks

Results: Clean Performance

Baseline vs hardened performance on **clean data**

- E-GraphSAGE tested on unperturbed flow graphs

F1-score

Dataset	**Threat**	Baseline	Hardened
CTU-13	*Neris*	0.880	**0.891**
	Rbot	0.989	**0.989**
	Virut	0.945	**0.945**
	Menti	0.996	**0.996**
	Murlo	0.984	**0.991**
	Average	0.959	**0.962**

F1-score

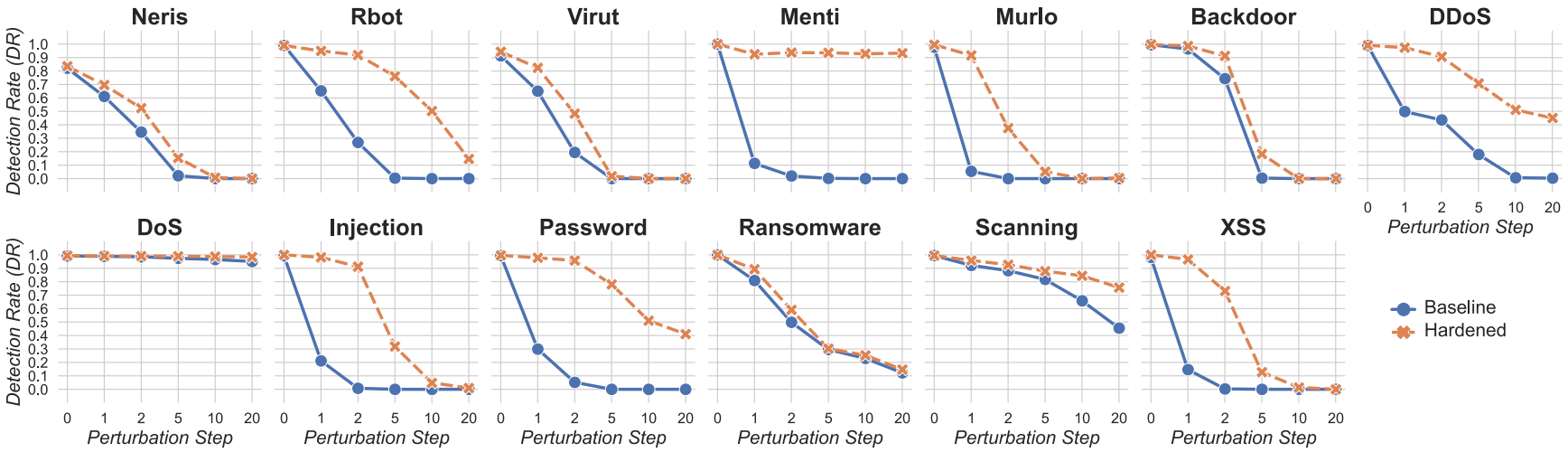
Dataset	**Threat**	Baseline	Hardened
TON-IoT	*Bkdr*	0.997	**0.997**
	DDoS	**0.992**	0.991
	DoS	0.993	**0.994**
	Inj	0.991	**0.994**
	Pswd	**0.994**	0.992
	Rans	0.998	**0.999**
	Scan	0.996	**0.996**
	XSS	0.987	**0.995**
	Average	0.994	**0.995**

- Hardened classifiers achieve comparable F1-scores to baseline
- Our defense does not degrade standard detection performance

Results: Adversarial Robustness (1)

Baseline vs hardened robustness against **C2x$_B$ attacks**

- E-GraphSAGE tested on perturbed graphs with new benign edges

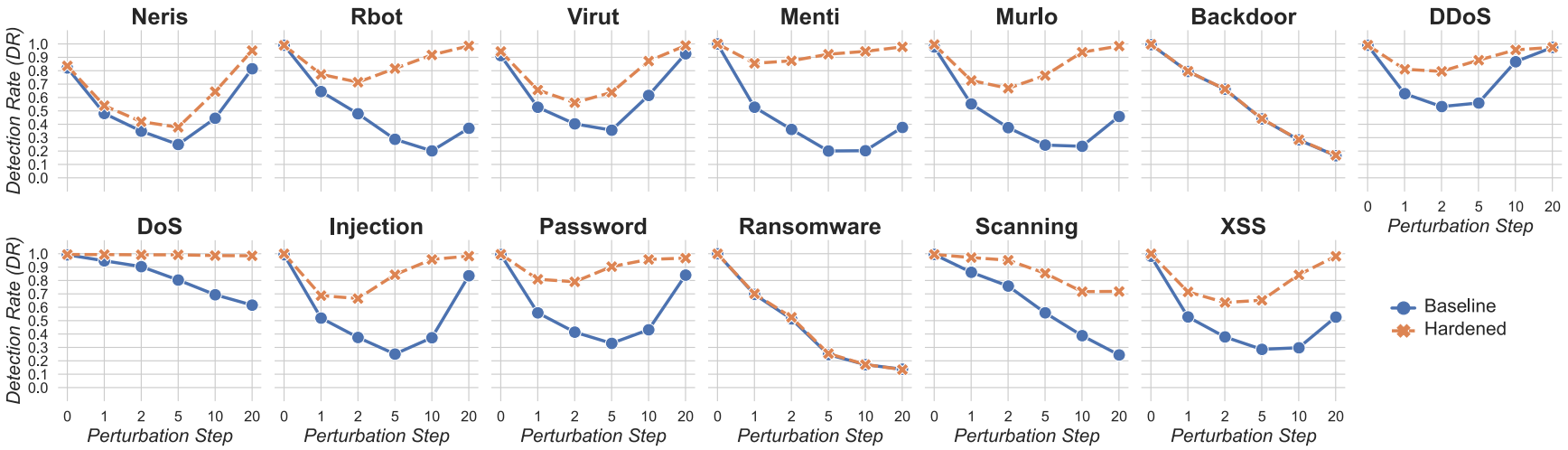


- Hardened classifiers show slower decline in DR
- Our framework enhances structural awareness

Results: Adversarial Robustness (2)

Baseline vs hardened robustness against **C2x$_M$ attacks**

- E-GraphSAGE tested on perturbed graphs with new malicious edges

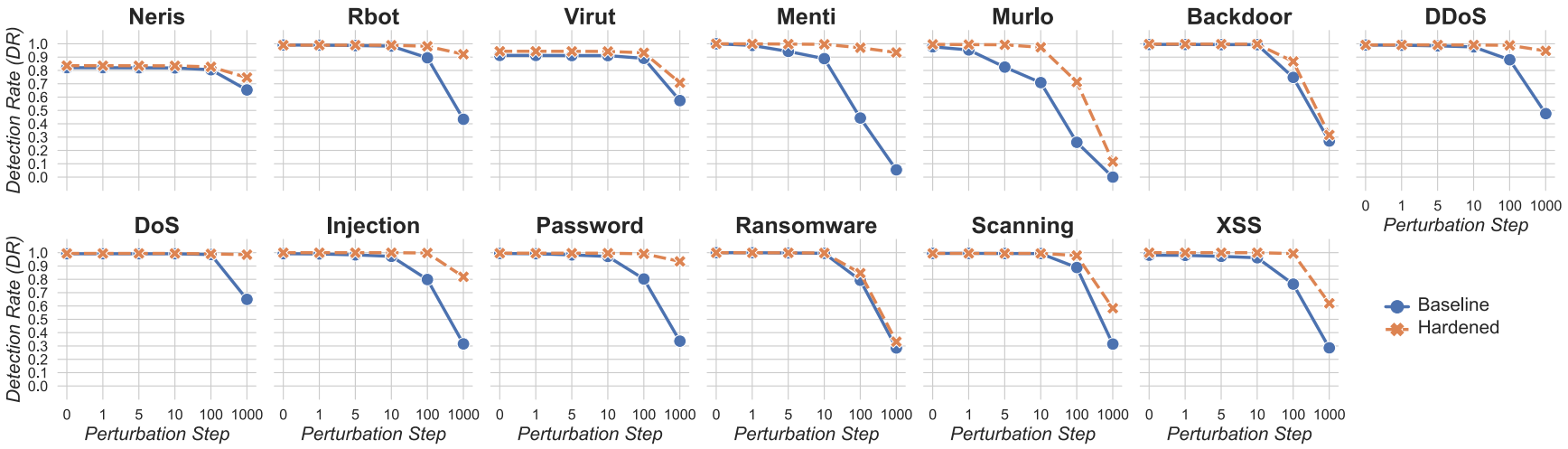


- Hardened classifiers outperform baseline across perturbations
- Our method improves generalization under large modifications

Results: Adversarial Robustness (3)

Baseline vs hardened robustness against **add node attacks**

- E-GraphSAGE tested on perturbed graphs with new unseen nodes



- Hardened classifiers remain stable with hundreds of new injected nodes
- Our strategy strengthens robustness against unseen graph structures

Conclusion

GNN-based NIDS constitute a **major advancement** in cybersecurity

- Reliance on graphs makes them vulnerable to structural attacks

We presented a novel two-phase **adversarial training** defense

- Our framework replaces topological attributes of benign netflows

We validated our proposal through an **extensive testbed**

- Results show improved robustness with no degradation on clean traffic

Source code available

https://github.com/dimgalli/defending-gnn-nids.git