

Détection de Cyberattaques Avancées par GNN

Apport des Graphes pour l'identification d'attaques avancées

Ludovic Andrieu, Léo Ritchie, Marion Duprez, Ruby Krishnaswamy, Frédéric Dang Tran

NETWORKS / NECTAR / équipe IRON

Présentation au
“Workshop about Telemetry, Provenance and Security Applications”
du 14 octobre 2026

Sommaire

- **Contexte et Enjeux** : Orange Innovation & La menace des Mouvements Latéraux.
- **Modélisation en Graphes de Provenance & détection d'anomalies via GNN**
- **Nos contributions: features, graphs, model & architecture**
- **Nos résultats**
- **Nos analyses: données et prédictions**
- **Conclusions et Perspectives**

Contexte : La detection des menaces



Le Défi de la Furtivité

Suite à une intrusion, les attaquants (APT) explorent l'infrastructure. C'est le **Mouvement Latéral (TA0008)**.

Pour rester invisibles, ils privilégient le *Living off the Land* : l'utilisation d'outils d'administration légitimes de l'OS.

Les approches classiques par signatures sont inopérantes. Il faut filtrer le bruit massif des SOC's et identifier des séquences d'actions suspectes sans saturer les analystes de fausses alertes.

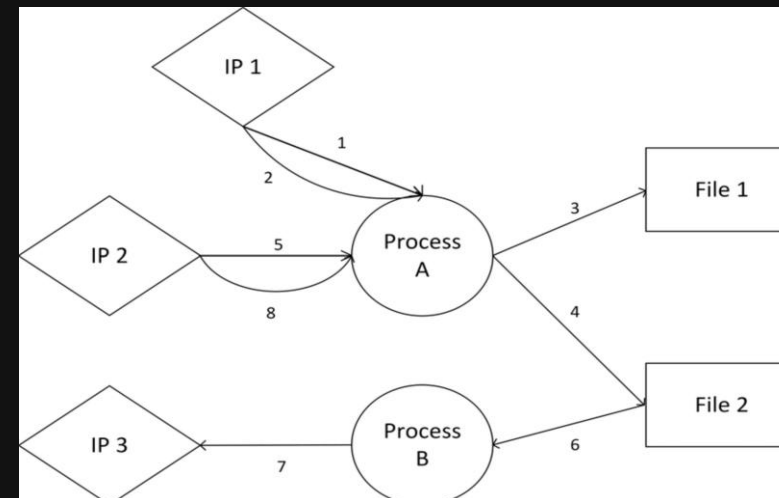
Objectifs:

- Développement d'un outil de détection d'attaques et mouvements latéraux à termes
- garder très bas le nombre de fausses alertes (Faux Positifs) dont les SOC's sont surchargés.

Graphes de **Provenance** & détection d'anomalies par **GNN**

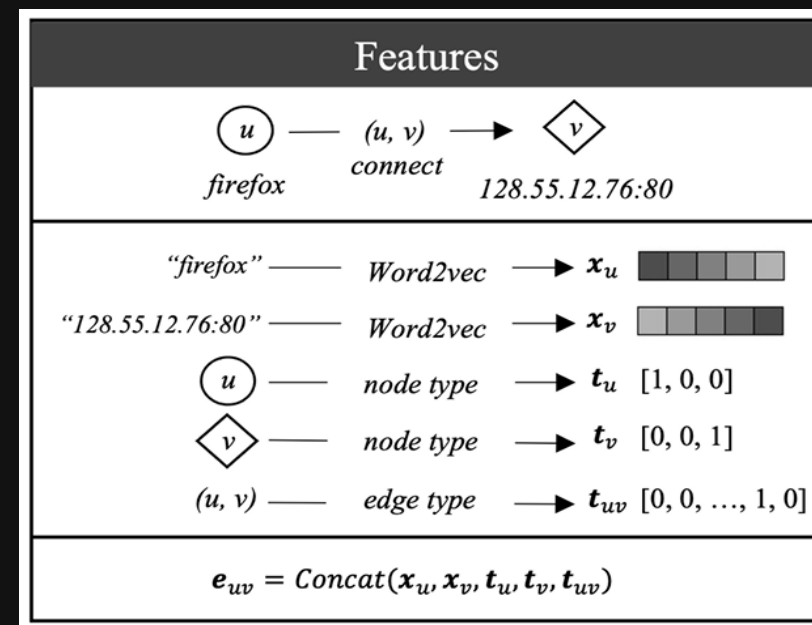
Modélisation Systémique

- Transformation des journaux systèmes bruts (logs) en une structure graphe.
- **Nœuds (Nodes)** : Processus, Fichiers, Sockets.
- **Arêtes (Edges)** : Les appels système (syscalls).
- **Atout** : Préserve la causalité stricte des événements pour reconstruire le cheminement d'une attaque.



Détection d'Anomalie par GNN

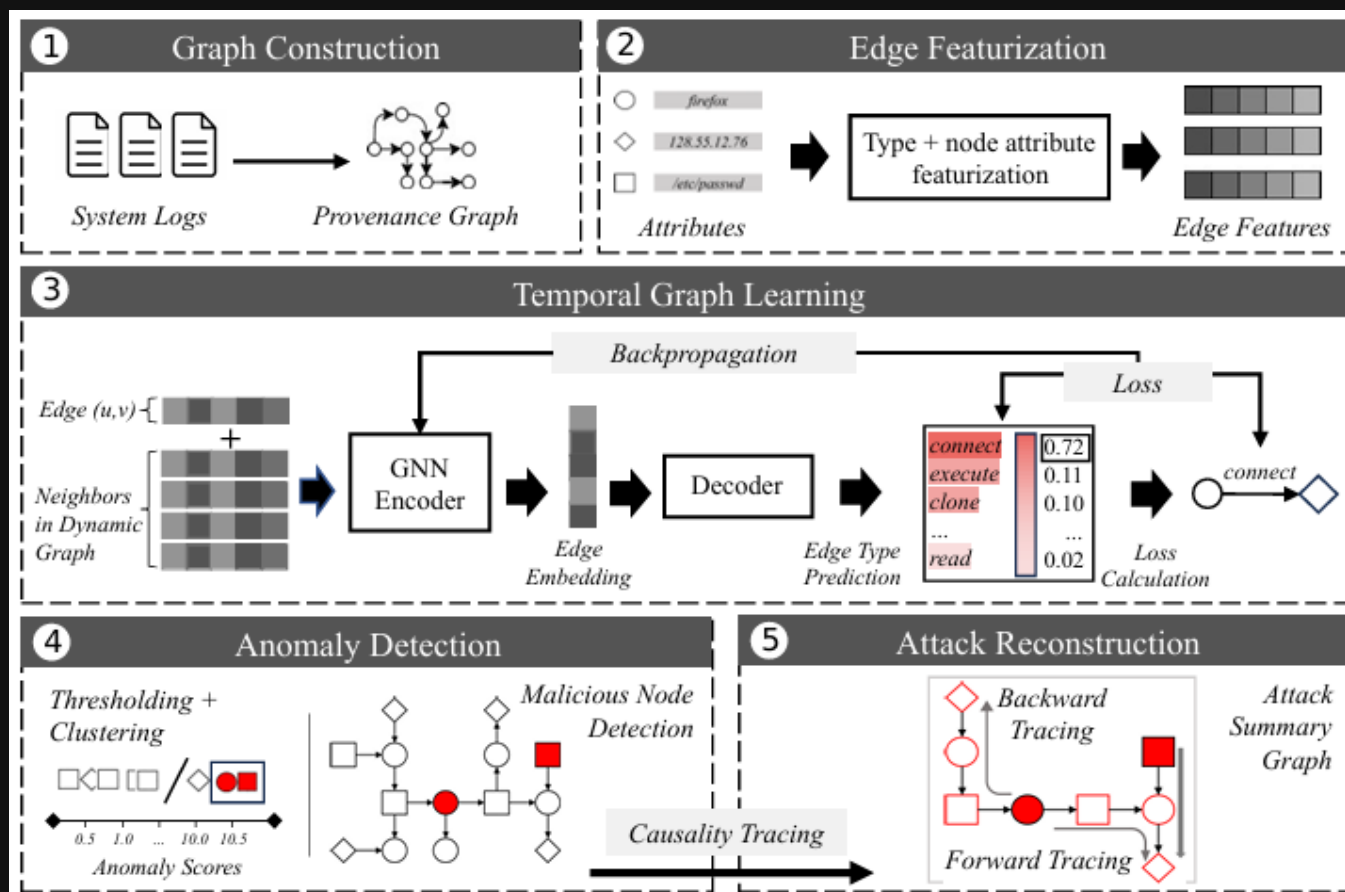
- L'architecture de base, pilotée via le framework expérimental **PIDSMaker**, s'articule autour d'un encodeur / décodeur.
- Le modèle s'entraîne uniquement sur des données bénignes pour apprendre la "normalité" du réseau.
- **Objectif typique** : Prédire le type d'arête entre deux noeuds. Une erreur de prédiction (Loss) élevée devient notre score d'anomalie.



Pipeline du Modèle Orthrus

L'architecture de base, pilotée via le framework expérimental **PIDSMaker**, s'articule autour d'un encodeur complexe.

- > **Selection des voisins:** récupération de context passé supplémentaire pour chaque noeuds du batch
- > **Encodeur GNN à attention locale:** Utilise TransformerConv (Graph Transformer) pour apprendre l'embedding de chaque noeud
- > **Décodeur :** Reconstitue le type de chaque arête pour évaluer la normalité de l'interaction.



Nos contributions: **features, graphs, model & architecture**

Features

- **ajouts d'appels systemes:** extension de 10 nouveaux appels systèmes supportés par le framework pour étendre les données exploitées.
- **time encoding :** utilisation d'un encodage linéaire 2D bijectif, préservant la stricte causalité des événements
- **propriétés structurelles :** métriques de centralité pour enrichir les nœuds de leurs propriétés macroscopiques, inaccessibles au champ de vision local du GNN.

Graphs

- **sélection de voisins:** ajouts de noeuds et arêtes supplémentaires
- **réduction des arêtes dupliquées** entre 2 nœuds avec même timestamp
- **agrégation des arêtes** de même type entre deux noeuds

Model

- prédiction de la **distribution** des appels systèmes entre deux noeuds du graphe.

Architecture

- **Directionnalité:** séparation explicite du traitement (attention) entre les flux entrants et sortants.
- **Attention globale à la transformer :** cross attention entre les noeuds et les triplets (src, dst, edge) construite à partir des embeddings des nœuds.

Nos contributions: **Les résultats**

- > **ajouts d'appels systemes:** bonnes améliorations de la détection en fonction du dataset.
- > **time encoding:** maintient les performances de détection tout en réduisant la complexité d'encoding trigonométrique.
- > **propriétés structurelles:** aide à réduire les faux positifs en stabilisant la décision, mais n'améliore pas la detection (pas plus de TP).
- > **agrégations des arêtes:** améliore fortement la détection sur tous les datasets.
- > **sélection de voisins:** bonnes améliorations de la détection en fonction du dataset.
- > **réduction des arêtes dupliquées:** n'améliore pas la détection mais réduit fortement le volume du jeu de données.
- > **directionnalité:** n'atteint de bons résultats qu'en forçant artificiellement une asymétrie extrême (propagation quasi-unidirectionnelle).
- > **attention globale:** performances équivalentes atteintes + bonnes perspectives, mais encore trop de variance entre les runs.

Nos analyses: données et predictions

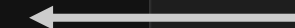
Données

Espace Word2Vec: 72% des fichiers et 100% des adresses IP d'attaque sont sémantiquement atypiques (au-delà du 95e percentile). **Cependant, ce décalage sémantique ne suffit pas à lui seul** : l'analyse combinée avec la structure du graphe reste indispensable.

Type de paire	Distance moyenne
attaque-attaque	0,241
attaque-bénin	0,300
bénin-bénin	0,249

Prédictions

- CaptumExplainer : *Captum* : Montre que les attaques sont caractérisées par la **nature des relations** (type d'arête), tandis que les **faux positifs** proviennent souvent d'une sémantique (Word2Vec) inhabituelle.
- GNNExplainer : Révèle que les motifs d'attaques forment des sous-graphes décisionnels compacts et étalés temporellement, à l'inverse de l'activité bénigne qui forme de vastes réseaux en étoile très concentrés.
- **Transport optimal** : Le calcul des distances (Gromov-Wasserstein) prouve mathématiquement que les sous-graphes d'attaque possèdent une signature topologique propre et mesurable, justifiant empiriquement l'usage des GNN.



Conclusions et Perspectives



Amélioration du PIDSmaker

- architecture
- jeux de données
- protocole d'évaluation



Synthéticité & discrétion

- données synthétiques
 - processus manifestes
- amélioration artificielle des perf



Popularité vs adoption faible

manque de naturalisme
→ **besoin de collecter des données réelles**



Guidelines

- modèle “one-fits-all”?
- covertness (co-design menace/bénin)
- étiquetage prudent des graphes
- réduction du bruit

Merci pour votre attention !

