Evaluating Explainability of Graph Neural Networks for Network Intrusion Detection with Structural Attacks

Dimitri Galli, **Andrea Venturi, Isabella Marasco, Mirco Marchetti**

dimitri.galli@unimore.it, andrea.venturi@unimore.it, isabella.marasco4@unibo.it, mirco.marchetti@unimore.it

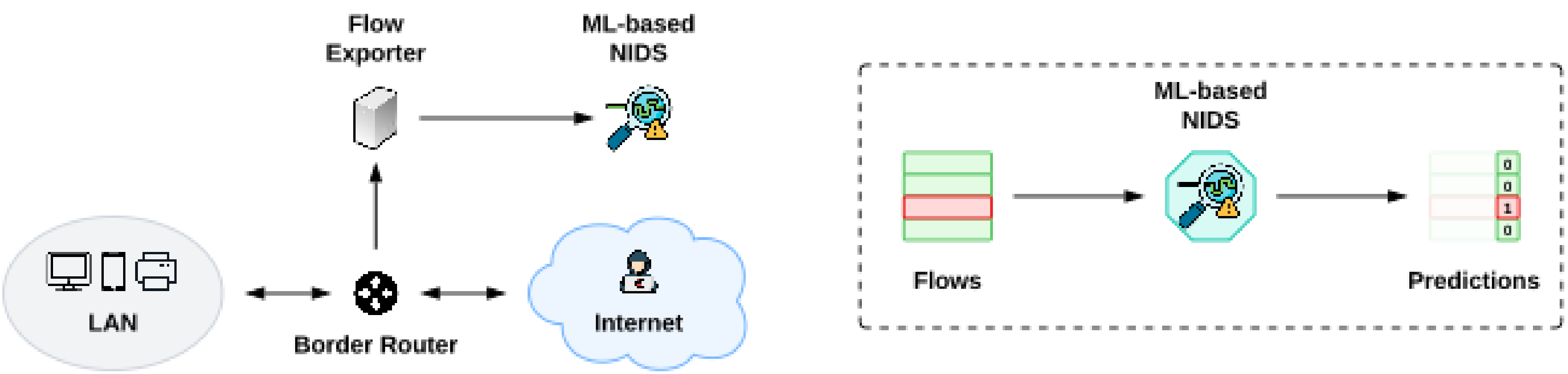
ITASEC25

February 3-8, 2025 - Bologna, Italy

ML-based NIDS

ML can enhance the detection capabilities of modern cyber threat detectors

- Traditional ML-based NIDS analyze features of individual **flows**



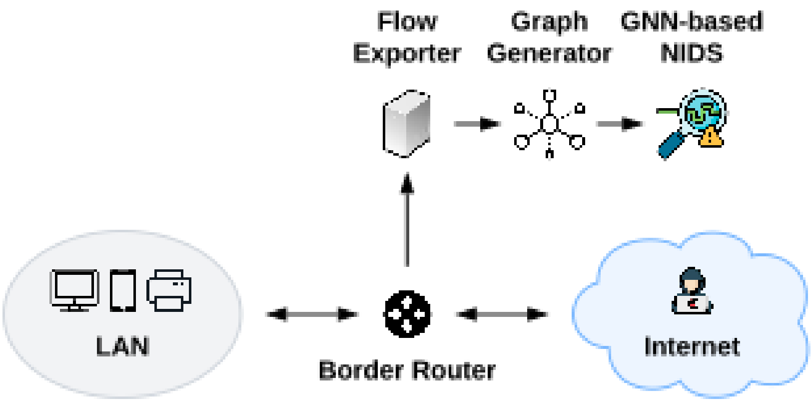
Limitations:

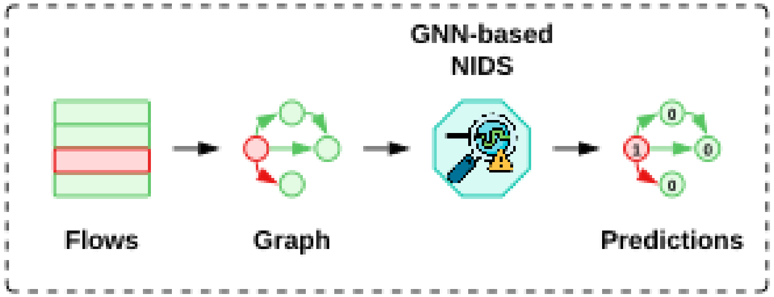
- ML algorithms fail to capture interdependencies in multi-flow attacks
- ML classifiers are vulnerable to adversarial manipulations of netflow features

GNN-based NIDS

GNN can improve performance by learning flow features and structural similarities

- GNN-based NIDS analyze network topology represented as **graphs**





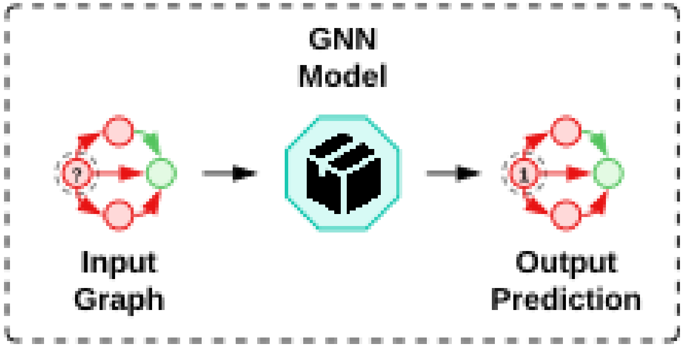
Limitations:

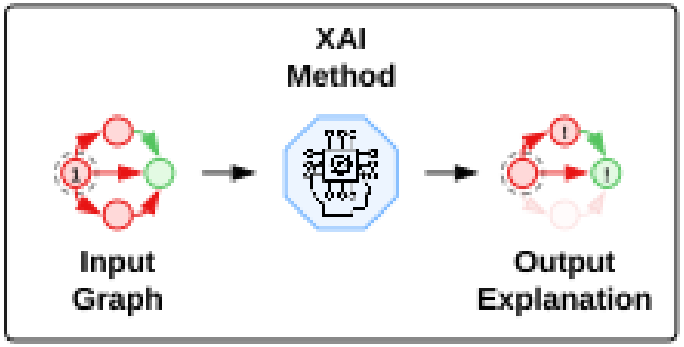
- GNN are opaque, acting as black boxes and lacking transparency
- GNN are vulnerable to adversarial perturbations of graph topology

XAI

XAI helps security practitioners understand GNN predictions

- Explainability methods define **masks** that contain relevance scores
- Explainers identify **subgraphs** that contribute most to intrusion detections





Approaches to evaluate explanations:

- Supervised approaches compare explanations with ground truth
- Unsupervised approaches evaluate how explanations impact predictions

Challenges in evaluating explanations:

- Generating ground truth labels is expensive
- Isolating subgraphs leads to breaking the network topology

Contributions

We develop an **evaluation framework** with key properties:

- Agnostic, i.e., independent of explainability methods
- Flexible, i.e., usable without ground truths
- Practical, i.e., useful in realistic scenarios

We present an **innovative methodology** to evaluate XAI methods in GNN-based NIDS

- Explainers identify important components within the graph
- Influential netflow records change the graph structure
- Perturbed network graphs fool the cyber detector

We propose a **case study** to validate our approach

- Two popular real-world datasets
- Thirteen SOTA attack-specific detectors
- Five different post-hoc explainers

Methodology

We compare **XAI methods** based on:

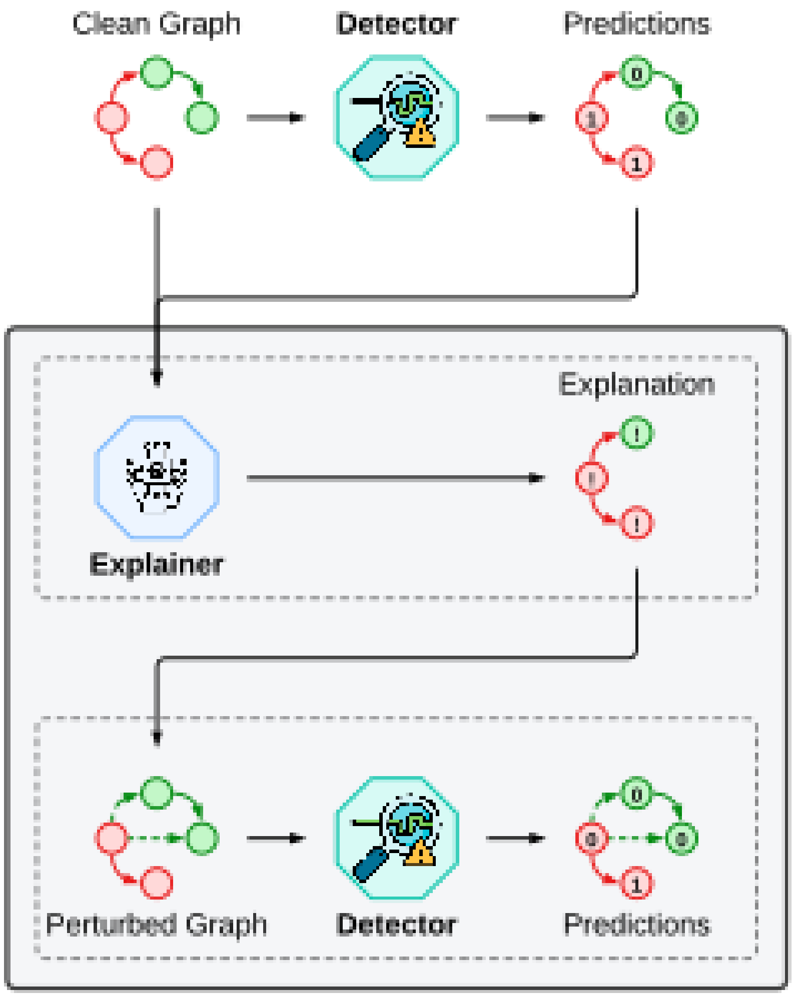
- Accuracy in identifying key components within the graph structure
- Effectiveness in evading GNN detectors through adversarial attacks

Explaining phase

- Explanations are extracted to identify structural vulnerabilities, offering insights into the GNN model

Evaluation phase

- Explanations are injected into the graph, modifying the resultant network topology



Explaining

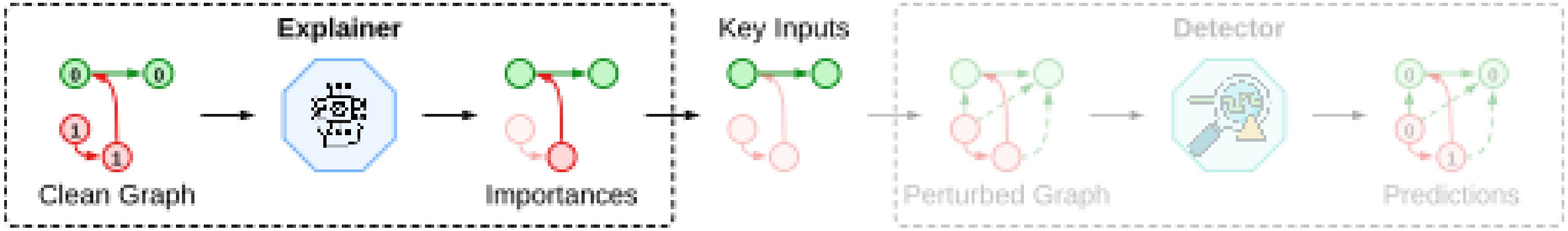
We apply explainers to the graph to extract key components

Each explainer generates an **explanation mask**

- Explanatory subgraph whose elements have relevance values

Flows are ranked to identify the most important **legitimate records**

- Network communications that contribute most to detector predictions



Evaluation

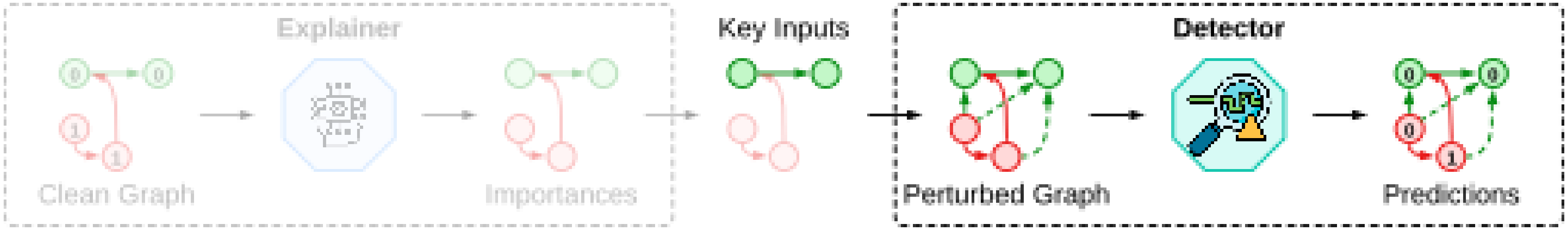
We assess explanations by measuring how well they evade detection

Attackers alter graph patterns carrying out **structural attacks**

- Important legitimate communications are injected into the graph

Manipulated graph is fed to the detector leading to **misclassifications**

- Most effective explanations are those that enable most successful attacks



Case Study

We consider two real-world **traffic datasets**:

- *CTU-13:* enterprise network traces that contain botnet traffic
- *ToN-IoT:* IoT network traces that include attack traffic

We evaluate thirteen attack-specific **cyber detectors**:

- *GraphSAGE:* inductive GNN model designed for large-scale graphs

We test five post-hoc **explainability methods**:

- *Dummy Explainer:* assigns random scores to graph components
- *Integrated Gradients:* calculates explanations by integrating gradients
- *Saliency:* computes importances by measuring gradients
- *GNNExplainer:* defines subgraphs by estimating the mutual information
- *GraphMask:* generates subgraphs by iteratively removing edges

Detectors Performance

We evaluate GNN detectors on clean network graphs

- Graphs are built from test sets and fed to GraphSAGE instances

CTU-13	
Botnet	***F1-score***
Neris	0.846
Rbot	0.989
Virut	0.943
Menti	0.953
Murlo	0.946
Average	**0.935**

ToN-IoT	
Attack	***F1-score***
Bkdr	0.999
DDoS	0.995
DoS	0.994
Inj	0.991
Pswd	0.998
Rans	0.995
Scan	0.994
XSS	0.995
Average	**0.995**

GraphSAGE is a solid target for structural attacks

Explainers Performance (1)

We evaluate GNN explainers on manipulated network graphs

- Graphs are perturbed with relevant nodes and submitted to GraphSAGE instances

Dataset	Threat	*DE*	*IG*	*SA*	*GE*	*GM*
CTU-13	*Neris*	0.071	**0.104**	0.058	0.058	0.056
	Virut	0.101	**0.143**	0.103	0.059	0.103
	Menti	0.457	**0.728**	0.538	0.402	0.529
	Murlo	0.668	**0.900**	0.798	0.675	0.478
ToN-IoT	*Bkdr*	0.151	**0.203**	0.035	0.135	0.164
	Inj	0.145	**0.226**	0.075	0.123	0.208
	Pswd	0.104	**0.223**	0.026	0.103	0.173
	Rans	0.183	**0.247**	0.233	0.186	0.167
	Scan	0.114	**0.184**	0.097	0.112	0.101
	XSS	0.190	**0.274**	0.187	0.177	0.222

IG allows more effective attacks than those exploiting random samples

Explainers Performance (2)

We evaluate GNN explainers on manipulated network graphs

- Graphs are perturbed with relevant nodes and submitted to GraphSAGE instances

Dataset	Threat	*DE*	*IG*	*SA*	*GE*	*GM*
CTU-13	*Rbot*	0.181	0.233	**0.237**	0.169	0.194

SA identifies netflow features that rarely influence GNN model predictions

Explainers Performance (3)

We evaluate GNN explainers on manipulated network graphs

- Graphs are perturbed with relevant nodes and submitted to GraphSAGE instances

Dataset	Threat	*DE*	*IG*	*SA*	*GE*	*GM*
ToN-IoT	*DDoS*	0.108	0.187	0.189	0.102	**0.273**
	DoS	0.012	0.018	0.011	0.012	**0.023**

GM exposes structural vulnerabilities when dealing with highly structured attacks

Conclusions

Lack of standardized evaluation approaches for XAI in GNN-based NIDS

We propose an evaluation framework tailored to real-world scenarios
- Explainability method defines an explanatory graph highlighting relevant flows
- Explainer performance depends on the severity of explanation-guided attacks

We test our methodology through a case study involving different explainers
- IG consistently generates explanations leading to targeted attacks
- Other explanations are not representative of topological vulnerabilities

Future research should validate our results across different settings and strategies