GNN-IDS: Graph Neural Network based Intrusion Detection System

IIC3696 Tópicos Avanzados en Aprendizaje de Máquina
2024-2

José Antonio Valladares Betancourt
Pontificia Universidad Católica de Chile

Zhenlu Sun
André M. H. Teixeira
Salman Toor

Uppsala University, Suecia

2024 International Conference on Availability, Reliability and Security (ARES 2024)
30 de julio - 02 de agosto de 2024
Vienna, Austria

Tabla de contenido

Sistemas de Detectión de Intrusiones (IDS)

Identificar y generar alertas por comportamiento intrusivo en las redes de comunicación.

IDS basado en firmas

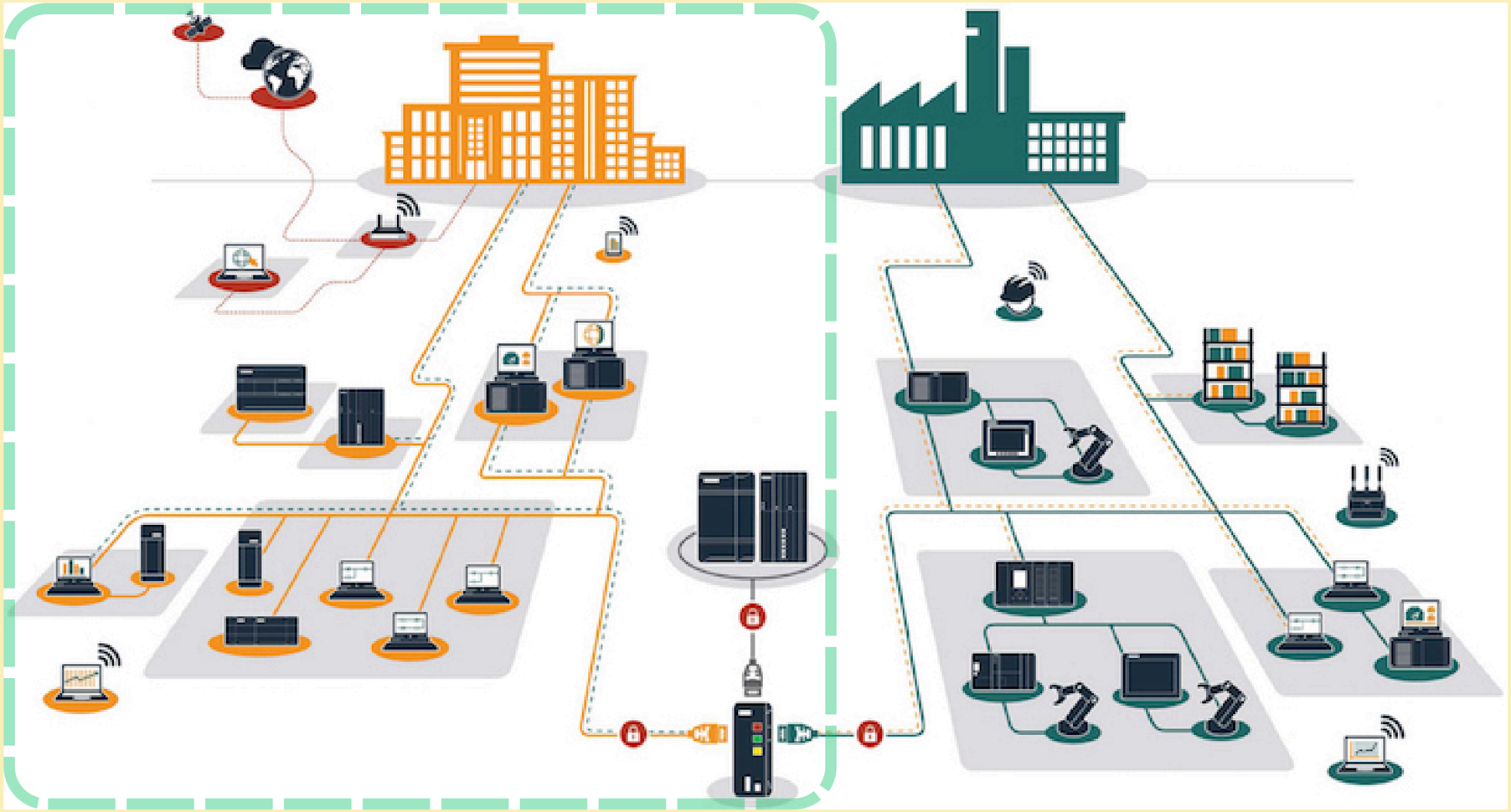
- Compara los patrones con una base de conocimientos predefinida
  - **No detecta ataques desconocidos**

IDS basado en anomalías

- Basado en estadísticas, conocimiento de expertos o aprendizaje automático (ML), identifica comportamientos anómalos
  - **Falsos Positivos**
  - **Trata cada muestra como aislada, no considera relaciones entre muestras**

Red

Este trabajo solo abarca redes IT.



Usos de GNN en IDS

Enfoque en la red: Representan la red como grafo (nodos: hosts, enlaces: comunicación)

- **Grafos de flujo**
  - Enlaces: vector de características con flujos de red (5-tuple: dirección IP de origen y destino, puertos de origen y destino, protocolo)

- **Grafos de paquetes**
  - Cada paquete capturado constituye un enlace individual

- **Grafos de autenticación**
  - Los enlaces representan solicitudes de autenticación y pueden tener atributos

Usos de GNN en IDS

Enfoque en huéspedes: Se enfocan en el análisis de la actividad interna de un sistema

- **Grafos de procedencia:**
  - Los nodos representan entidades del sistema y los enlaces representan las interacciones

- **Grafos de Llamadas al Sistema:**
  - Los nodos representan las llamadas al sistema individuales realizadas por un proceso, y los enlaces representan la relación secuencial entre las llamadas

GNN-IDS

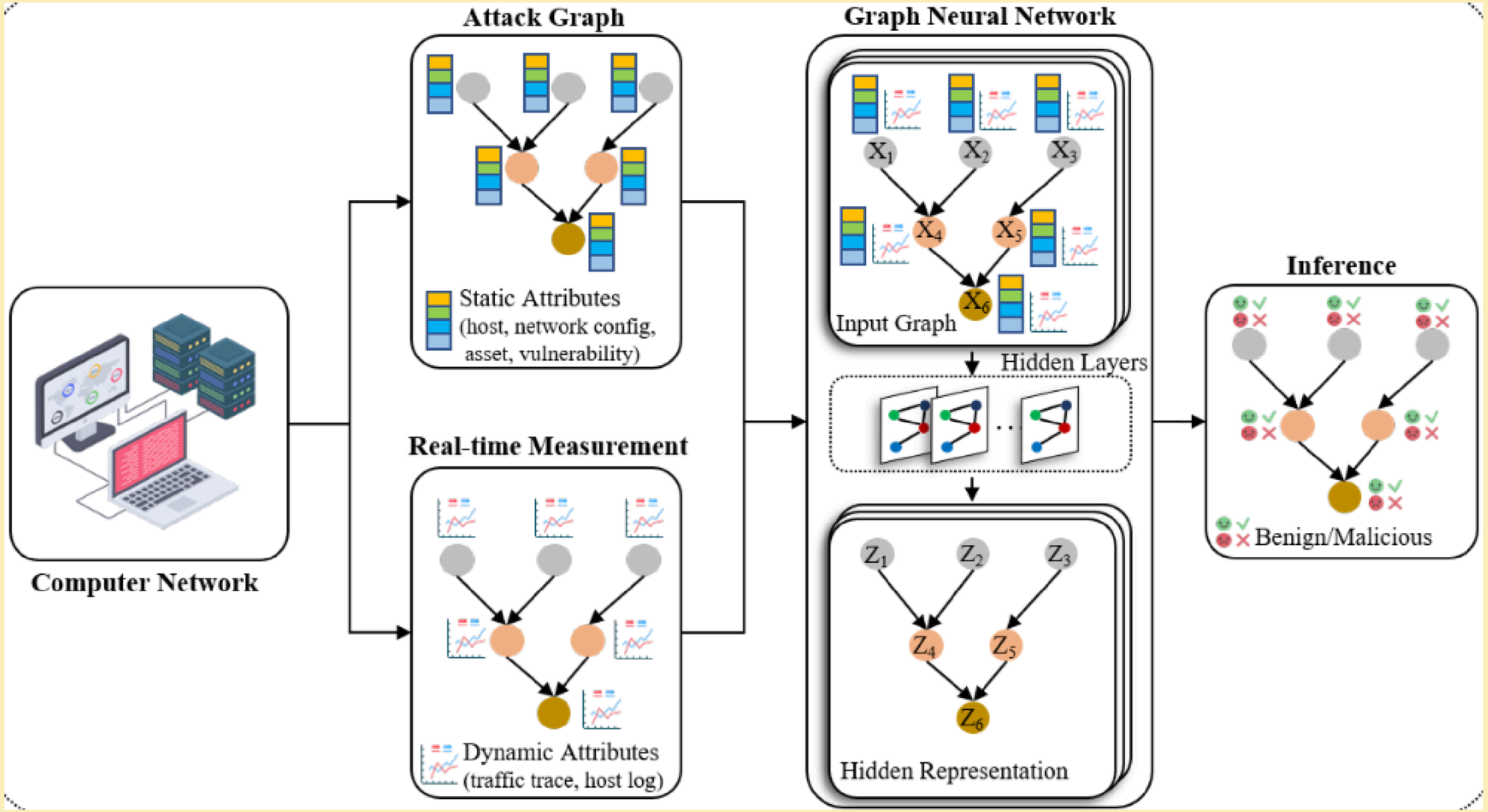
Objetivos

Clasificación de nodos: detectar host bajo ataque y acciones del atacante

- Incertidumbre: predicciones confiables
- Explicabilidad: encontrar la importancia de las características
- Robustez: observar el comportamiento del modelo ante entradas ruidosas

Sin embargo...

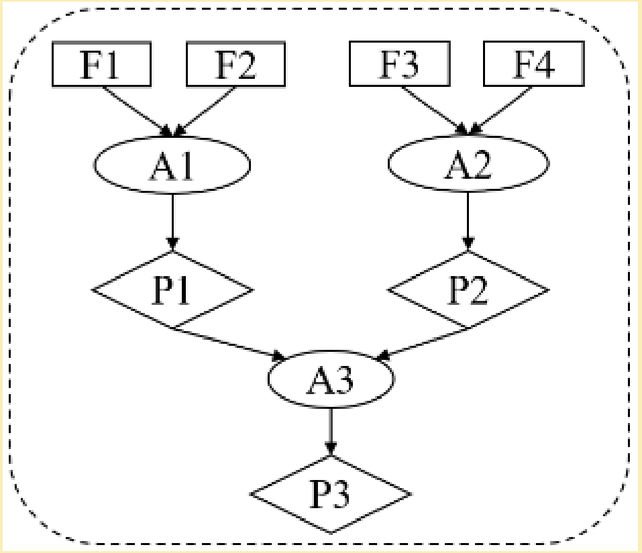
GNN-IDS


Attack Graph
Graph Neural Network
Inference
Computer Network
Static Attributes
(host, network config,
asset, vulnerability)
Real-time Measurement
Dynamic Attributes
(traffic trace, host log)
Input Graph
Hidden Layers
Hidden Representation
Benign/Malicious

Grafo de Ataque

- Grafo dirigido acíclico
- Formado a partir de información de la red
- Modela posibles caminos de ataque

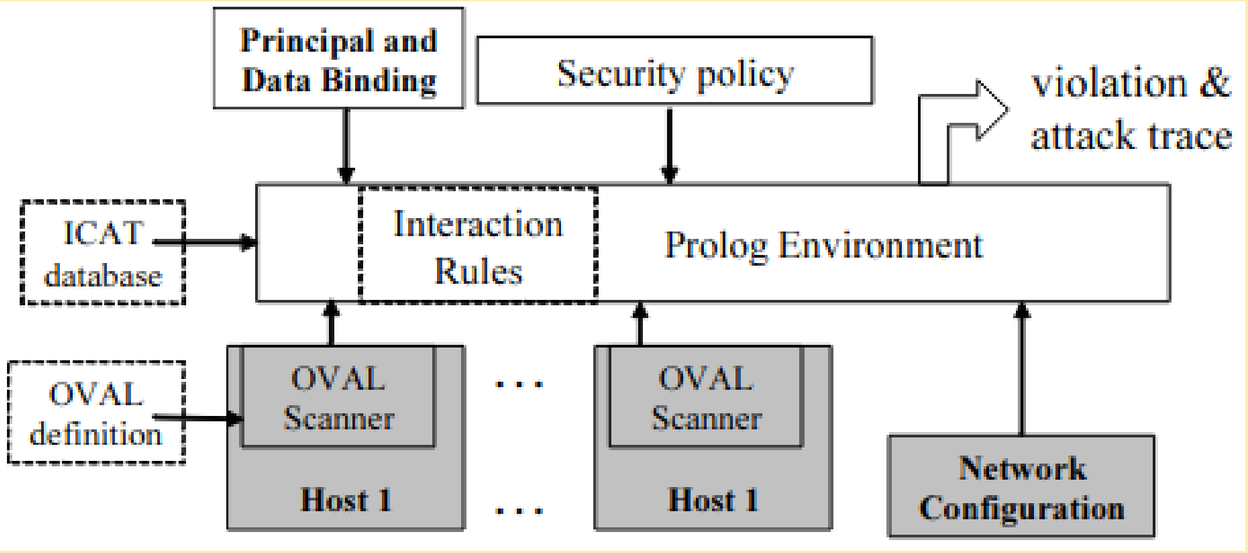
- F: Fact node (información de la red)
- A: Action node (como el atacante puede comprometer la red)
- P: Privilege node (consecuencias de cada fase de ataque)

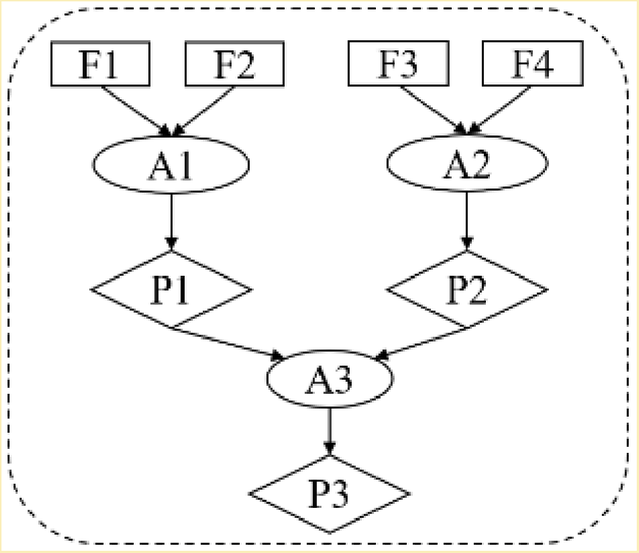


Grafo de Ataque
MulVAL

MulVAL [1] (multi-host, multi-stage vulnerability analysis)

- se forma a partir de especificaciones de vulnerabilidades (OVAL e ICAT)
- utiliza lenguaje de programación Datalog
- entrega un grafo de ataque con cláusulas textuales





1- Ou, X., Govindavajhala, S., & Appel, A. W. (2005, August). MulVAL: A logic-based network security analyzer. In USENIX security symposium (Vol. 8, pp. 113-128).

Grafo de Ataque
MulVAL

Advisories: vulnerabilidades reportadas existentes en los dispositivos.

```
vulExists(webServer, ’CAN-2002-0392’, httpd).
vulProperty(’CAN-2002-0392’, remoteExploit, privilegeEscalation).
```

Configuración de host: software y servicios activos, y su configuración.

```
networkService(webServer, httpd, TCP, 80, apache).
```

Configuración de red: configuración de routers y firewalls.

```
hacl(internet, webServer, TCP, 80).
```

Principals: identidad de usuarios de la red.

```
hasAccount(user, projectPC, userAccount).
hasAccount(sysAdmin, webServer, root).
```

Grafo de Ataque
MulVAL

Interacción: modelo de interacción entre todos los componentes.

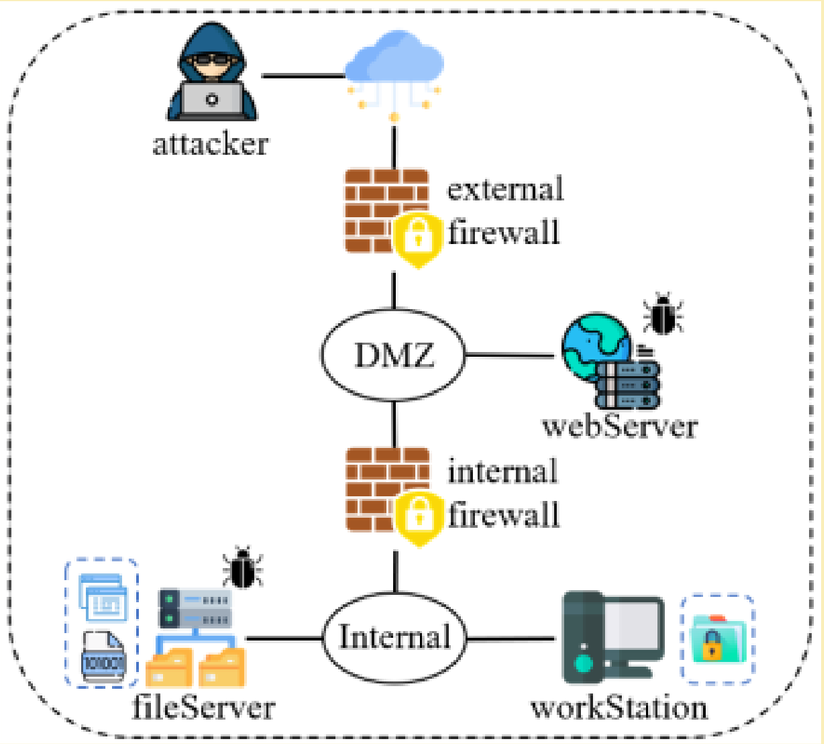
```
execCode(Attacker, Host, Priv) :-
         vulExists(Host, VulID, Program),
         vulProperty(VulID, remoteExploit, privEscalation),
         clientProgram(Host, Program, Priv),
         malicious(Attacker).
```

Política: permisos de acceso deseados.

```
allow(Everyone, read, webPages).
allow(systemAdmin, write, webPages).
```

Grafo de Ataque
Red usada

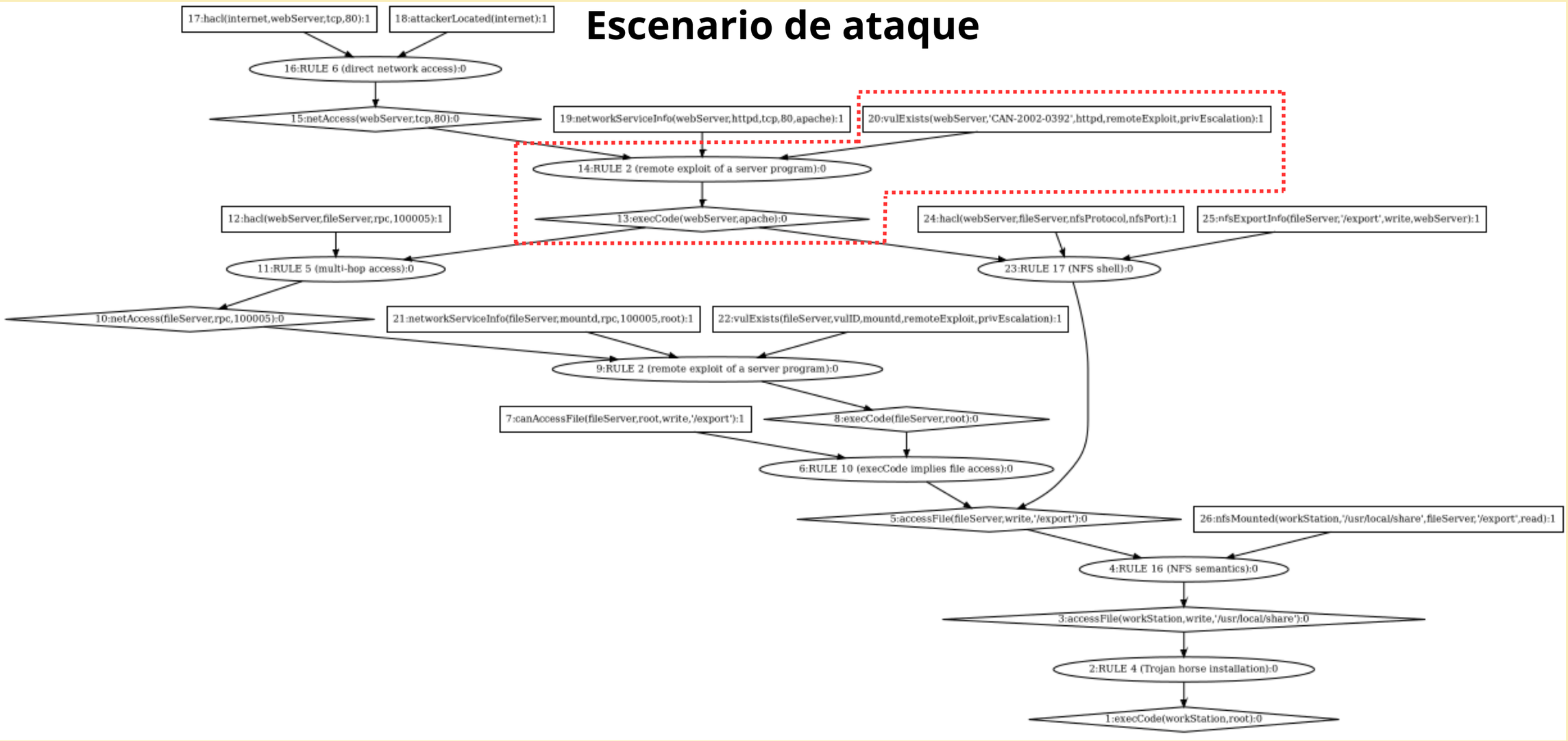
- Zona externa, zona desmilitarizada (DMZ) y zona interna separadas por firewall
- *webServer* y *fileServer* manejados por administradores
- Usuarios normales usan *workStation* para acceder a ejecutable en *fileServer*
- Vulnerabilidades:



Server	CVE ID	Impacted Protocol	Description
webServer	CVE-2002-0392	HTTP	Allow remote execution of code or DoS attack
fileServer	CVE-2003-0252	NFS	Allow remote execution of code or DoS attack

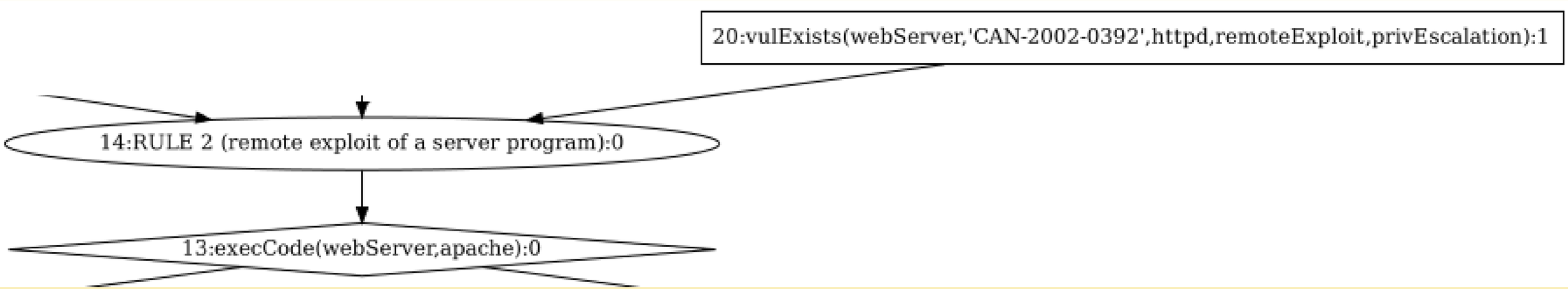
Grafo de Ataque

Escenario de ataque



1. Atacante con acceso a internet explota remotamente vulnerabilidad *CVE-2002-0392* para obtener acceso local al *webServer*.

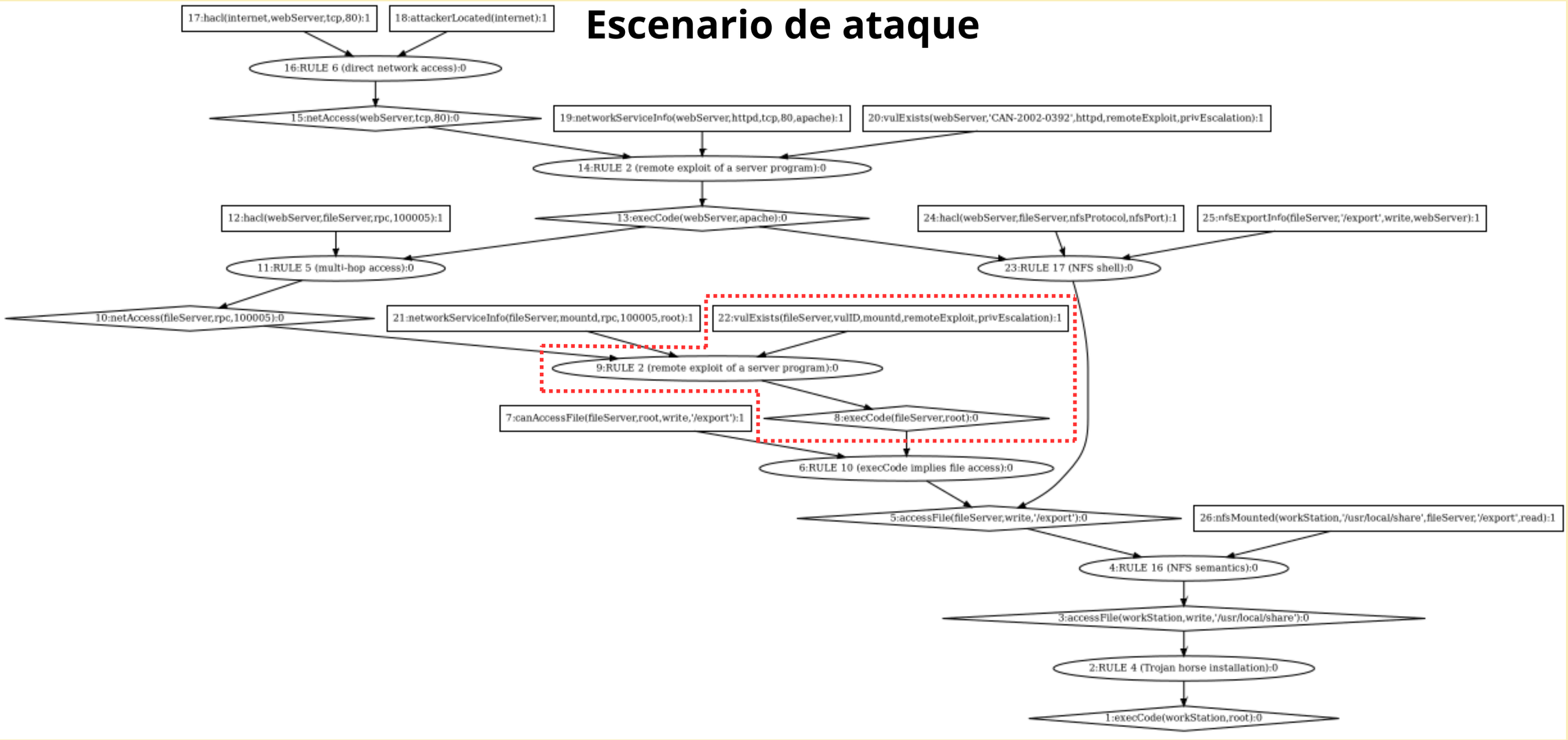
Grafo de Ataque
Escenario de ataque



1. Atacante con acceso a internet explota remotamente vulnerabilidad *CVE-2002-0392* para obtener acceso local al *webServer*.

Grafo de Ataque

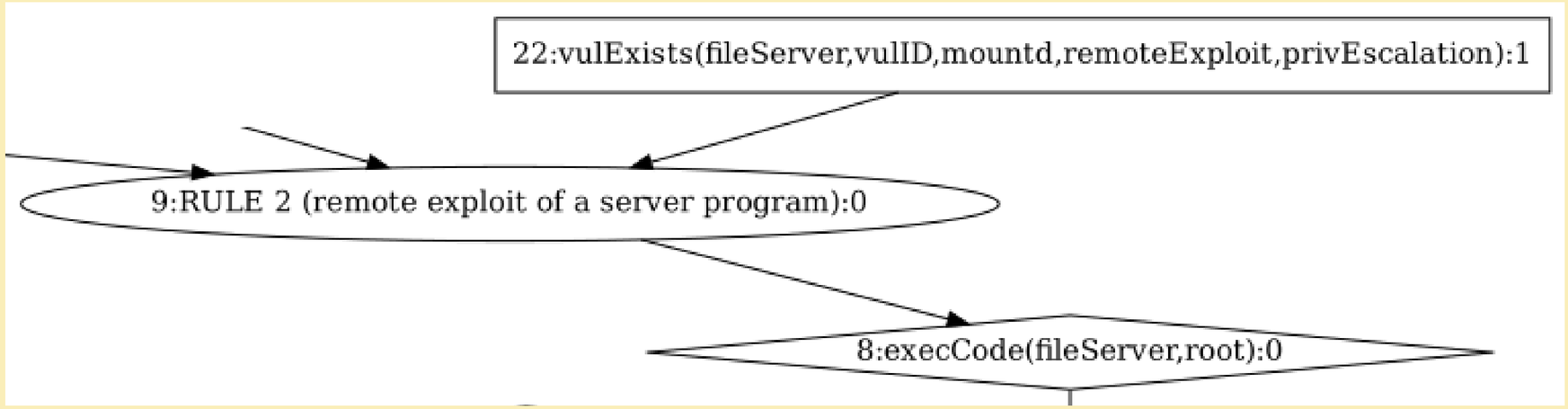
Escenario de ataque



2. Atacante explota vulnerabilidad CVE-2003-0252 en fileServer para obtener privilegios de root.

Grafo de Ataque

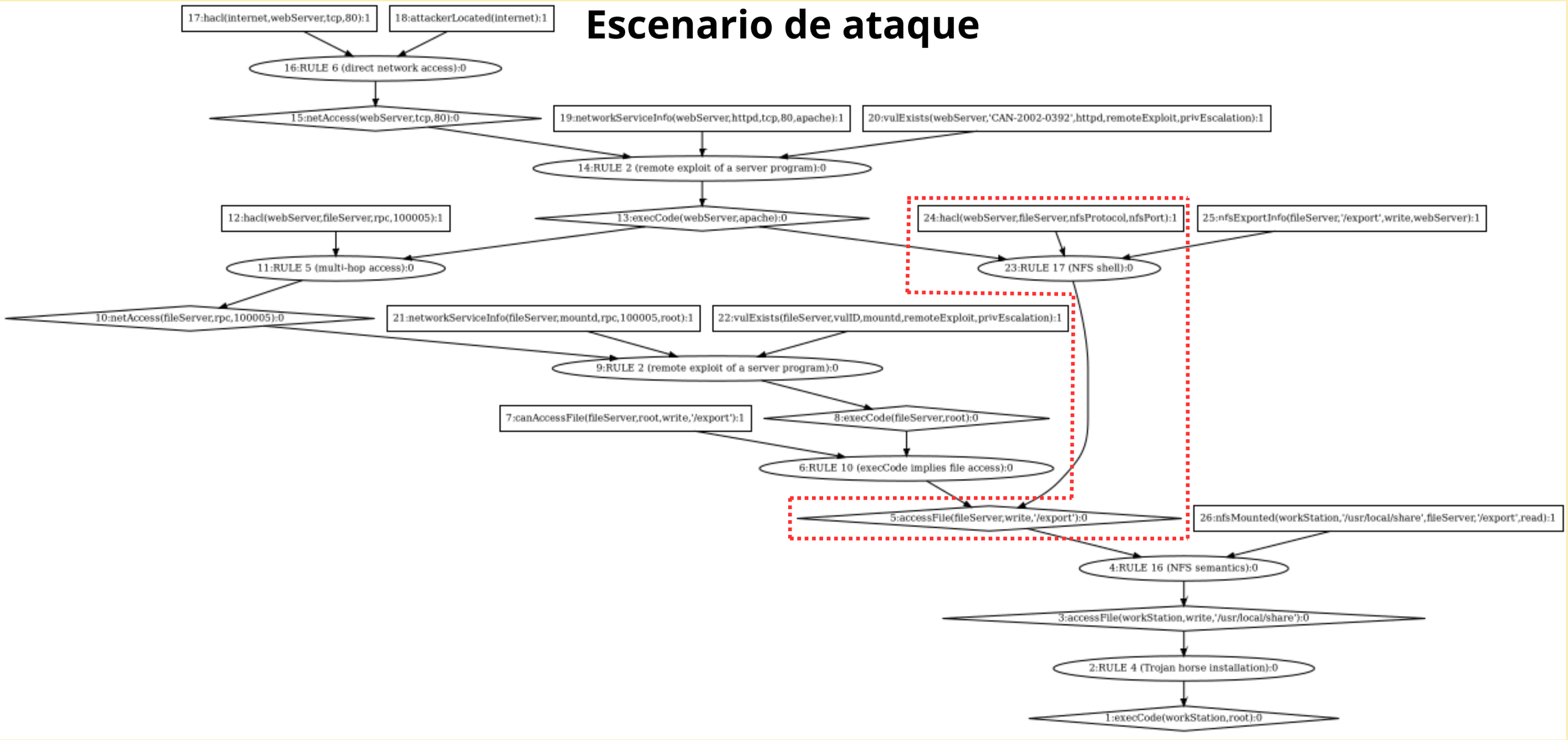
Escenario de ataque



2. Atacante explota vulnerabilidad CVE-2003-0252 en fileServer para obtener privilegios de root.

Grafo de Ataque

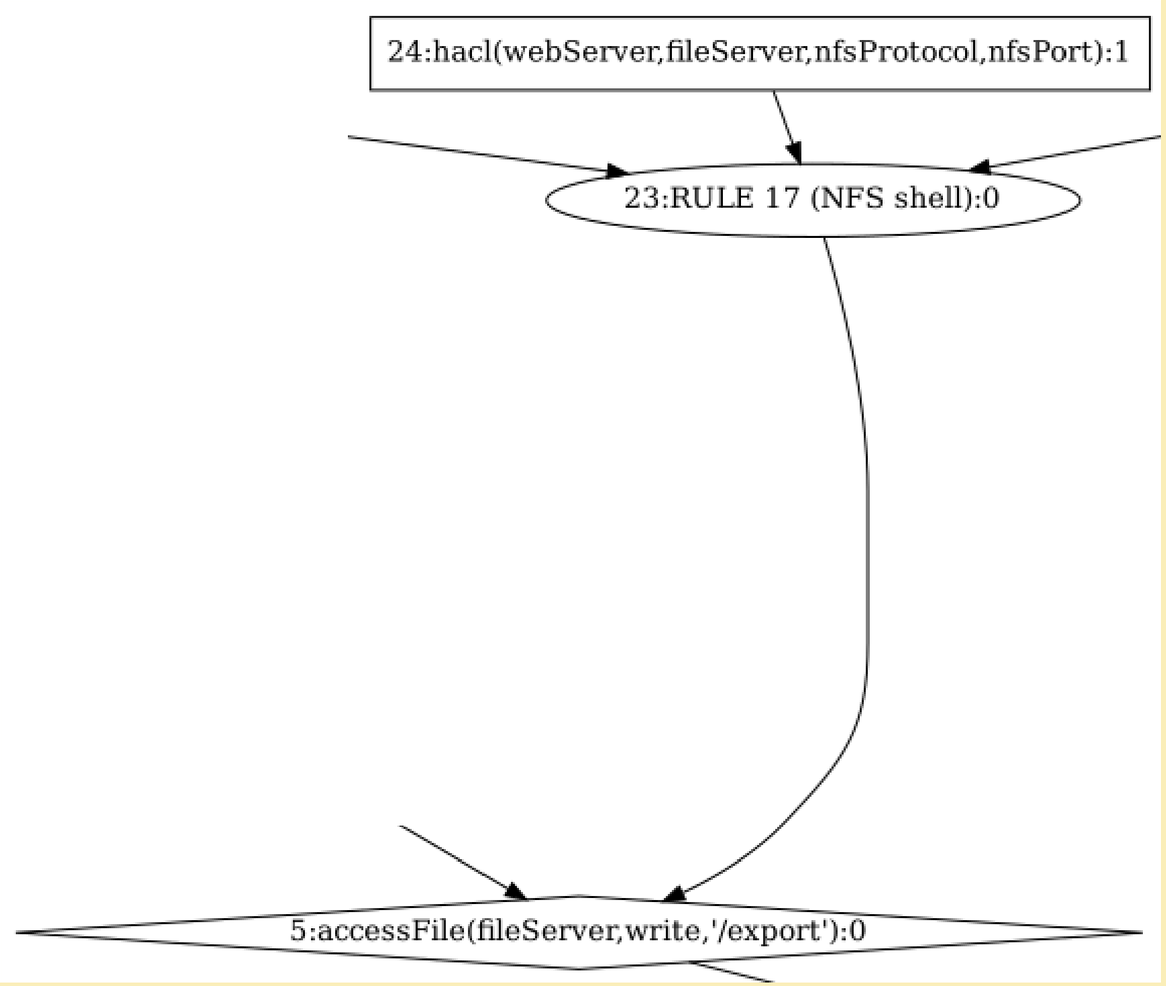
Escenario de ataque



3. Atacante podría modificar archivos en fileServer usando protocolo NFS (Network File System) si la tabla de exportación NFS está configurada incorrectamente.

Grafo de Ataque

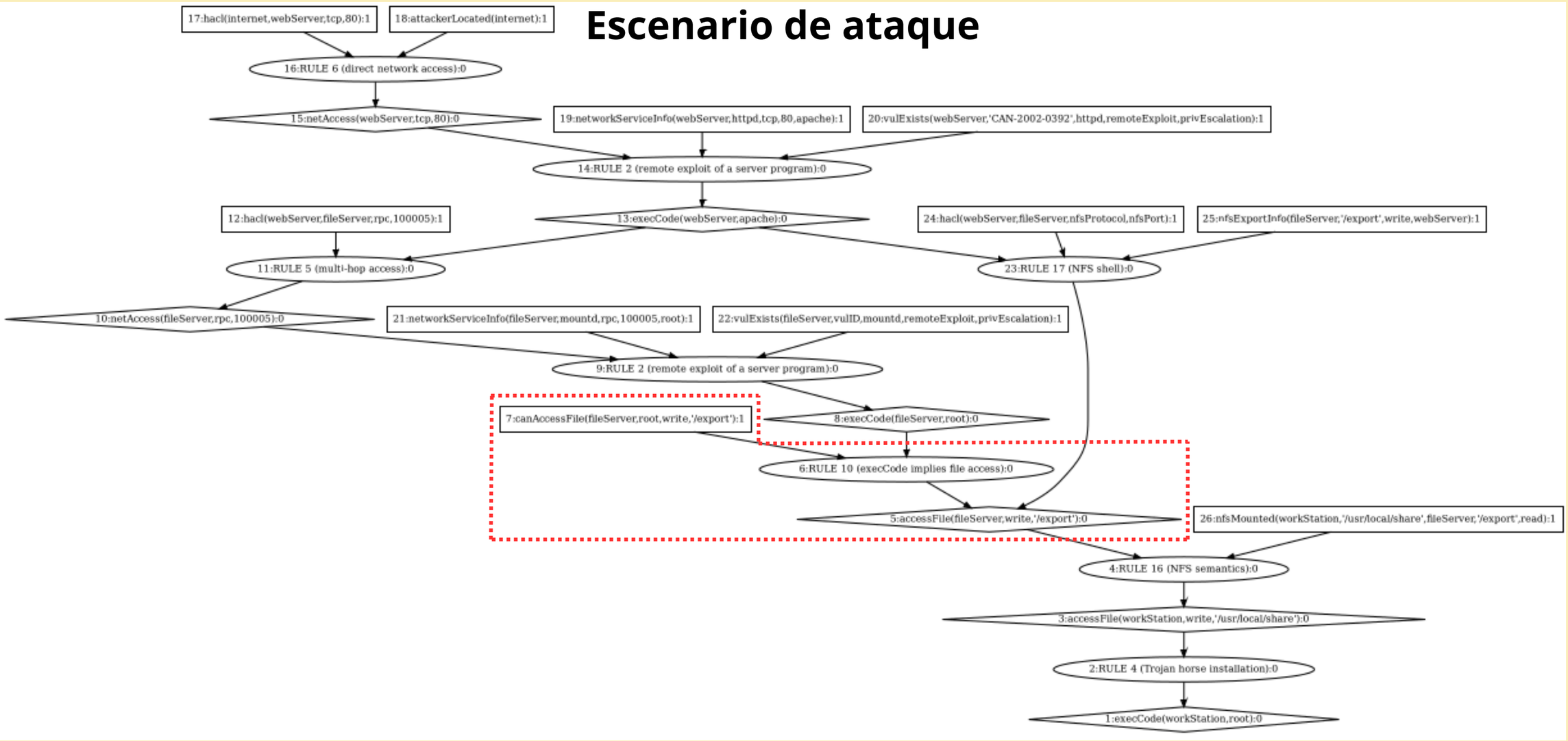
Escenario de ataque



3. Atacante podría modificar archivos en fileServer usando protocolo NFS (Network File System) si la tabla de exportación NFS está configurada incorrectamente.

Grafo de Ataque

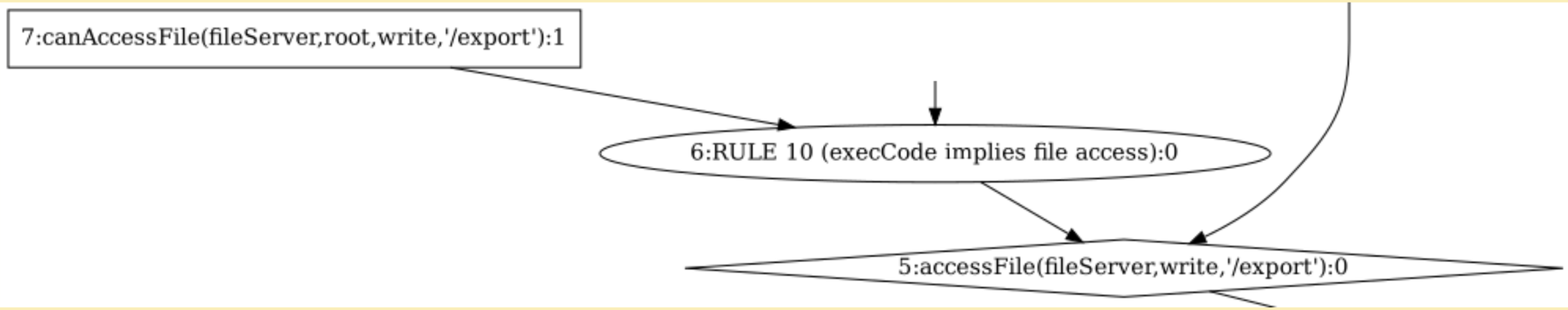
Escenario de ataque



4. Tras escalar privilegios, el atacante instala un malware tipo caballo de Troya en los binarios ejecutables del fileServer.

Grafo de Ataque

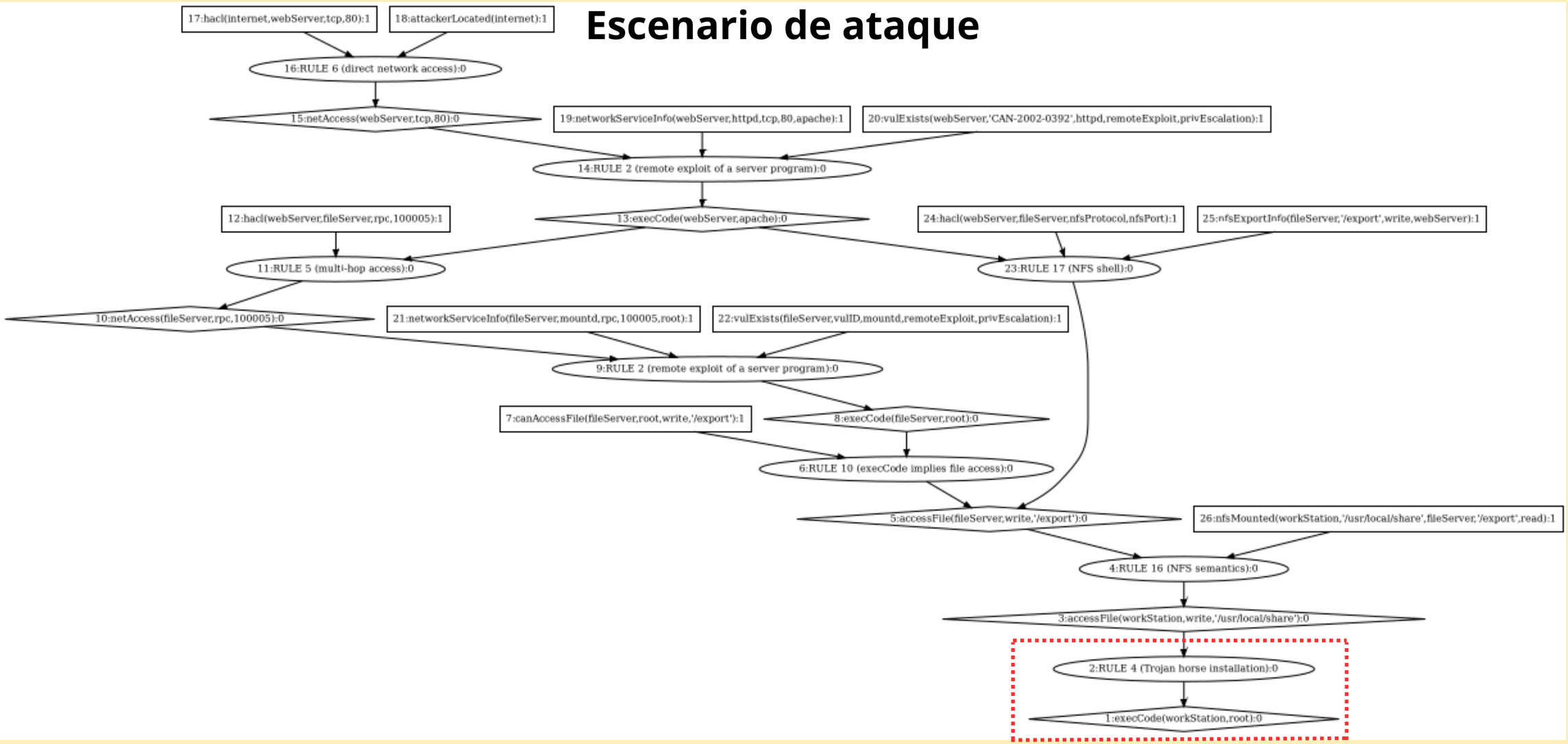
Escenario de ataque



4. Tras escalar privilegios, el atacante instala un malware tipo caballo de Troya en los binarios ejecutables del fileServer.

Grafo de Ataque

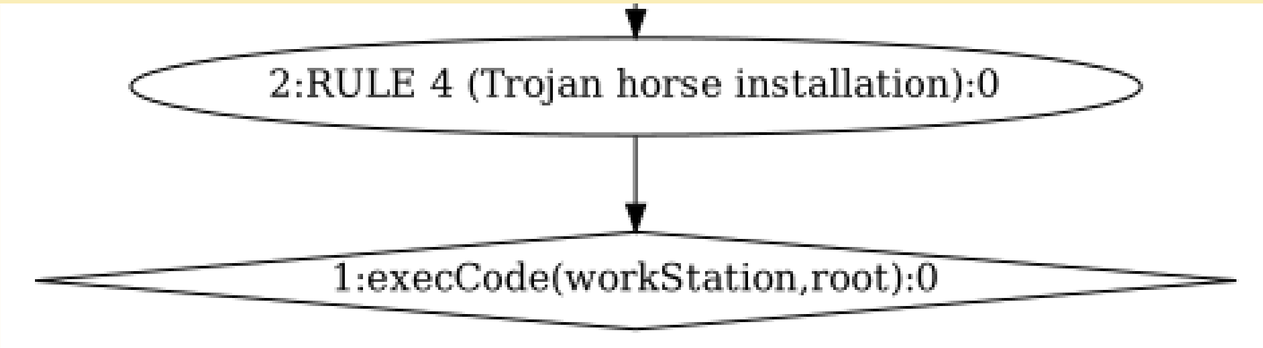
Escenario de ataque



5. Como los usuarios comunes utilizan workStation para computación, es probable que el caballo de Troya se monte y ejecute en el workStation, causando filtración de datos.

Grafo de Ataque

Escenario de ataque



5. Como los usuarios comunes utilizan workStation para computación, es probable que el caballo de Troya se monte y ejecute en el workStation, causando filtración de datos.

Grafo de Ataque
One-hot Encoding

Input: nodes V, edges E, statements $\{S_v | \forall v \in V\}$
Output: encoded attack graph $\mathcal{G} = (V, E, \mathbf{F})$

```
1:  i ⇐ 0
2:  for v ∈ V do
3:    for w ∈ S_v do
4:      if w ∉ C then
5:        C_i ⇐ w
6:        i ⇐ i + 1
7:      end if
8:    end for
9:  end for
10: D ⇐ |C|
11: for v ∈ V do
12:   i ⇐ 0
13:   while i < D do
14:     if C_i ∈ S_v then
15:       f_vi ⇐ 1
16:     else
17:       f_vi ⇐ 0
18:     end if
19:     i ⇐ i + 1
20:   end while
21: end for
```

w : token

Sv : enunciado del nodo

C : corpus

D : tamaño del diccionario

fv : vector de características codificado

Mediciones en tiempo real

Datos públicos

- Fragmento del dataset CIC-IDS2017.
- 40500 registros
- 78 características
- 7 clases de ataque

Datos sintéticos

- Criterios de generación de features:
  - Distribución Gaussiana
  - Distribución de Poisson
  - Combinación no lineal de las anteriores

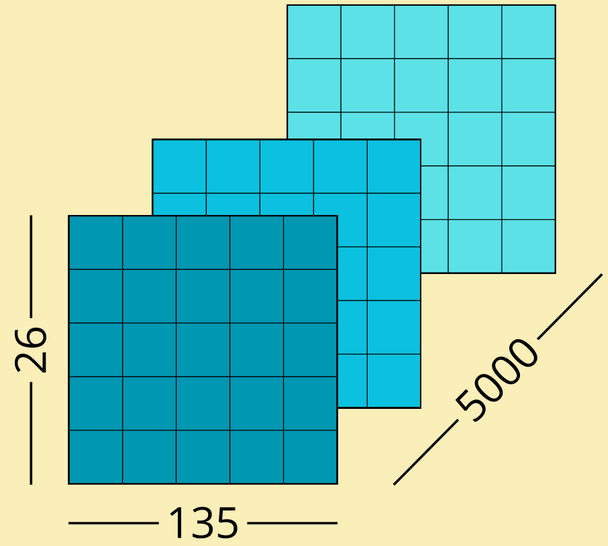
Uso de mediciones:

$$\mathbf{x}_v^t = \begin{cases} \mathbf{f}_v \parallel \mathbf{m}_{action}^t & \text{if } action \in S_v \text{ and } v \in \{\text{privilege nodes}\} \\ \mathbf{f}_v \parallel \{\mathbf{0} \in \mathbb{R}^K\} & \text{otherwise} \end{cases}$$

Mediciones en tiempo real

Resumen de datos

- Tamaño del vocabulario: 57
- Cantidad de nodos: 26
- Cantidad de enlaces: 26
- Cantidad de privilege nodes: 7 (asociado a los tipos de ataque)
- Cantidad de muestras (cantidad de grafos): 5000 (por cada dataset)
- Cantidad de mediciones: 78



Modelos

- Neural Network (NN)

- Graph Convolution Network (GCN): aprender representaciones de cada nodo combinando uniformemente sus características con las de sus vecinos

- Graph Convolution Network with learnable Edge Weights (GCN-EW): asignar importancia variable a conexiones entre nodos

- Graph Attention Network (GAT): utilizar mecanismos de atención para asignar pesos adaptativos a las conexiones de cada nodo

$$h_v^l = f^l(h_v^{l-1}, AGG^l(\{h_u^{l-1}, \forall u \in N_u\}))$$

$$h_v = \Phi(h_v^l)$$

$$h_G = R(h_v^l | v \in V)$$

—

$$h_i' = \sigma(W^T \sum_{j \in N_i \cup \{i\}} \frac{w_{ji}}{\sqrt{\hat{d}_j \hat{d}_i}} h_j)$$

$$\hat{d}_i = 1 + \sum_{j \in N_i \cup \{i\}} w_{ji} \cdot w_{ji} \in R$$

—

$$a_{ij} = \frac{exp(LeakyReLU(a^T[Wh_i \| Wh_j]))}{\sum_{k \in N_i} exp(LeakyReLU(a^T[Wh_i \| Wh_k]))}$$

$$h_i' = \sigma(\sum_{j \in N_i} a_{ij} W h_j)$$

$$h_i' = \|_{k=1}^K \sigma(\sum_{j \in N_i} a_{ij}^k W^k h_j)$$

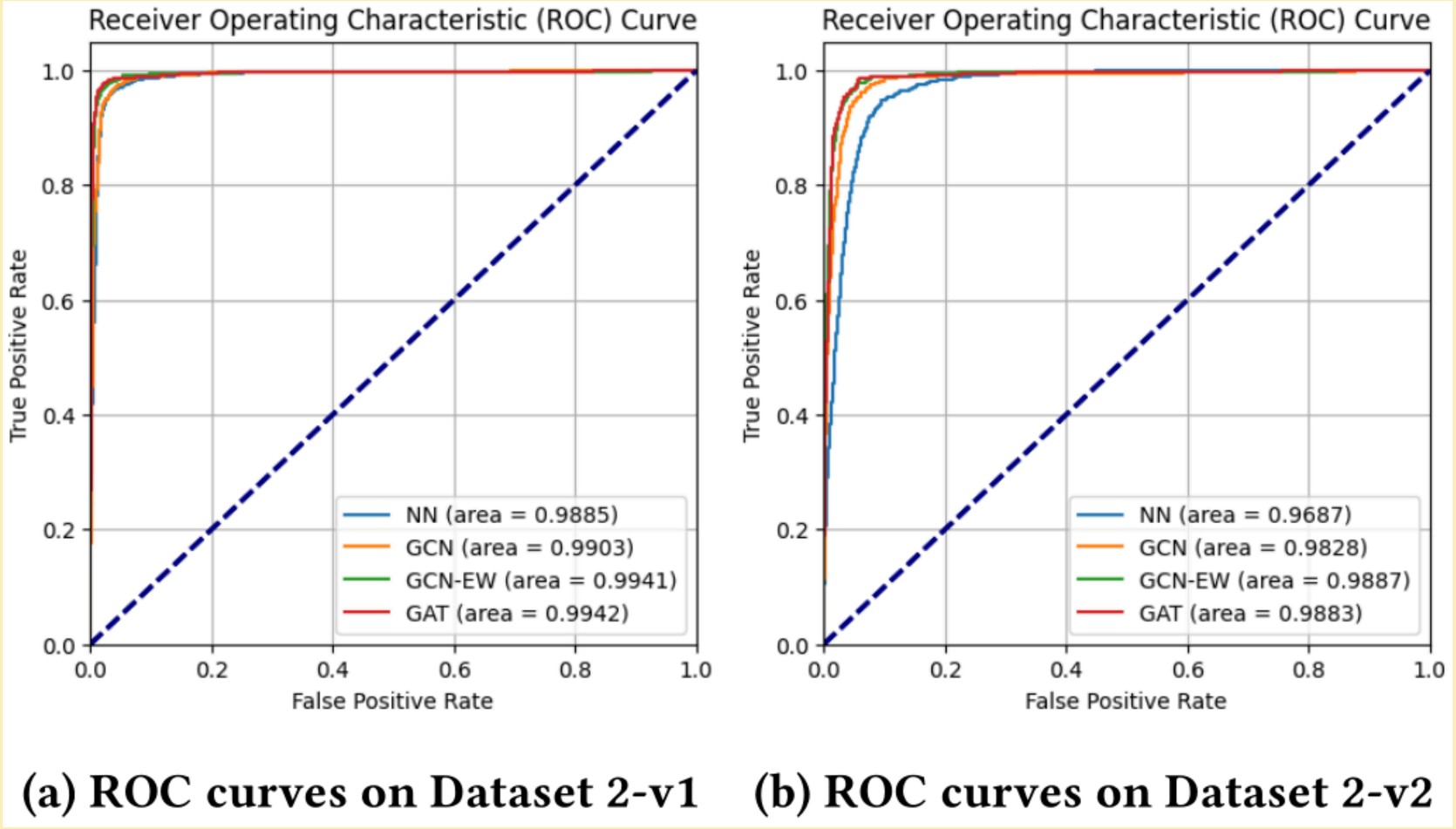
Resultados

Table 4: Statistics on standard metrics (in percentage) of NN and GNN models evaluated on Datasets 1 and 2

Dataset	Model	Prec	Recall	F1-score	AUC	FPR
1	NN	71.35	85.82	75.62	93.76	11.37
	GCN	83.47	94.24	87.80	98.33	4.95
	GCN-EW	**86.84**	95.47	**90.53**	**98.94**	**3.63**
	GAT	86.61	**95.49**	90.39	98.87	3.73
2	NN	88.25	97.08	92.04	98.98	3.27
	GCN	91.76	97.74	94.49	99.40	2.10
	GCN-EW	94.14	98.02	95.97	99.65	1.40
	GAT	**94.74**	**98.72**	**96.62**	**99.72**	**1.27**

Resultados

Incertidumbre: Usan 2 variantes del dataset de datos públicos con ruido Gaussiano para comparar predicciones entre ellos.



(a) ROC curves on Dataset 2-v1 (b) ROC curves on Dataset 2-v2

Resultados

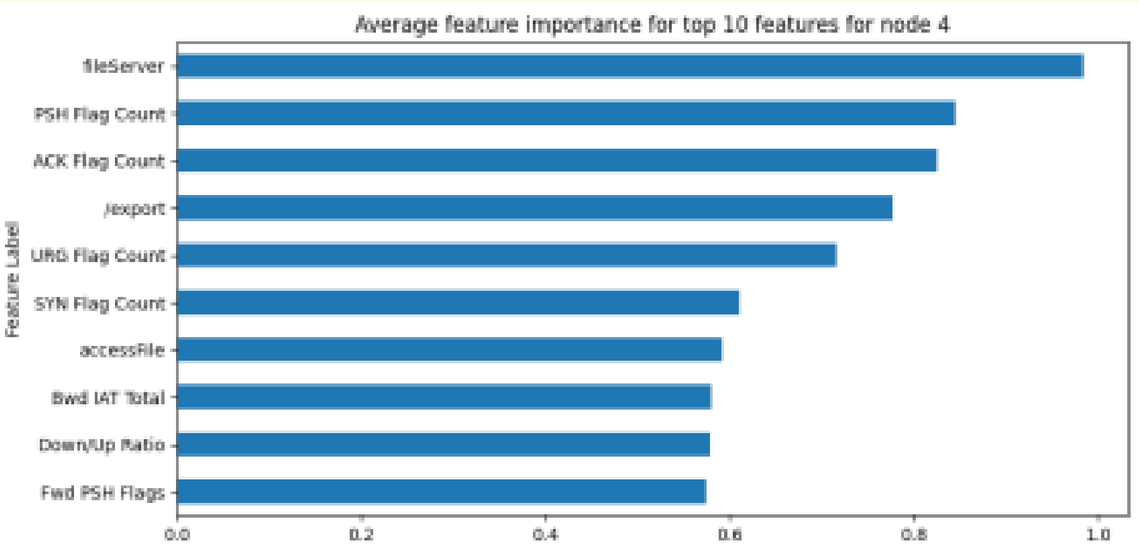
Robustez: Usan otra variante de dataset con ruido Gaussiano para comparar predicciones con dataset original.

Table 5: Performance changes (in percentage) of NN and GNN models evaluated on noisy test set of Datasets 1 and 2

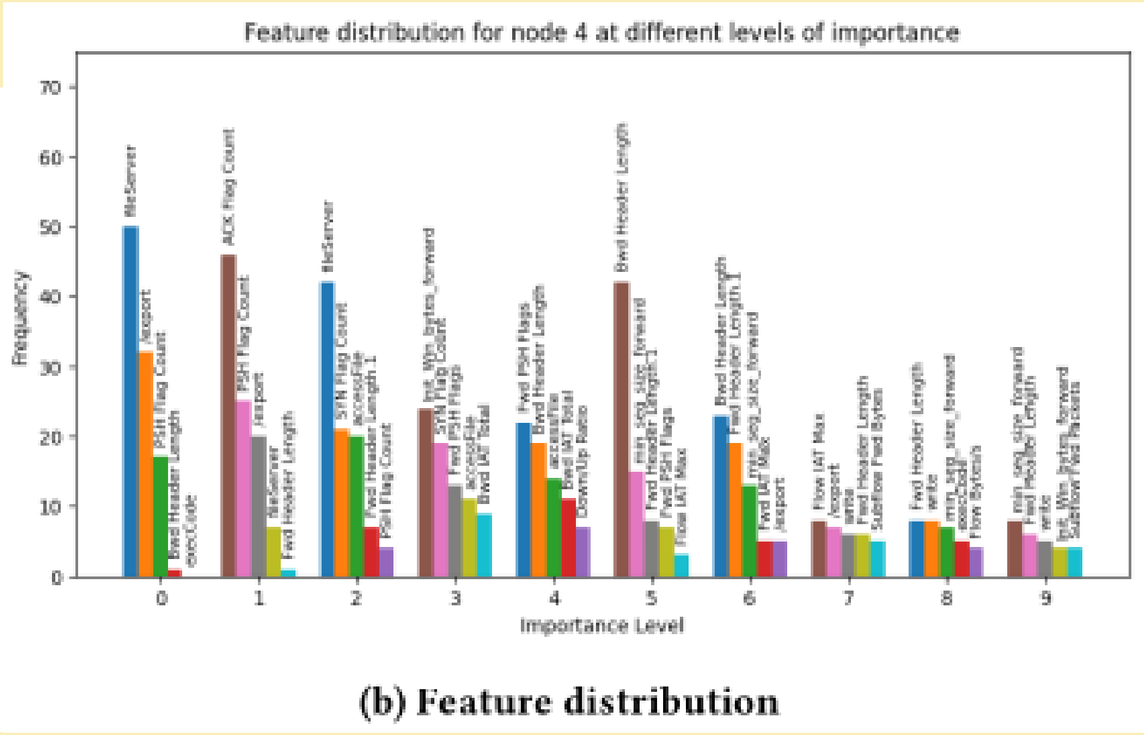
Dataset	Model	Prec	Recall	F1-score	AUC	FNR
1	NN	-1.03	-2.43	-1.36	-2.15	+4.86
	GCN	-0.56	-1.31	-0.81	-0.30	+2.57
	GCN-EW	-0.65	-2.00	-1.15	-0.48	+4.00
	GAT	**-0.39**	**-1.21**	**-0.69**	**-0.22**	**+2.42**
2	NN	-1.32	-4.29	-2.47	-1.25	+8.57
	GCN	**-0.25**	**-0.80**	**-0.48**	**-0.13**	**+1.57**
	GCN-EW	-0.31	-1.43	-0.826	-0.19	+2.86
	GAT	-0.63	-3.00	-1.72	-0.29	+6.00

Resultados

Explicabilidad: Usan GNNEXPLAINER. Muestran features de mayor importancia en predicciones sobre 100 muestras de ataque tipo 4 con modelo GAT. Aparecen features del dataset y del grafo de ataque.



(a) Average feature importance



(b) Feature distribution

Conclusiones

Todos los grafos tienen los mismos features relativos al grafo de ataque porque es un solo grafo de ataque.

Las mediciones se usan en todos los nodos de privilegio. No se discrimina cuales son los nodos involucrados en cada tipo de ataque.

No se hace clasificación de nodos sino de grafos.

No queda clara la contribución del grafo de ataque.

Se usan pocas mediciones, solo 5000.

Generar el grafo de ataque puede ser difícil en redes complejas.