RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs inria

FRANCE 2030 PROGRAMME DE RECHERCHE CYBERSÉCURITÉ

Discovering Critical Vulnerability Paths with RL Agents: Enhancing Generalization and Scalability

Franco Terranova - Ph.D. Student[1]

Under the supervision of Abdelkader Lahmadi[1] and Isabelle Chrisment[1]

SuperVIZ Meeting - Campus Cyber La Défense, Paris, France

11/03/2025

[1] Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France

anr agence nationale de la recherche

Introduction

- **Proactive methods**: valuable alternative to reactive methods aiming to anticipate attacks before they happen

Introduction

- **Proactive methods**: valuable alternative to reactive methods aiming to anticipate attacks before they happen
- Several technologies already exist, mainly for the generation of attack graphs (Fig. 1)
- **Drawbacks**:
    - Requires algorithms to run on top of the attack graphs
    - Regeneration when network changes
    - Requires full information of the network

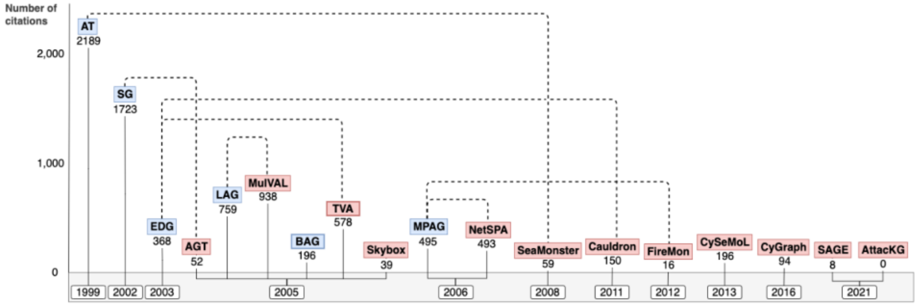
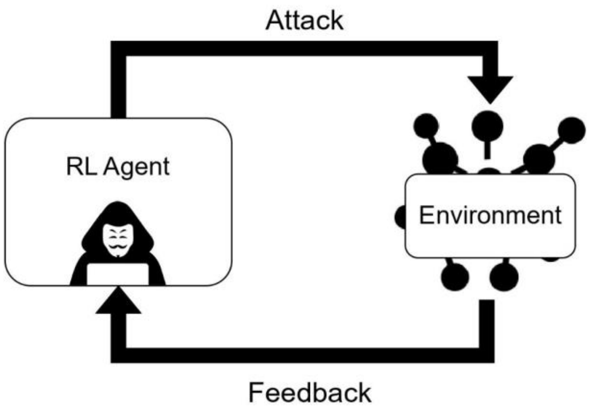


Fig.1: Evolution of attack graph generation methods (red) as well as attack graph representations (blue): publication year and number of citations [1]

[1] D. Tayouri, N. Baum, A. Shabtai, and R. Puzis, “A Survey of MulVAL Extensions and Their Attack Scenarios Coverage.” arXiv, Aug. 11, 2022. Accessed: May 29, 2024. [Online]. Available: http://arxiv.org/abs/2208.05750

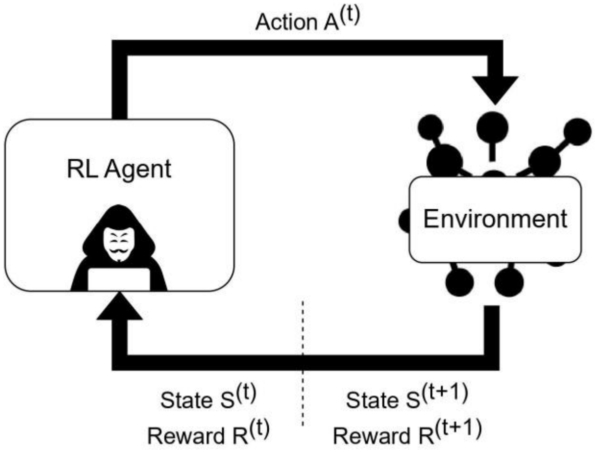
Reinforcement Learning

- **Machine Learning** (ML) solution to approximate the attacker strategy
- **Best candidate:** Reinforcement Learning (RL)



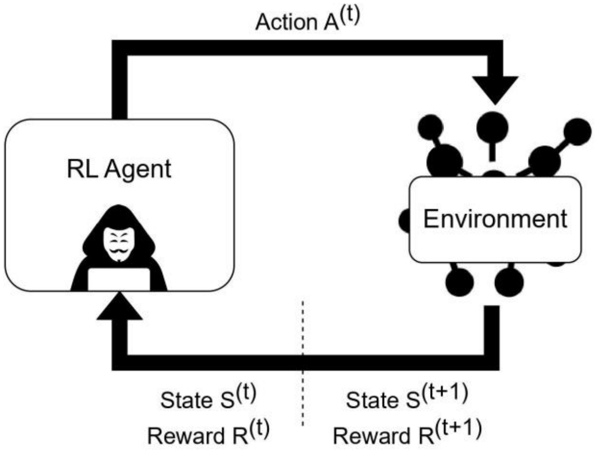
Reinforcement Learning

- **Machine Learning** (ML) solution to approximate the attacker strategy
- **Best candidate:** Reinforcement Learning (RL)
- **Learning** occurs via trial and error



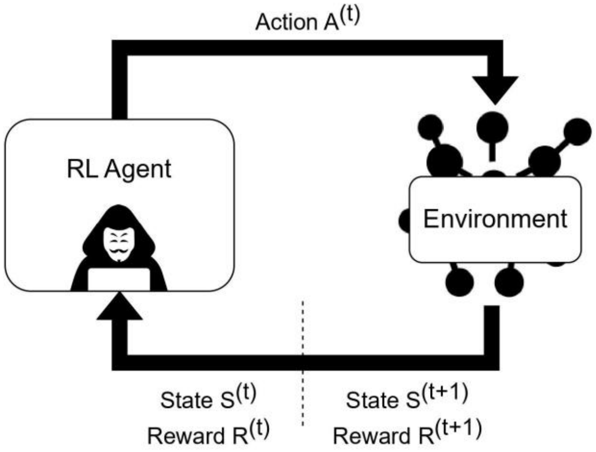
Reinforcement Learning

- **Machine Learning** (ML) solution to approximate the attacker strategy
- **Best candidate:** Reinforcement Learning (RL)
- **Learning** occurs via trial and error
- **Advantages**:
  - Agent works as prioritization algorithm itself
  - Environment can change during execution
  - Agent can work with partial observability

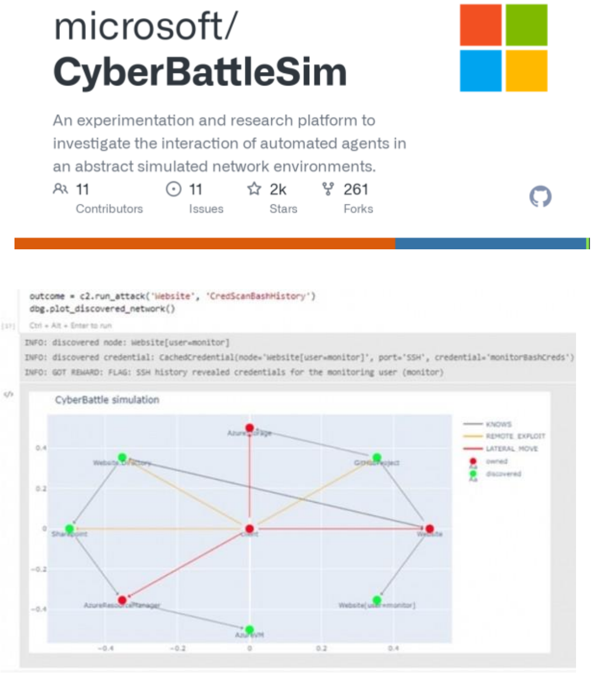


Reinforcement Learning

- **Machine Learning** (ML) solution to approximate the attacker strategy
- **Best candidate:** Reinforcement Learning (RL)
- **Learning** occurs via trial and error
- **Advantages**:
    - Agent works as prioritization algorithm itself
    - Environment can change during execution
    - Agent can work with partial observability
- **Deep RL:** neural networks to approximate the agent



RL Environment



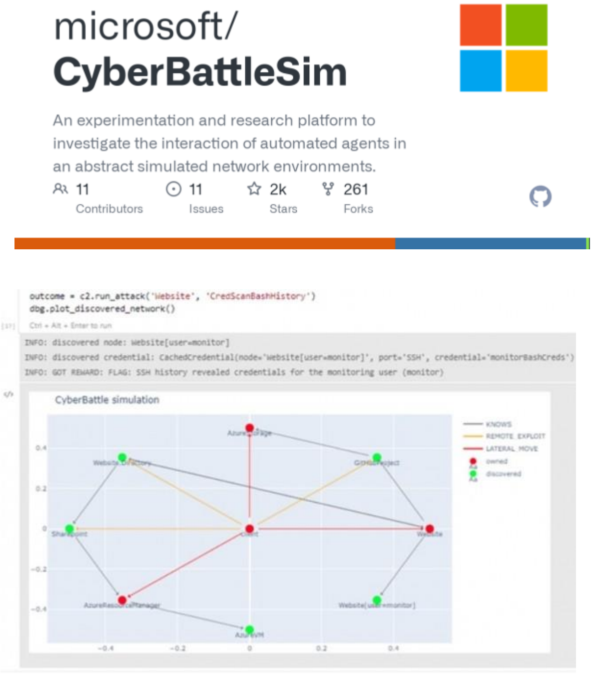
- Why **simulations**? Real-world training is costly
- **Target Environment:** Microsoft CyberBattleSim [2]

[2] Team., M.D.R.: Cyberbattlesim (2021), created by Seifert C., Betser M., Blum W., Bono J., Farris K., Goren E., Grana J., Holsheimer K., Marken B., Neil J., Nichols N., Parikh J., Wei H.

RL Environment

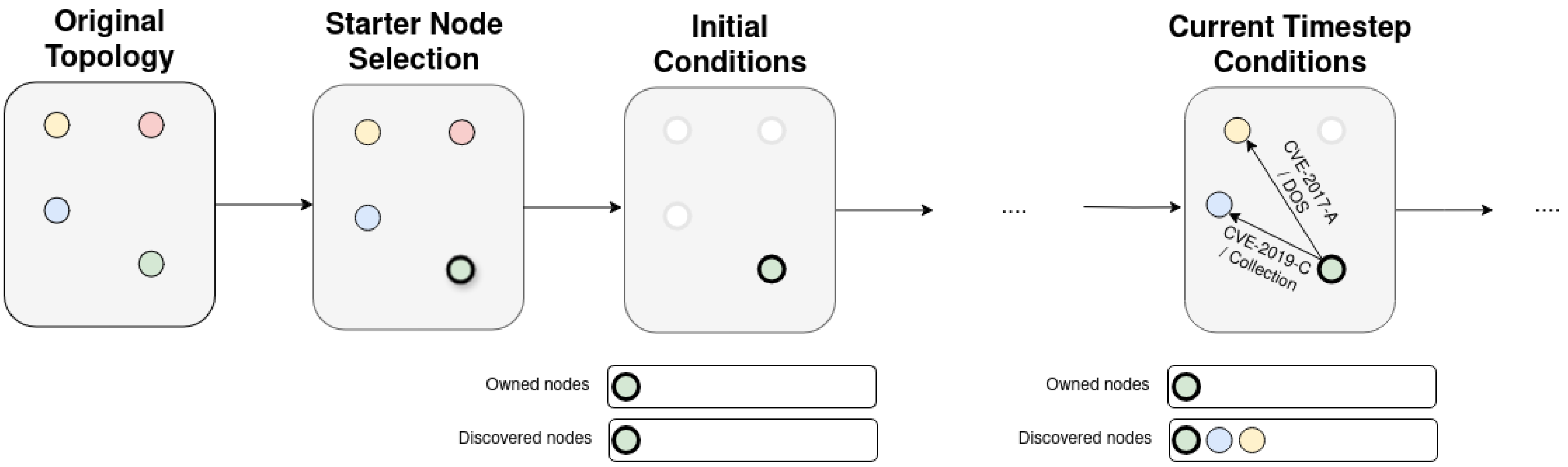
- Why **simulations**? Real-world training is costly
- **Target Environment:** Microsoft CyberBattleSim [2]

- **States**: Graph representation of the network
- **Actions**: *(Source node, Target node, Vulnerability)*
- **Reward**: Linear function of outcomes



[2] Team., M.D.R.: Cyberbattlesim (2021), created by Seifert C., Betser M., Blum W., Bono J., Farris K., Goren E., Grana J., Holsheimer K., Marken B., Neil J., Nichols N., Parikh J., Wei H.

CyberBattleSim Episode



Contributions

1. **More realistic simulation environment:** Close the sim-to-real gap
2. **Generalizable & scalable RL agents** for the task
    - Achieve independence from application graph and vulnerability set
    - Achieve invariance from their size and ordering
3. Improved RL **training & evaluation framework**

Goal: Learn from simulation and deploy in real-world scenarios

C1: Environment Realism

- **Environment realism** affects the patterns the agent learns
- Previous issues:
  - **Random scattering** of vulnerabilities in nodes and across the environment
  - **Vulnerability set** is typically fictitious and small
  - E.g. Original CyberBattleSim scenarios include <= 10 vulnerabilities each

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE CYBERSÉCURITÉ

C1: Scenario Generation

Shodan Report

`has_vuln:True cloud.provider:"Amazon"`

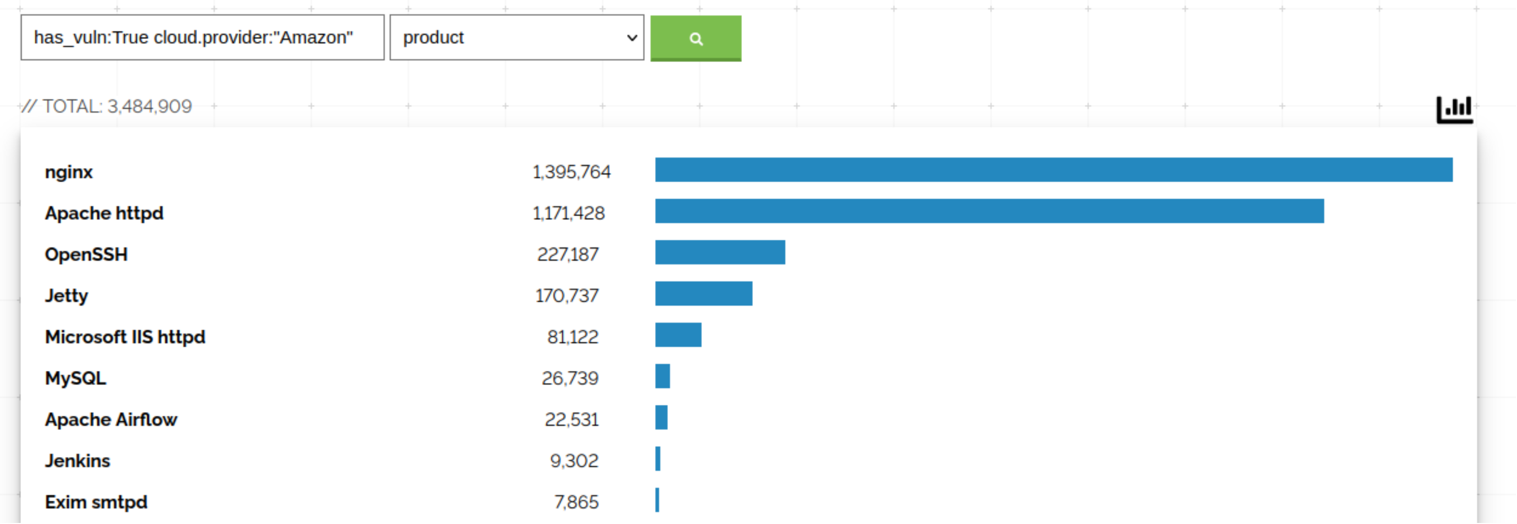
Total: 3,484,909

// GENERAL

Countries

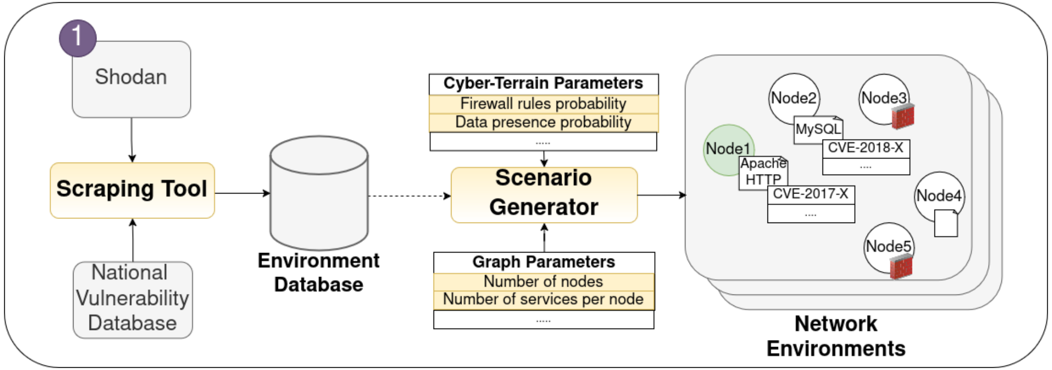
United States	1,538,944
Japan	322,876
India	278,947
Ireland	223,764
Germany	172,497

C1: Scenario Generation



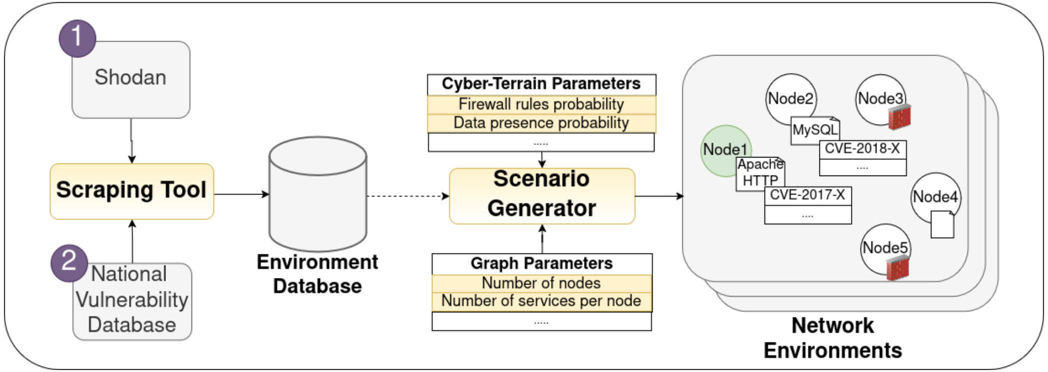
C1: Scenario Generation

1. **Shodan Scraping**:
   - Use customizable query to gather statistics
   - Determine service set and allocation frequency
2. Use **NVD** to gather **vulnerabilities**
3. **Scenario Generation:**
   - Use graph parameters, cyber-terrain parameters, and statistics for allocation



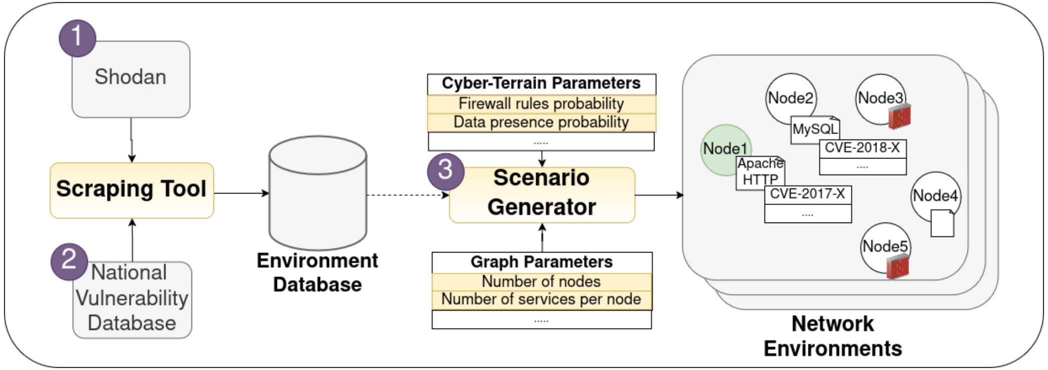
C1: Scenario Generation

1. **Shodan Scraping:**
   - Use customizable query to gather statistics
   - Determine service set and allocation frequency
2. Use **NVD** to gather **vulnerabilities**
3. **Scenario Generation:**
   - Use graph parameters, cyber-terrain parameters, and statistics for allocation



C1: Scenario Generation

1. **Shodan Scraping**:
   - Use customizable query to gather statistics
   - Determine service set and allocation frequency
2. Use **NVD** to gather **vulnerabilities**
3. **Scenario Generation:**
   - Use graph parameters, cyber-terrain parameters, and statistics for allocation



C1: Automated Outcome Mapping

- What limited **vulnerability** set **integration**?
  - Need to define outcomes to simulate for each vulnerability of the simulation environment

C1: Automated Outcome Mapping

- What limited **vulnerability** set **integration**?
  - Need to define outcomes to simulate for each vulnerability of the simulation environment

- Automate outcome mapping using multi-label classifier:

$$f_{\text{multi-label}}(vulnerability\ description) \rightarrow Outcomes$$

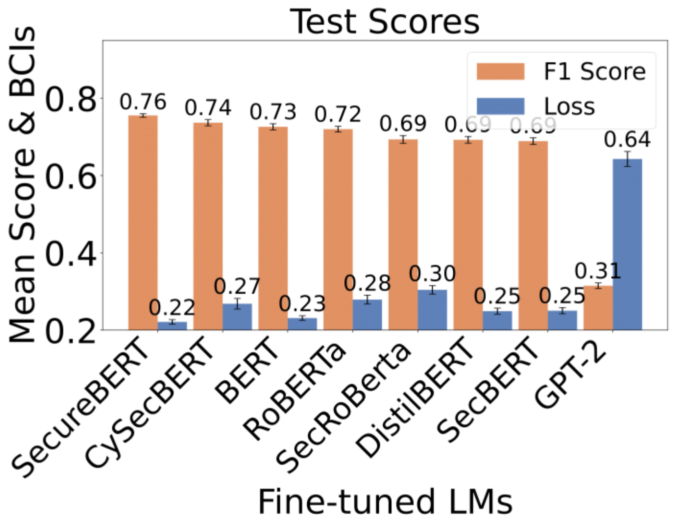
- **Fine-tuned Language Models** (LMs) to this aim

C1: Automated Outcome Mapping

- What limited **vulnerability** set **integration**?
  - Need to define outcomes to simulate for each vulnerability of the simulation environment
- Automate outcome mapping using multi-label classifier:

$$f_{\text{multi-label}}(\textit{vulnerability description}) \rightarrow \textit{Outcomes}$$

- **Fine-tuned Language Models** (LMs) to this aim



C2: Agent Reformulation

- **General Agent Formulation:**

$$\begin{aligned} \pi_{\text{general}}: &\quad \textit{graph encoding} \rightarrow \textit{(source, target, vulnerability)} \\ \text{such that} &\quad \max_{\pi} \mathbb{E}_{\tau \sim \pi} \left[\textit{Impact}(\text{Path } \tau \mid \text{Goal, InitialConditions}) \right] \end{aligned}$$

C2: Agent Reformulation

- **General Agent Formulation:**

$$
\begin{aligned}
\pi_{\text{general}}: &\quad \textit{graph encoding} \rightarrow \textit{(source, target, vulnerability)} \\
\text{such that} &\quad \max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\textit{Impact}(\text{Path } \tau \mid \text{Goal, InitialConditions})\right]
\end{aligned}
$$

- **Key Principle:** Generalizable observation & action spaces → Generalizable agents
- **Goal:** Learn mappings in spaces independent from graph and vulnerability set

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE CYBERSÉCURITÉ

C2: Agent Reformulation

Global Agent

$$\pi_{\text{global}} : \mathbb{R}^{N \times f} \rightarrow \mathbb{R}^{N \times N \times |V|}$$

- Dependence on graph structure
- Dependence on the vulnerability set V
- Global optimization

C2: Agent Reformulation

Global Agent

$$\pi_{\text{global}} : \mathbb{R}^{N \times f} \rightarrow \mathbb{R}^{N \times N \times |V|}$$

- Dependence on graph structure
- Dependence on the vulnerability set V
- Global optimization

Local Agent [3]

$$\pi_{\text{local}} : \mathbb{R}^{2 \times f + g} \rightarrow \mathbb{R}^{|V| + |S|}$$

- No dependence on graph structure
- Dependence on the vulnerability set V
- Local optimization

[3] Franco Terranova, Abdelkader Lahmadi, and Isabelle Chrisment. 2024. Leveraging Deep Reinforcement Learning for Cyber-Attack Paths Prediction: Formulation, Generalization, and Evaluation. In Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '24). Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3678890.3678902

C2: Agent Reformulation

Global Agent

$$\pi_{\text{global}} : \mathbb{R}^{N \times f} \to \mathbb{R}^{N \times N \times |V|}$$

- Dependence on graph structure
- Dependence on the vulnerability set V
- Global optimization

Local Agent [2]

$$\pi_{\text{local}} : \mathbb{R}^{2 \times f + g} \to \mathbb{R}^{|V| + |S|}$$

- No dependence on graph structure
- Dependence on the vulnerability set V
- Local optimization

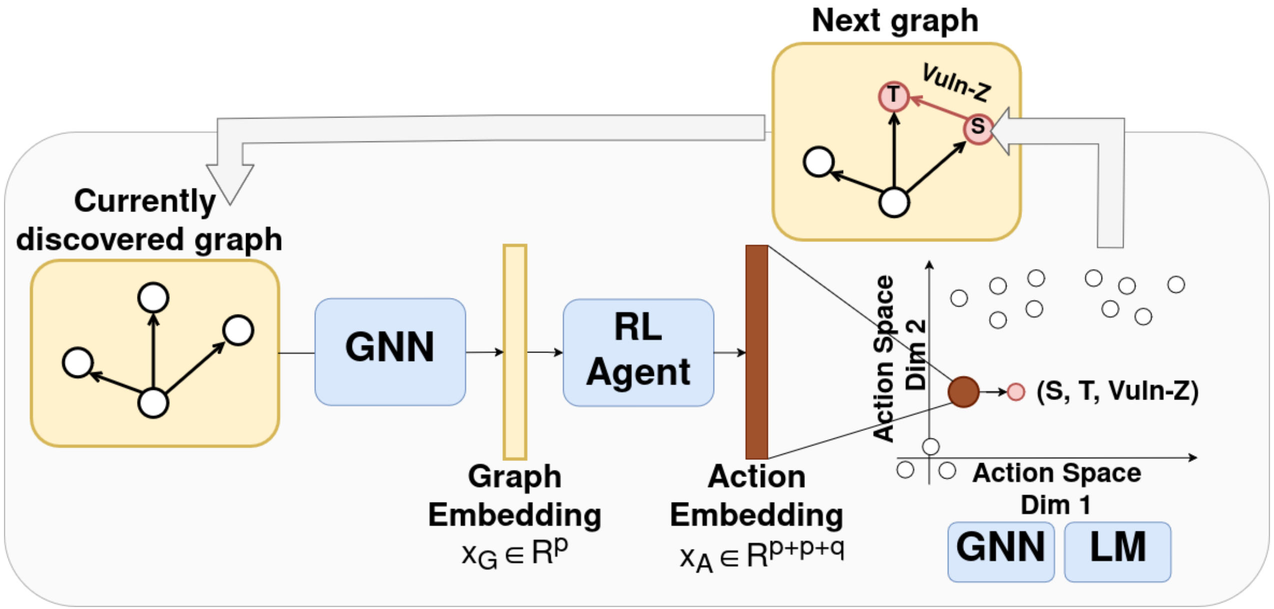
Continuous Agent

$$\pi_{continuous} : \quad \mathbb{R}^{p} \to \mathbb{R}^{p+p+q}$$

- No dependence on graph structure
- No dependence on the vulnerability set V
- Global optimization

[2] Franco Terranova, Abdelkader Lahmadi, and Isabelle Chrisment. 2024. Leveraging Deep Reinforcement Learning for Cyber-Attack Paths Prediction: Formulation, Generalization, and Evaluation. In Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '24). Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3678890.3678902

C2: Agent Reformulation



Contributions

1. **More realistic simulation environment:** Close the sim-to-real gap
2. **Generalizable & scalable RL agents** for the task
   - Achieve independence from application graph and vulnerability set
   - Achieve invariance from their size and ordering
3. Improved RL **training & evaluation framework**
   - **Domain randomization:** training & evaluation on large set of different scenarios
   - **Starter node randomization**
   - **Complexity** scenario set **splitting** in training, validation, and test sets

Experimental Benchmark

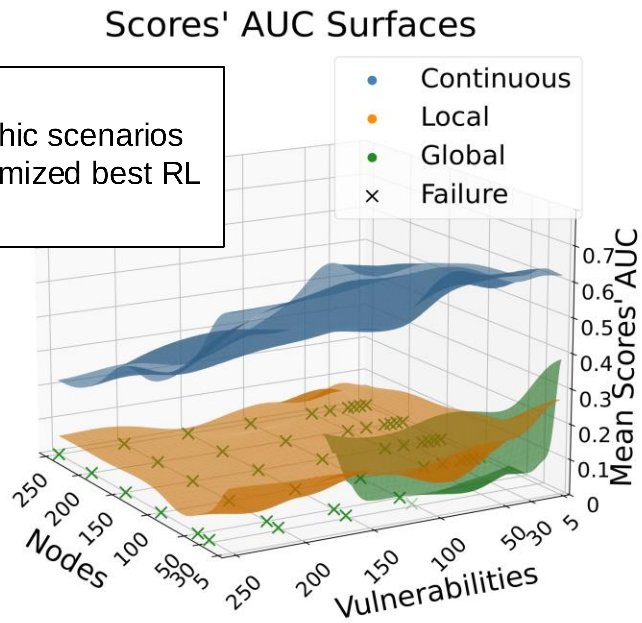
- Parameterized **Reward Function**:

$$R(o,a) = B(o,a,Goal) - \sum_{i=1}^{|Penalties|} P_i(o,a,Goal)$$

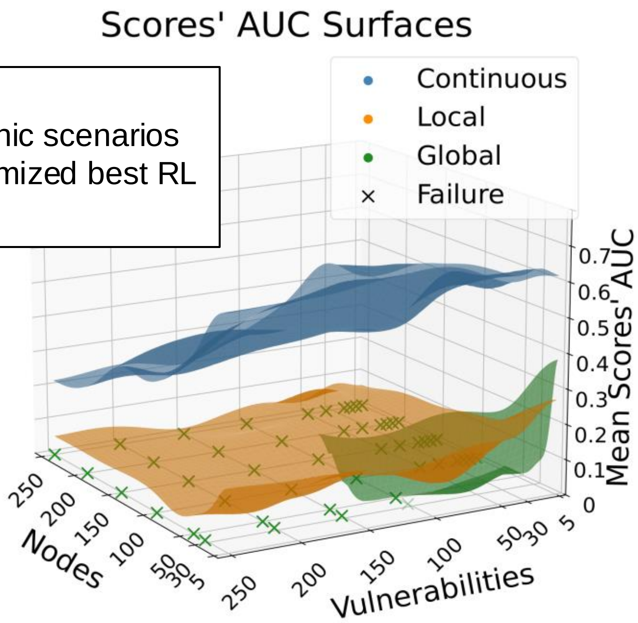
- **Experimental Benchmark:**
    - **Threat Models/Goals**: Control, Discovery, Disruption
    - **Environment database:** 172 service versions, 829 unique vulnerabilities
    - **RL Libraries:** StableBaselines3 algorithms + RLiable for evaluation

Scalability

Settings:
- 300 syntethic scenarios
- Hyper-optimized best RL algorithm



Scalability

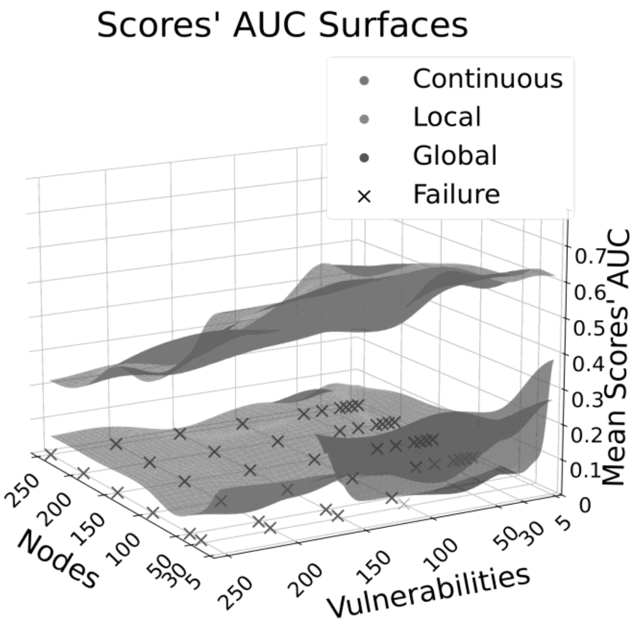


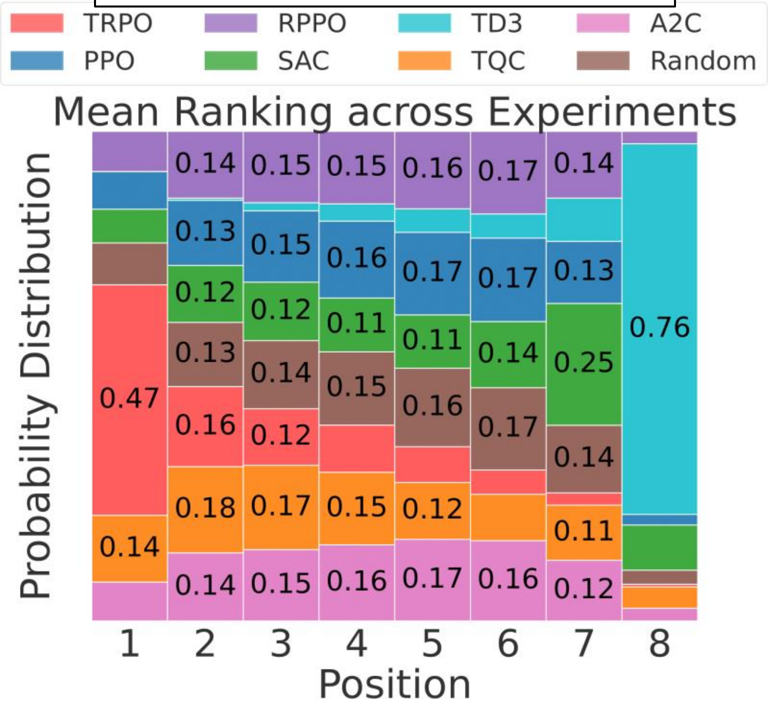
Settings:
- 300 syntethic scenarios
- Hyper-optimized best RL algorithm

Results:
- 7.4x improvement wrt local approach
- 11.2x improvement wrt global approach

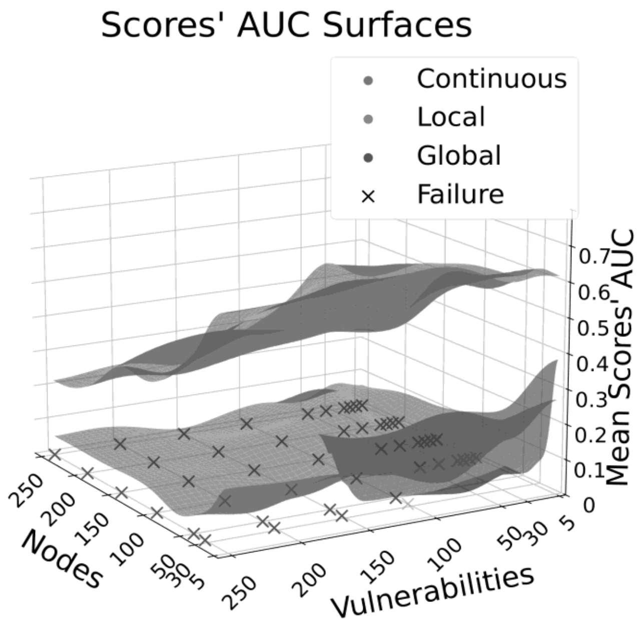
Generalization

Settings:
- 500 syntethic scenarios
- Hyper-optimized version of each RL algorithm

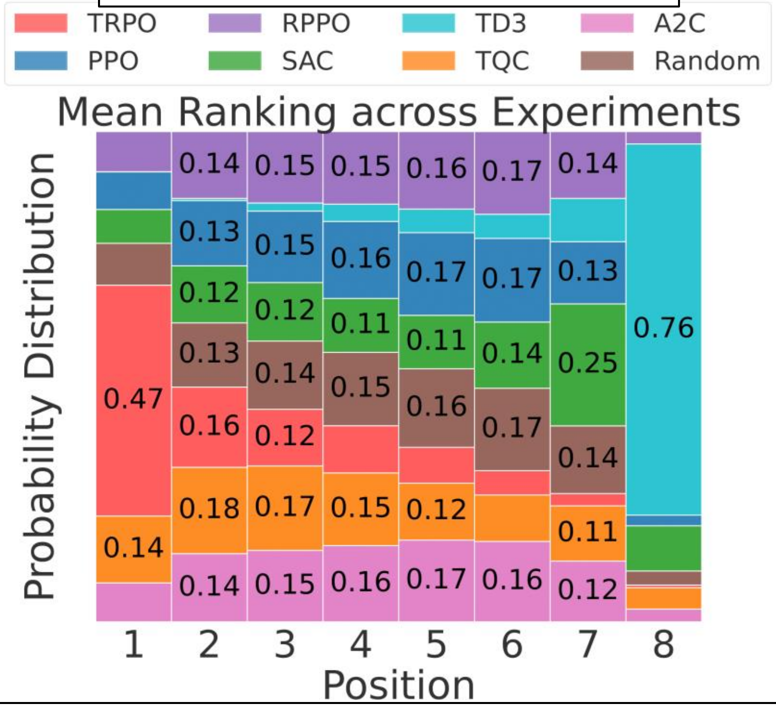




Generalization



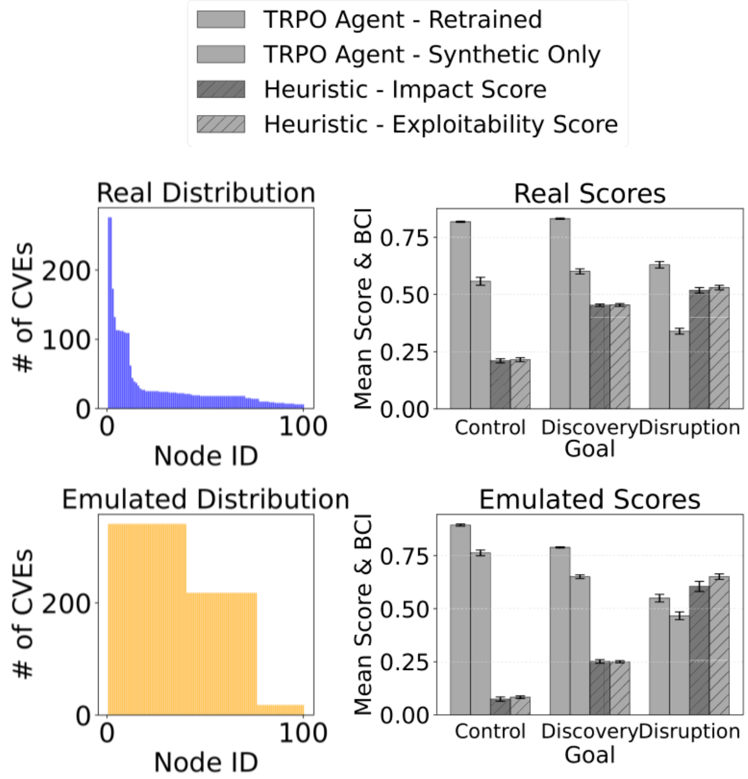
Settings:
- 500 syntethic scenarios
- Hyper-optimized version of each RL algorithm



Results:
- TRPO as the outperforming algorithm for generalization
- 89% test-to-training generalization ratio

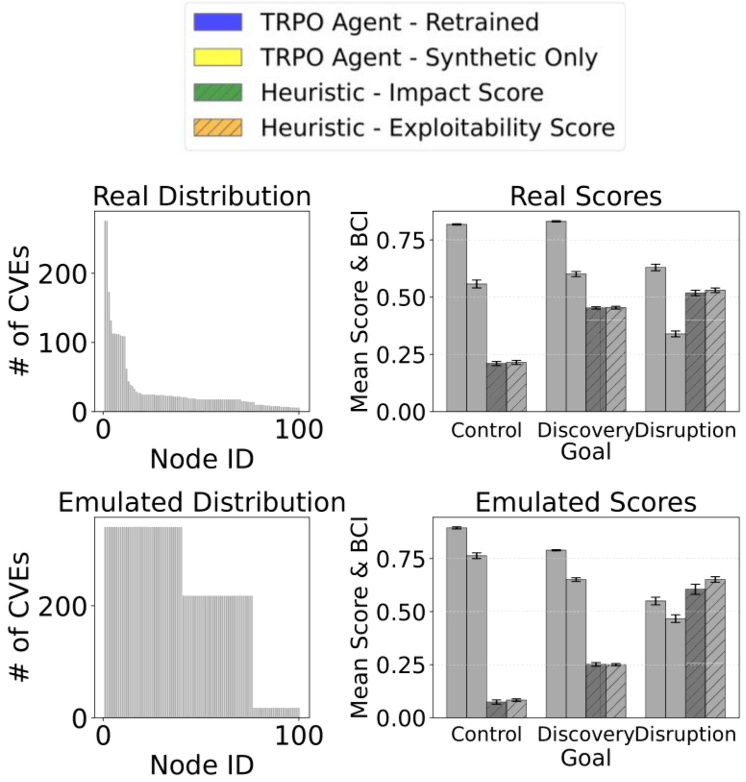
Deployment

- **Scenario set:** scenarios built from real-world and emulated-scan



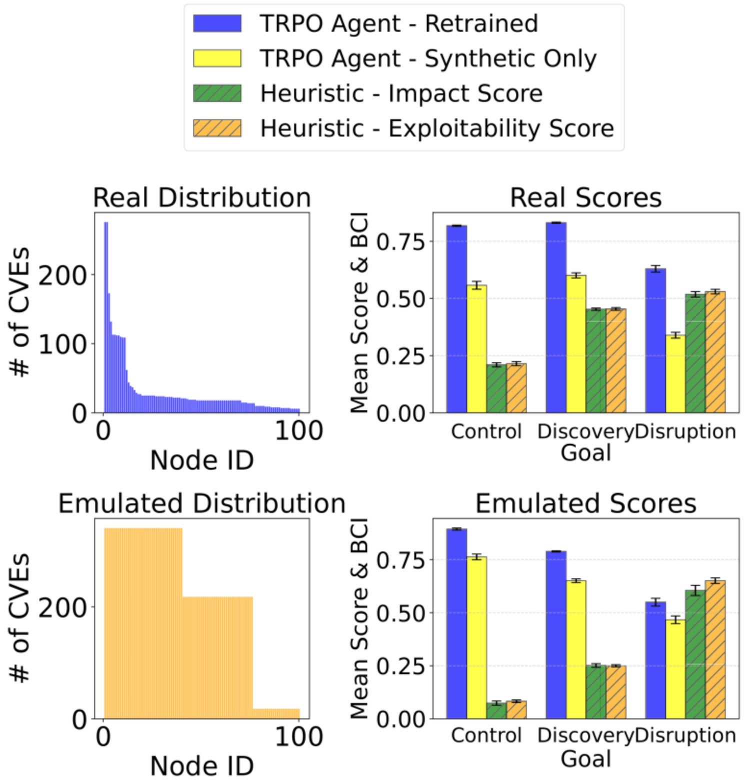
Deployment

- **Scenario set:** scenarios built from real-world and emulated-scan
- TRPO algorithm retrained vs only synthetic
- Heuristics maximizing Impact and Exploitability Scores



Deployment

- **Scenario set:** scenarios built from real-world and emulated-scan
- TRPO algorithm retrained vs only synthetic
- Heuristics maximizing Impact and Exploitability Scores



- **Results:**
  - Outperforming heuristics on control and discovery games
  - 75% synthetic-to-retrained agent score

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE CYBERSÉCURITÉ

Conclusions

- **Environment Realism:** Data Scraping, Scenario Generation, Automated Outcome Mapping
- **Agent Reformulation:** Learning in Continuous Invariant Spaces
    - **Improvements:** Scalability (avg. 9.3×) and Generalization (avg. 89%)
- **Training and Evaluation Pipeline** for more generalizable agents

- **Future Work**
    - **Competitive Multi-Agent RL:** Secure Virtual Machine Placement (in collaboration with University of Waterloo)
    - **Upcoming Release: C-CyberBattleSim** on GitHub

Secure Virtual Machine Placement with MARL

- **Context:** Virtualized networked system with a starting allocation of Virtual Machines (VMs) on top of Physical Machines (PMs)

Secure Virtual Machine Placement with MARL

- **Context:** Virtualized networked system with a starting allocation of Virtual Machines (VMs) on top of Physical Machines (PMs)
- **Attacker:** Find vulnerability paths inside the current allocation
- **Defender:** Reconfigures the allocation to decrease the paths and trade-off with other performance objectives
    - **Action Space:** (VM, PM, Isolation Level)

Secure Virtual Machine Placement with MARL

- **Context:** Virtualized networked system with a starting allocation of Virtual Machines (VMs) on top of Physical Machines (PMs)
- **Attacker:** Find vulnerability paths inside the current allocation
- **Defender:** Reconfigures the allocation to decrease the paths and trade-off with other performance objectives
    - **Action Space:** (VM, PM, Isolation Level)

- Turn-Based **Security Stackelberg Game**(TBSSG) with **competitive** agents
- At convergence, the defender agent will aim for a **reconfiguration strategy** of the environment

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

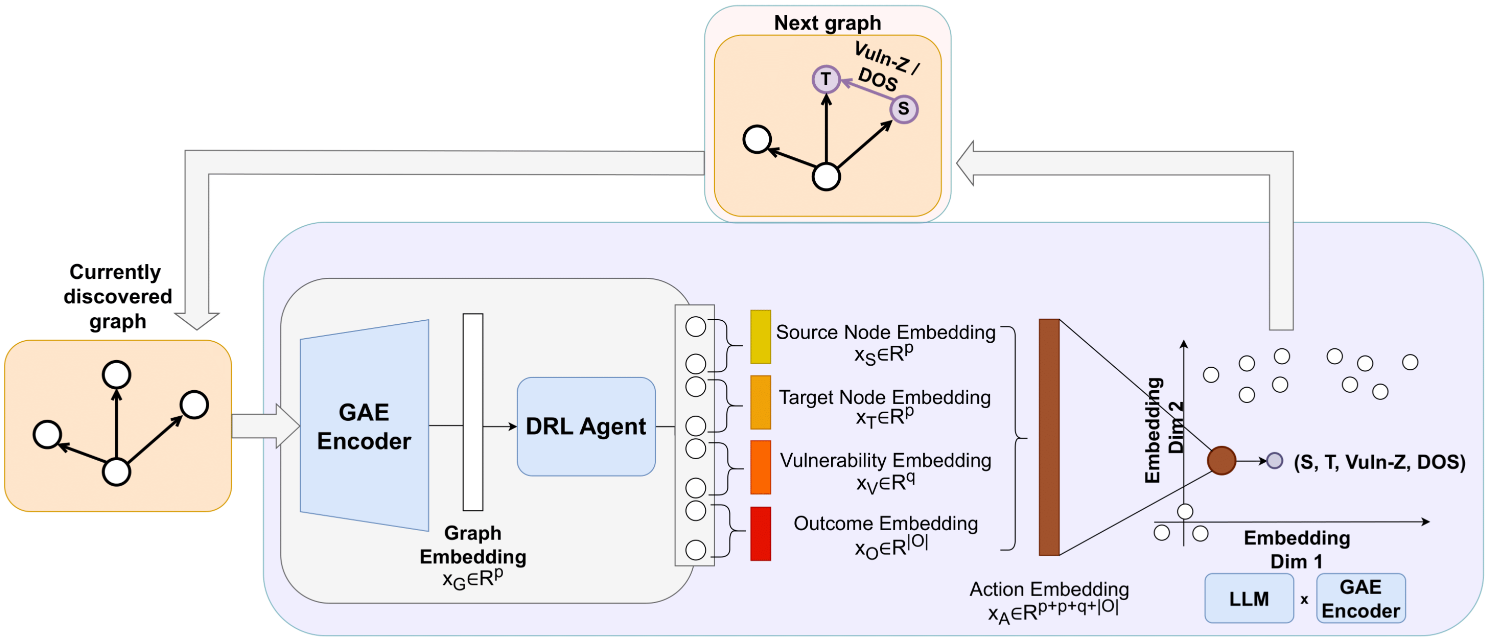
FRANCE 2030
PROGRAMME
DE RECHERCHE
CYBERSÉCURITÉ

Q&A Session

anr agence nationale de la recherche

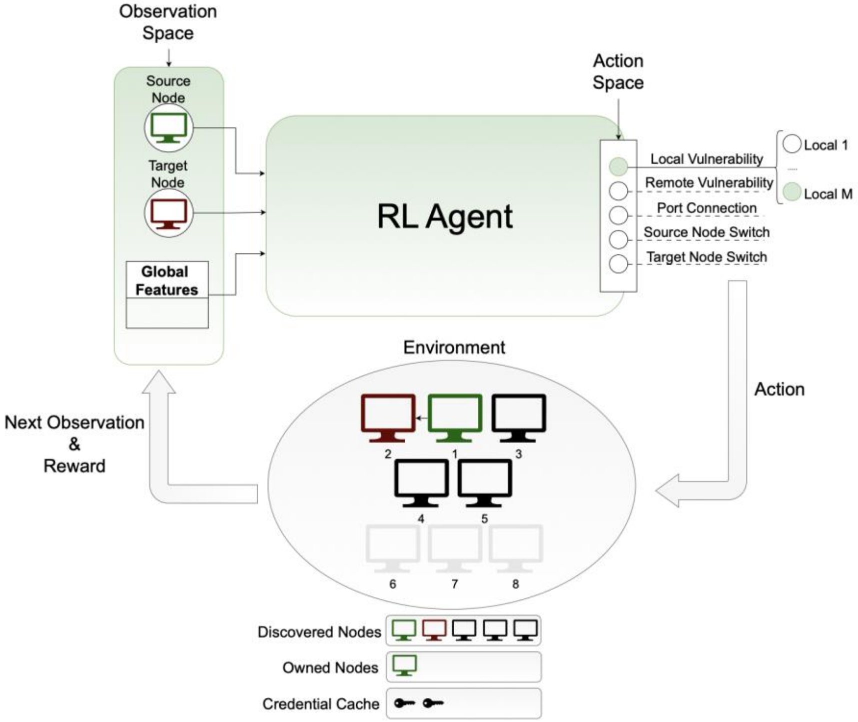
Backup

Pipeline



Next graph
Vuln-Z / DOS
T
S
Currently discovered graph
GAE Encoder
Graph Embedding
$x_G \in R^p$
DRL Agent
Source Node Embedding
$x_S \in R^p$
Target Node Embedding
$x_T \in R^p$
Vulnerability Embedding
$x_V \in R^q$
Outcome Embedding
$x_O \in R^{|O|}$
Action Embedding
$x_A \in R^{p+p+q+|O|}$
Embedding Dim 2
Embedding Dim 1
(S, T, Vuln-Z, DOS)
LLM
x
GAE Encoder

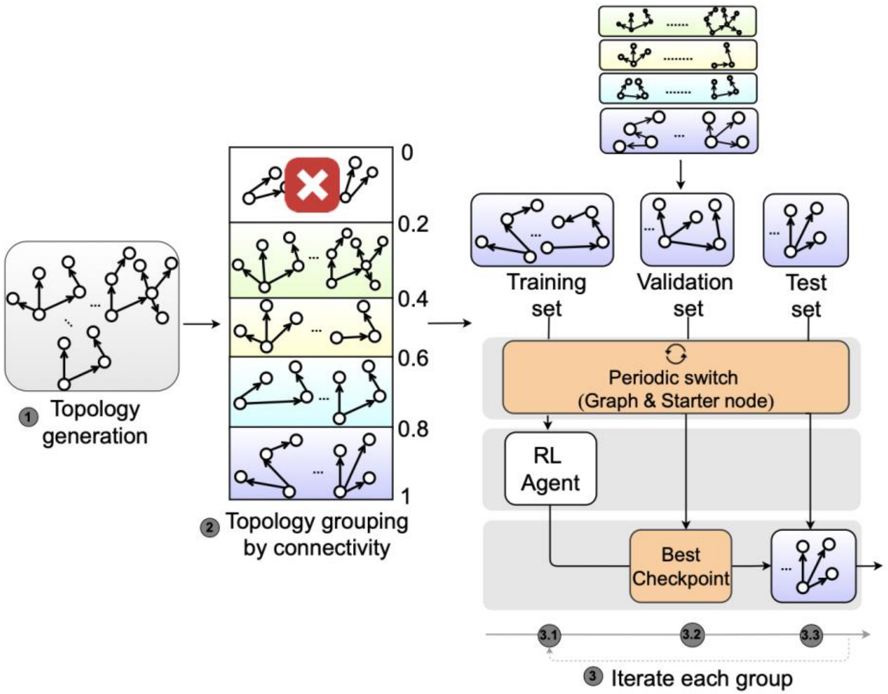
Local Agent


Observation Space
Source Node
Target Node
Global Features
RL Agent
Action Space
Local Vulnerability
Remote Vulnerability
Port Connection
Source Node Switch
Target Node Switch
Local 1
Local M
Action
Environment
2
1
3
4
5
6
7
8
Next Observation & Reward
Discovered Nodes
Owned Nodes
Credential Cache

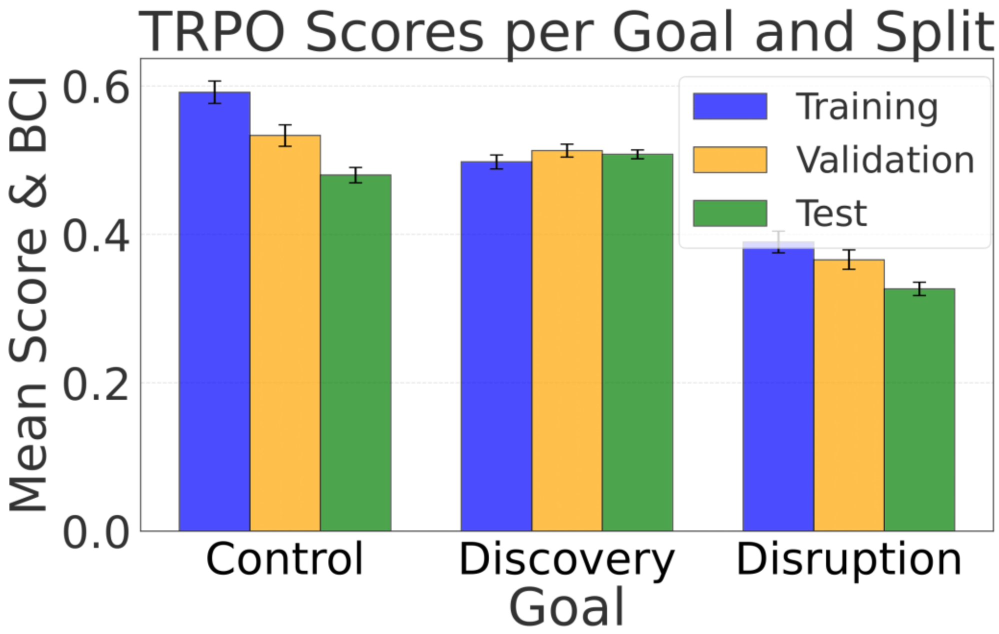
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE CYBERSÉCURITÉ

Scenario Generation Pipeline



TRPO Scores per Split and Goal



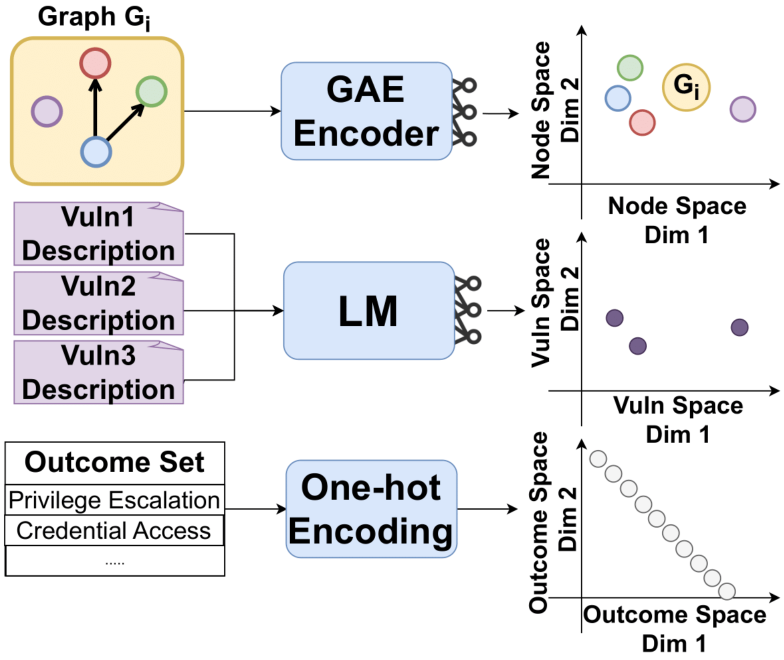
Embedding spaces

$$f_{\mathrm{GAE}}(\textit{feature vector} \in \mathbb{R}^f, \mathcal{G}) \rightarrow \mathrm{embedding}_{\mathrm{node}} \in \mathbb{R}^p$$

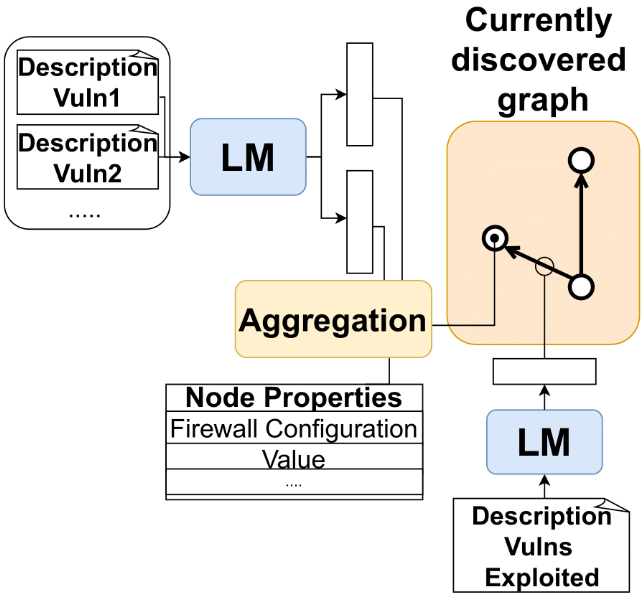
$$\mathrm{POOL}(\{\mathrm{embedding}_{\mathrm{node}_i} \in \mathbb{R}^p\}_{i \in \{1,\ldots,N\}}) \rightarrow \mathrm{embedding}_{\mathrm{graph}} \in \mathbb{R}^p$$

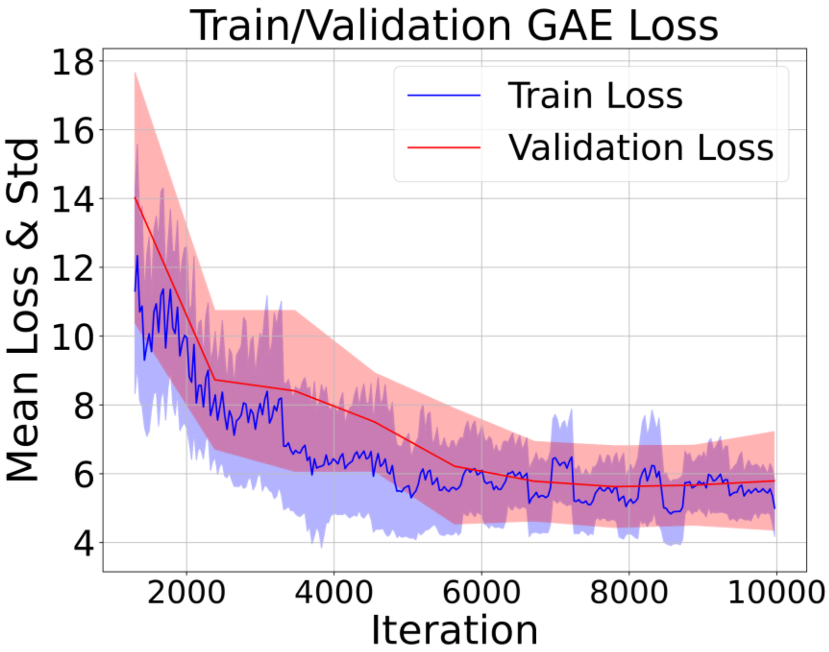
$$f_{\mathrm{LM}}(\textit{vulnerability description}) \rightarrow \mathrm{embedding}_{\mathrm{vulnerability}} \in \mathbb{R}^q$$

$$f_{\text{one-hot}}(o \in O) \rightarrow \mathrm{embedding}_{\mathrm{outcome}} \in \mathbb{R}^{|O|}$$

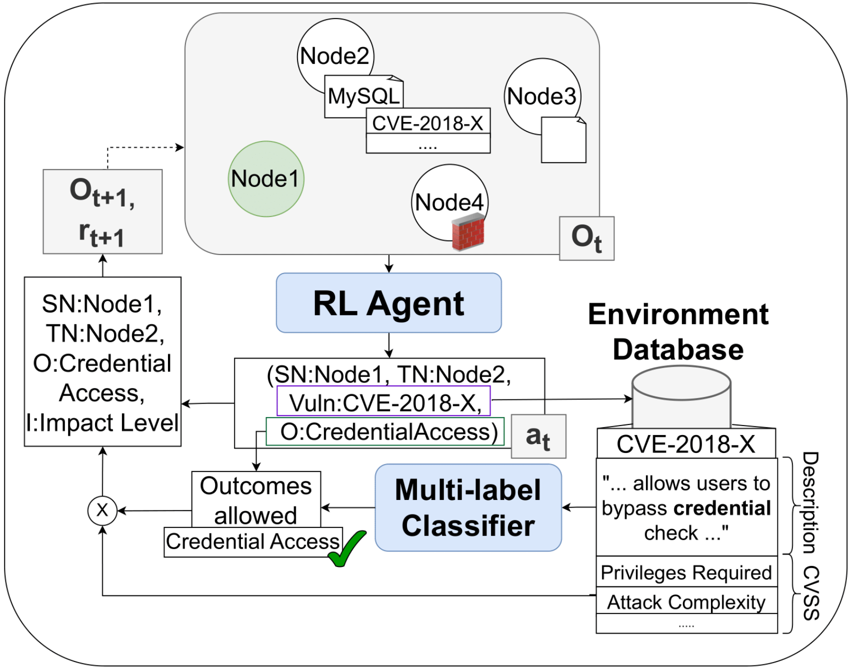


Graph Auto-Encoder

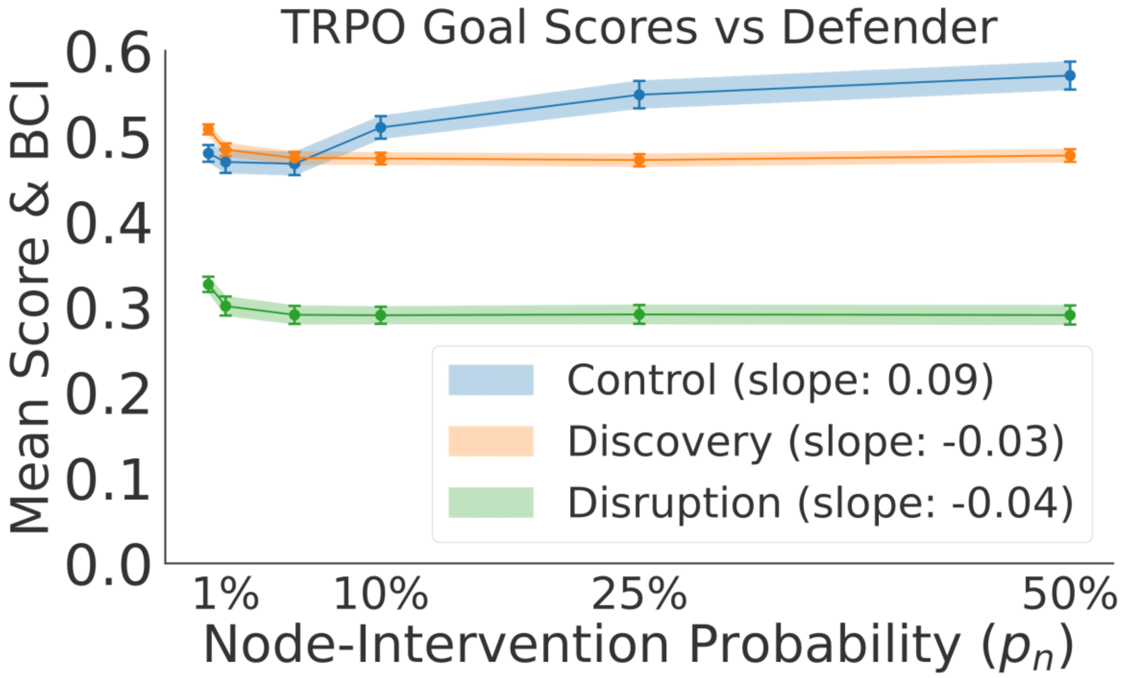




Automated Outcome Mapping



Dynamic Events Study



Environment Pipeline