Hardening Machine Learning based Network Intrusion Detection Systems with synthetic netflows

Andrea Venturi, Dimitri Galli, Dario Stabili, Mirco Marchetti

andrea.venturi@unimore.it, dimitri.galli@unimore.it, dario.stabili@unibo.it, mirco.marchetti@unimore.it

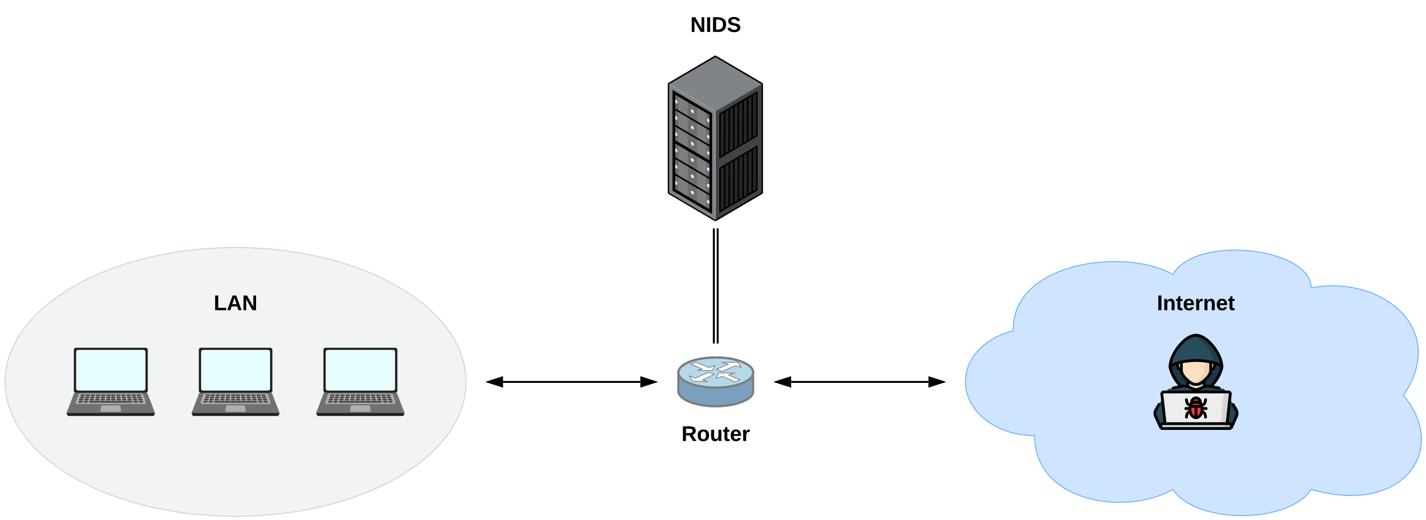
ITASEC24

April 8-12, 2024 - Salerno, Italy

ML-based NIDS

- Modern **NIDS** adopt **ML** algorithms to automate detection processes
  - Cyber detectors employ ML techniques to learn malicious patterns from network traffic

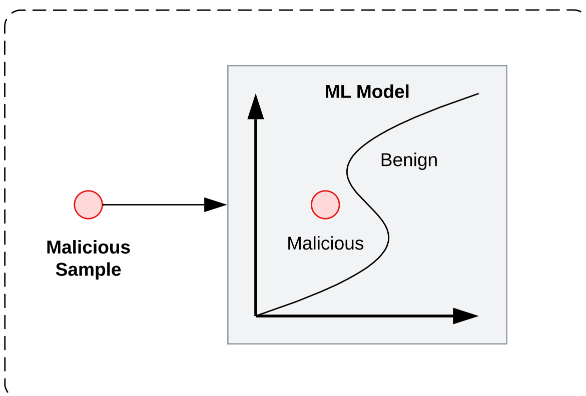
- Majority of ML-based NIDS analyze **data** represented as **netflows**
  - Netflows summarize the characteristics of the communication between two hosts in a network

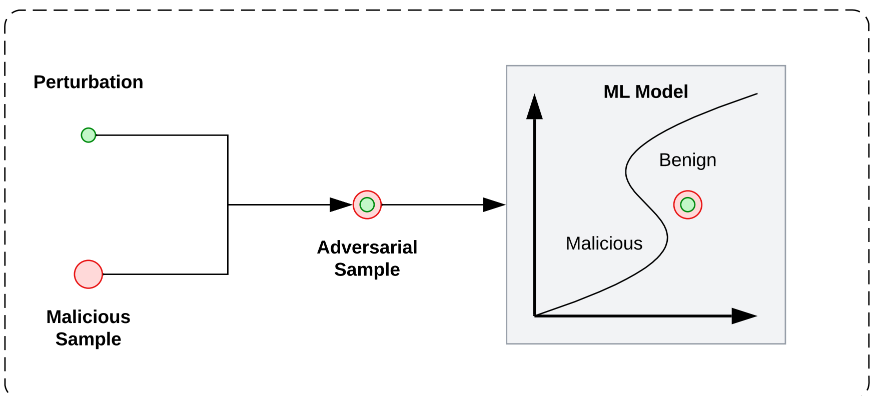


Adversarial attacks

- ML models are particularly vulnerable to **adversarial attacks**
  - Attackers can perturb malicious samples to trick classifiers into producing a misclassification

- ML-based NIDS can be hardened through **adversarial training**
  - Perturbed samples are included into the training set of vulnerable detectors



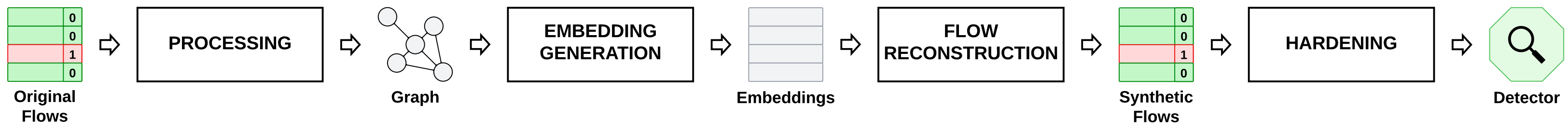


Methodology

- Adversarial training is affected by **limitations**:
    - Many adversarial examples must be generated
    - Perturbed records must represent all possible attacks

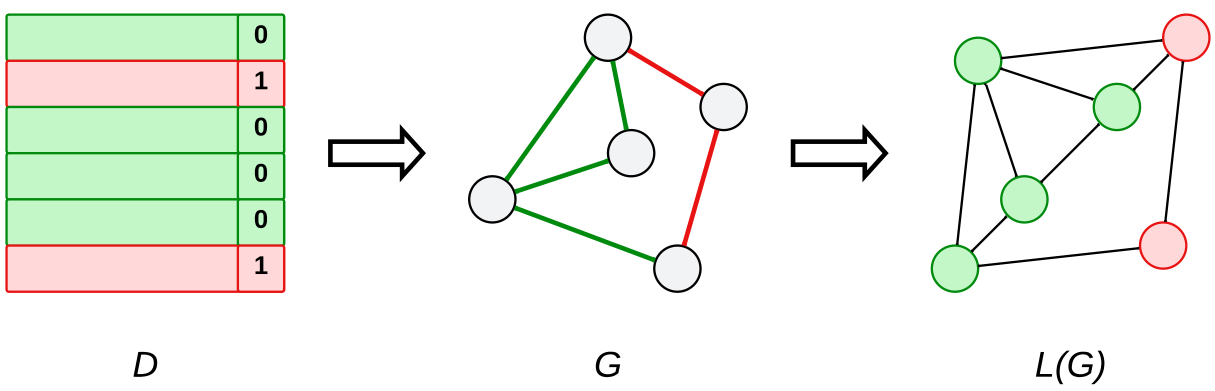
- We overcome these pitfalls by presenting an approach for generating synthetic traffic
    - GNN embeds the graph into a latent space considering the network topology
    - ML models reconstruct independently the initial samples introducing limited noise

- Generated netflows show small alterations in the attributes
    - Perturbed records are proper to be used as evasive samples
    - Reconstructed samples are injected into the dataset to get hardened systems



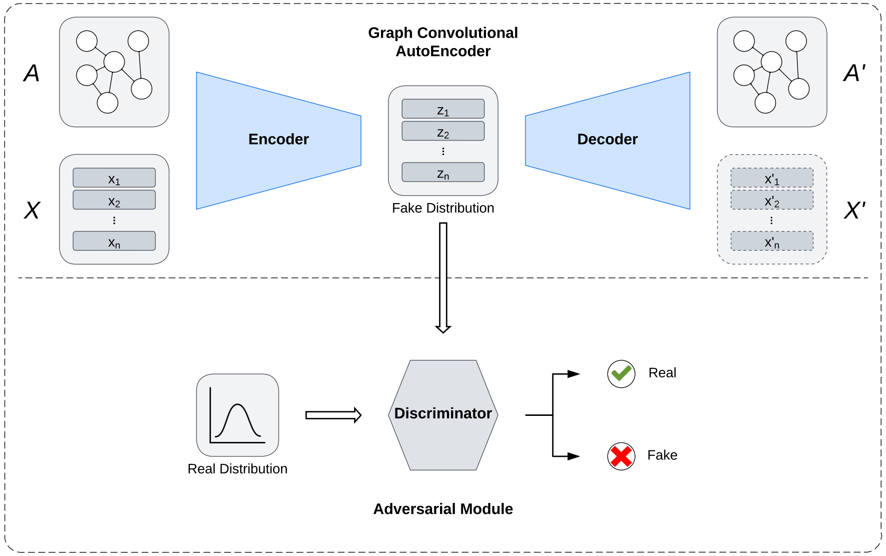
Processing

- Captures are converted into a specific graph structure
- Flows are associated to edges while endpoints in netflows correspond to graph nodes
  - This graph is referred to as *G*
- Linearization procedure transforms all edges into new vertices
  - This graph is referred to as *L(G)*



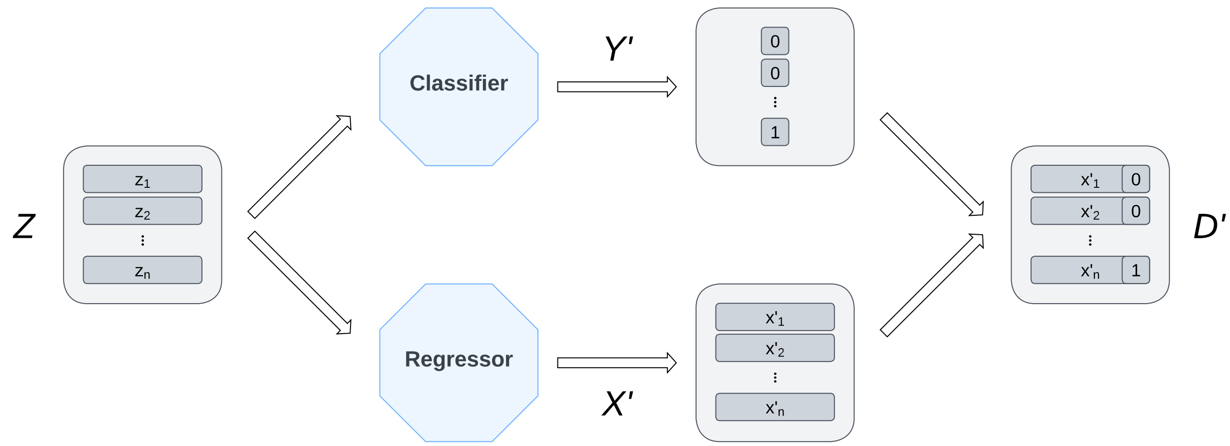
Embedding generation

- ARGA encodes both semantic and topological features of the graph into a latent space
- Encoder exploits the features and the topology for generating the latent variables
  - Encoder takes *A* and *X* as input to produce *Z*
- Decoder reconstructs the topology using the encoded latent variables
  - Decoder takes *Z* as input to reproduce *A*



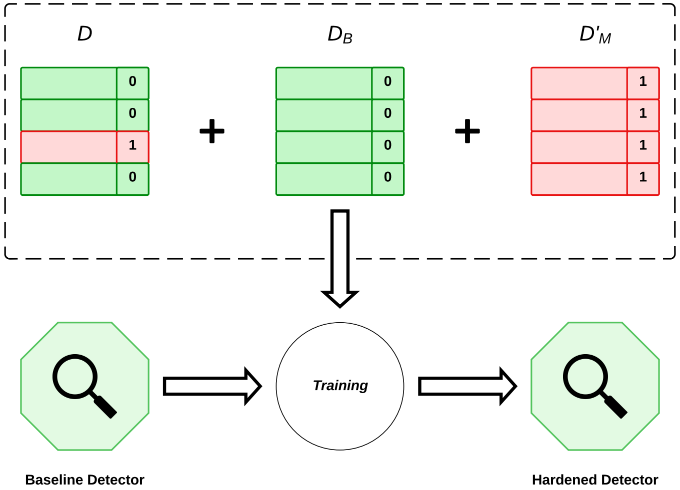
Flow reconstruction

- Dedicated ML models predict features and labels of the netflows to be reconstructed
- Classifier produces the labels of the new records
  - Classifier is trained over *Z* and *Y* to generate *Y*'
- Regressor reproduces the features of the original samples
  - Regressor is trained over *Z* and *X* to generate *X*'



Hardening

- Adversarial flows are injected into the dataset to harden ML-based NIDS
- Generated malicious netflows have perturbations due to error in reconstruction
    - D'_M includes adversarial attack candidates that could evade the models
- Generated malicious samples are inserted into the training set
    - D, D_B, and D'_M are merged to provide an augmented dataset



Case study

- Dataset employed in the evaluation is **ToN-IoT**
  - Medium-scale network traffic mixed with attack traffic
- Detectors implemented with **RF**
  - Binary classifiers tailored to detect attack-specific variants
- Embedding generation phase is based on **ARVGA_GD**
  - Encoder to generate latent variables + decoder to reconstruct topology
- Flow reconstruction stage is based on **RF**
  - Classifier to produce labels + Regressor to reproduce features

Standard evaluation

- Baseline and hardened classifiers are tested using flows from the original validation set
  - Goal is to validate whether considered detectors exhibit good performance in attack-free settings

Attack	***F1-score***	
	Baseline	**Hardened**
Backdoor	**1.000**	0.999
DDoS	0.981	**0.988**
DoS	0.995	**0.997**
Injection	0.991	**0.996**
Password	0.983	**0.987**
Ransomware	0.923	**0.939**
Scanning	0.998	**0.998**
XSS	0.978	**0.987**
average	0.981	**0.987**

- **Takeaways**:
  - Baseline detectors achieve performance scores in line with state-of-the-art
  - Hardened instances obtain higher performance scores than baseline systems

Adversarial evaluation

- Baseline and hardened classifiers are tested against evasive flows with increasing perturbation steps
  - Goal is to validate whether considered detectors exhibit good resilience toward adversarial attacks

Attack	*DR*	
	Baseline	**Hardened**
Backdoor	0.613	**0.747**
DDoS	0.781	**0.828**
DoS	0.465	**0.917**
Injection	0.887	**0.972**
Password	0.567	**0.624**
Ransomware	0.269	**0.587**
Scanning	0.727	**0.995**
XSS	0.925	**0.945**
average	0.654	**0.827**

- **Takeaways**:
  - Baseline detectors get insufficient performance scores
  - Hardened instances show more strength than baseline systems

Conclusions

- Adversarial attacks represent a serious threat to ML-based NIDS
    - Research is still at an early stage

- We propose an architecture based on ARGA and RF to automatically generate synthetic traffic
    - Generated records act as adversarial samples in adversarial training

- We apply our approach to a case study involving classifiers trained on a public real-world dataset
    - Results show the efficacy of our proposal in enhancing the robustness of the systems

References

- G. Apruzzese and M. Colajanni, "Evading Botnet Detectors Based on Flows and Random Forest with Adversarial Samples," in *2018 IEEE 17th International Symposium on Network Computing and Applications (NCA)*, Cambridge, MA, USA, 2018, pp. 1-8

- S. Pan, R. Hu, S. Fung, G. Long, J. Jiang, and C. Zhang, "Learning Graph Embedding With Adversarial Training Methods," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2475-2487, June 2020