Orthrus: Achieving High Quality of Attribution in Provenance-based Intrusion Detection Systems

Baoxiang Jiang, Tristan Bilot, Nour El Madhoun, Khaldoun Al Agha,
Anis Zouaoui, Shahrear Iqbal, Xueyuan Han, Thomas Pasquier

Xi'an Jiaotong University
USENIX SEC '25

Background: System Provenance

- Records interactions between system objects (e.g. processes, files, or sockets, etc.)

System Events

<Timestamp 1> Nginx, receive, IP1
<Timestamp 2> Nginx, open, index.html
<Timestamp 3> Nginx, read, index.html
<Timestamp 4> Nginx, send, IP1
<Timestamp 5> Nginx, close, Index.html
……

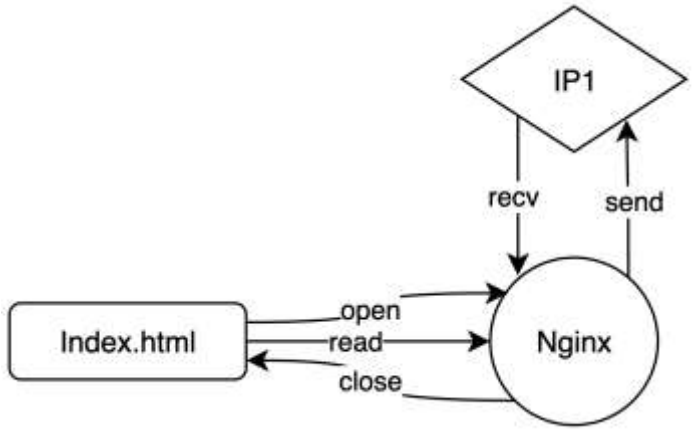
Background: System Provenance

- Records interactions between system objects (e.g. processes, files, or sockets, etc.)
- Represents system behaviors as a ***directed***, ***attributed*** and ***dynamic*** graph

System Events

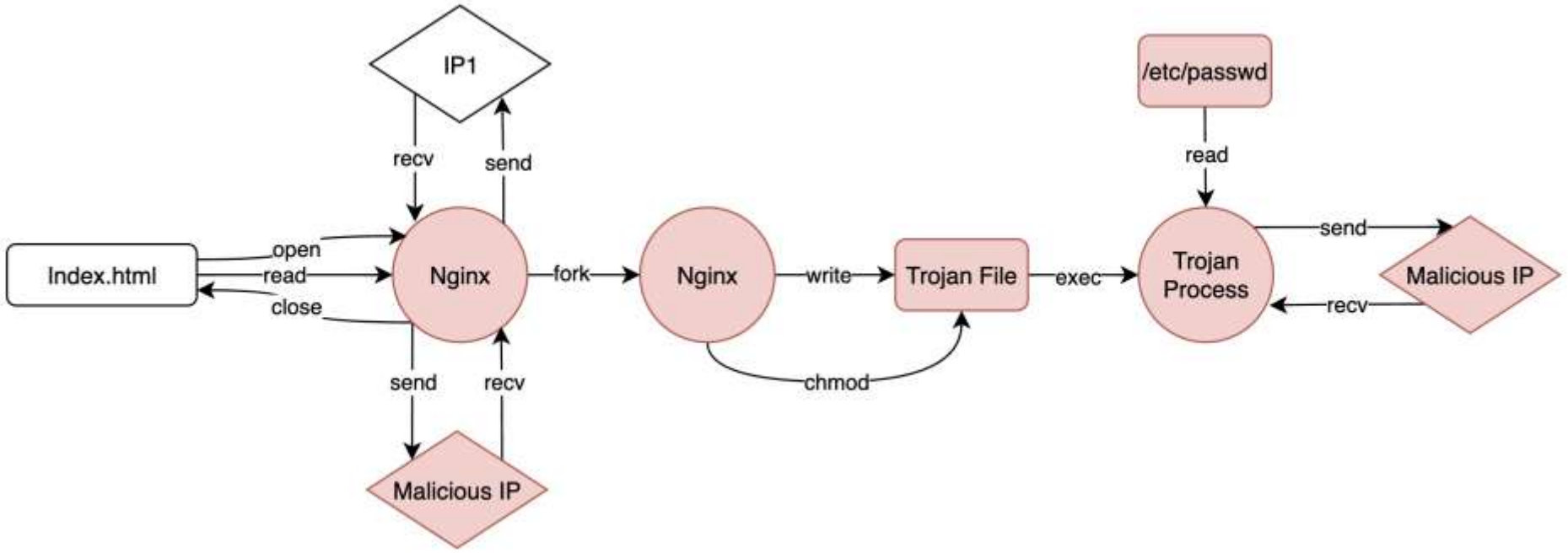
<Timestamp 1> Nginx, receive, IP1
<Timestamp 2> Nginx, open, index.html
<Timestamp 3> Nginx, read, index.html
<Timestamp 4> Nginx, send, IP1
<Timestamp 5> Nginx, close, Index.html
……

Provenance Graph



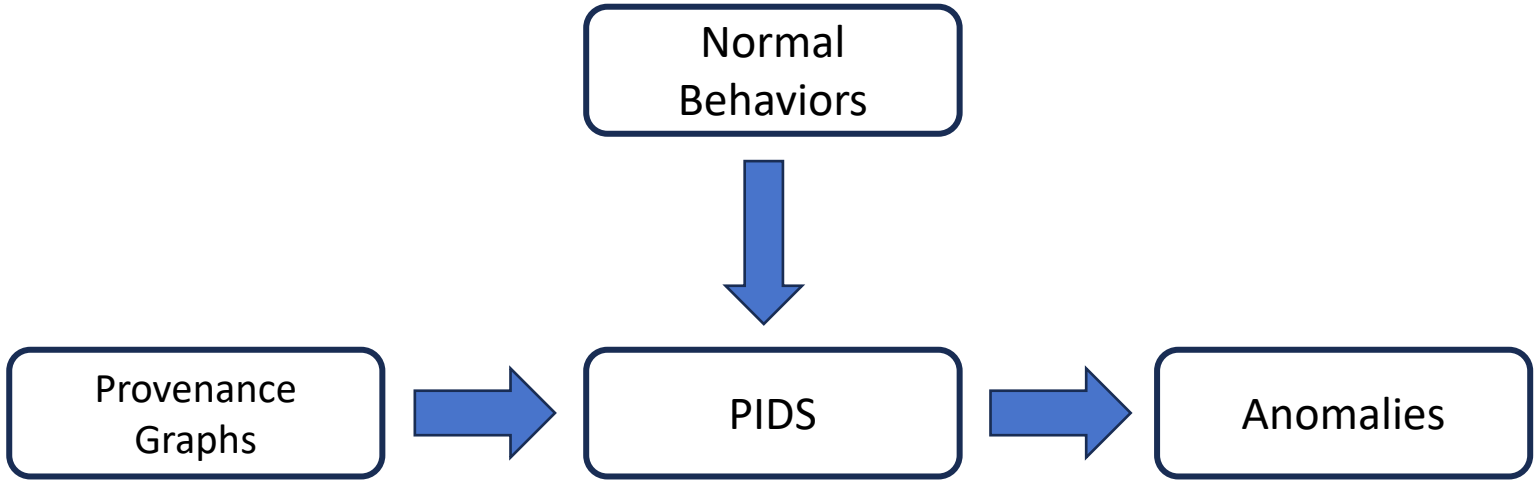
Background: System Provenance

- Records interactions between system objects (e.g. processes, files, or sockets, etc.)
- Represents system behaviors as a ***directed***, ***attributed*** and ***dynamic*** graph
- Cyber attacks → Abnormal system behaviors → Abnormal provenance graph structure



Background: Anomaly-based PIDS

- A Provenance-based Intrusion Detection System (PIDS) detects attacks in provenance graphs
- An anomaly-based PIDS aims to identify abnormal system behaviors as malicious



Motivation: Limitation in Signal Attribution Quality

SOTA systems report attack signals with low Quality of Attribution (QoA)

- Overwhelmingly contextual information causing alert fatigue and frequent burnout

Main factors:

- Significant data imbalance between classes
- Improper evaluation strategies

Motivation: Data Imbalance

Significant data imbalance between classes:

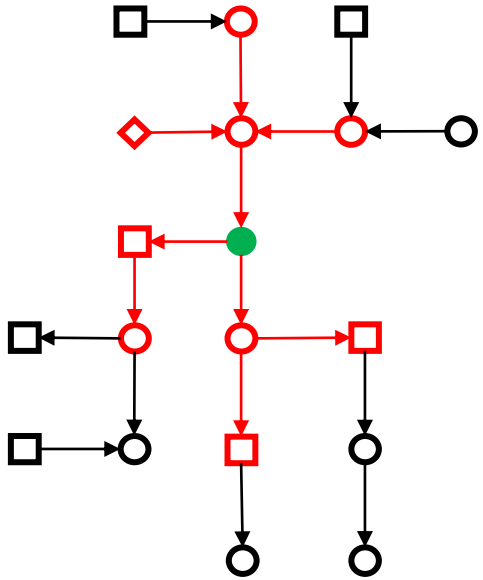
Prevalence of malicious nodes in the test set ranges from ~1:10,000 to ~1:1,000,000

Number of malicious nodes and total nodes in test set

Datasets	Test set	Malicious	Prevalence
E3-CADETS	268,153	68	2.5×10^{-4}
E3-THEIA	699,295	118	1.7×10^{-4}
E3-CLEARSCOPE	111,394	41	4.0×10^{-5}
E5-CADETS	3,111,378	123	4.0×10^{-5}
E5-THEIA	747,452	69	9.2×10^{-5}
E5-CLEARSCOPE	150,725	51	3.4×10^{-4}

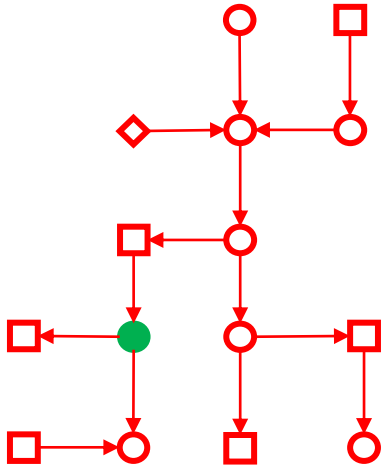
Motivation: Improper Evaluation Strategies

1) Neighborhood Approach



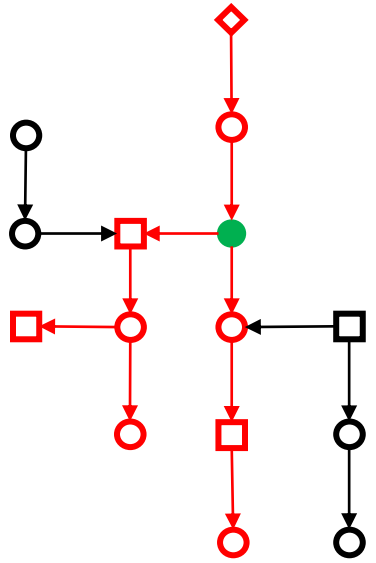
ThreaTrace (TIFS ’22), Flash (S&P ’24) and Magic (USENIX ’24) consider 2-hop neighbors of an attack node as malicious

2) Batch Approach



Kairos (S&P ’24) and EdgeTorrent (RAID ’23) consider all nodes in the same batch as malicious

3) Source Approach



R-CAID (S&P ’24) identifies the source node and all descendants are considered as malicious

Motivation: Improper Evaluation Strategies

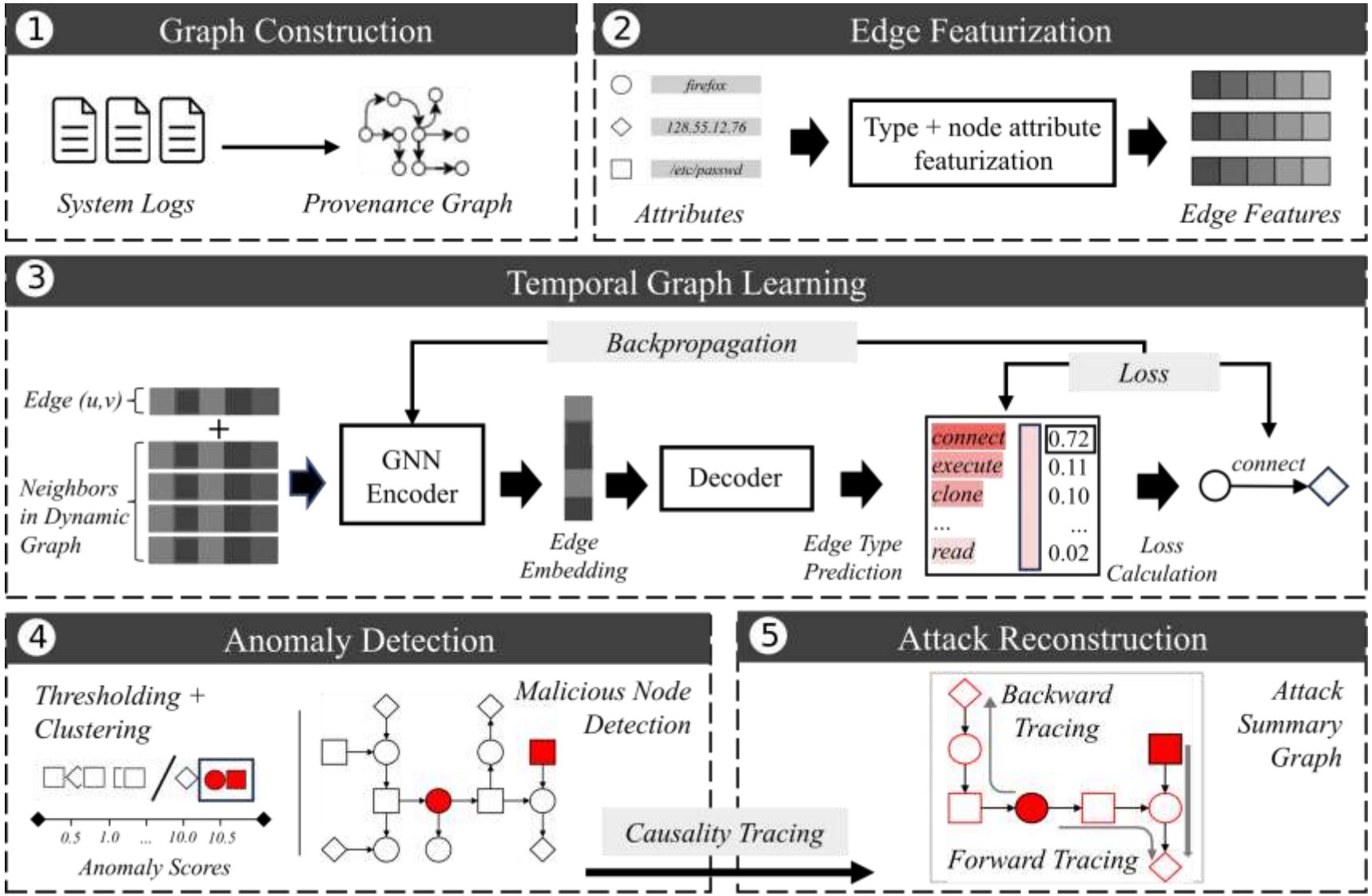
Difference between evaluation strategies:

- Evaluation strategies overestimate the number of malicious nodes
- Orthrus does **not** use any evaluation strategies

Number of malicious nodes with different evaluation strategies

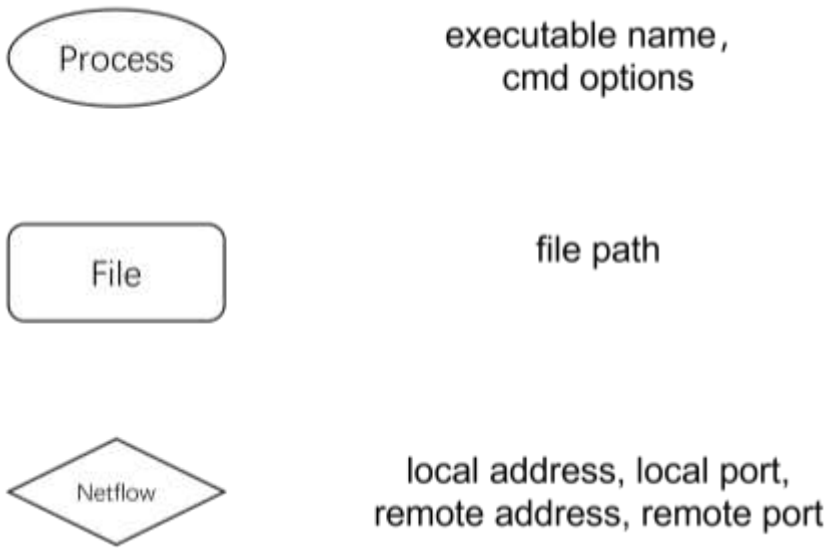
Datasets	Neighborhood	Batch	Source	Ours
E3-CADETS	12,852	4,929	2,062	68
E3-THEIA	25,362	51,098	35,794	118
E3-CLEARSCOPE	32,451	8,727	2,750	41
E5-CADETS	20,524	717,783	401,065	123
E5-THEIA	162,724	61,368	9,374	69
E5-CLEARSCOPE	48,488	8,636	1,020	51

Design: Overview of Orthrus

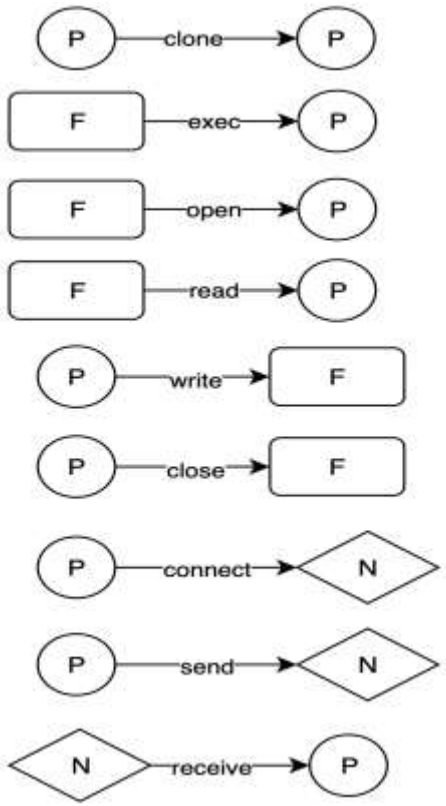


Design: Graph Construction

Considered System Objects

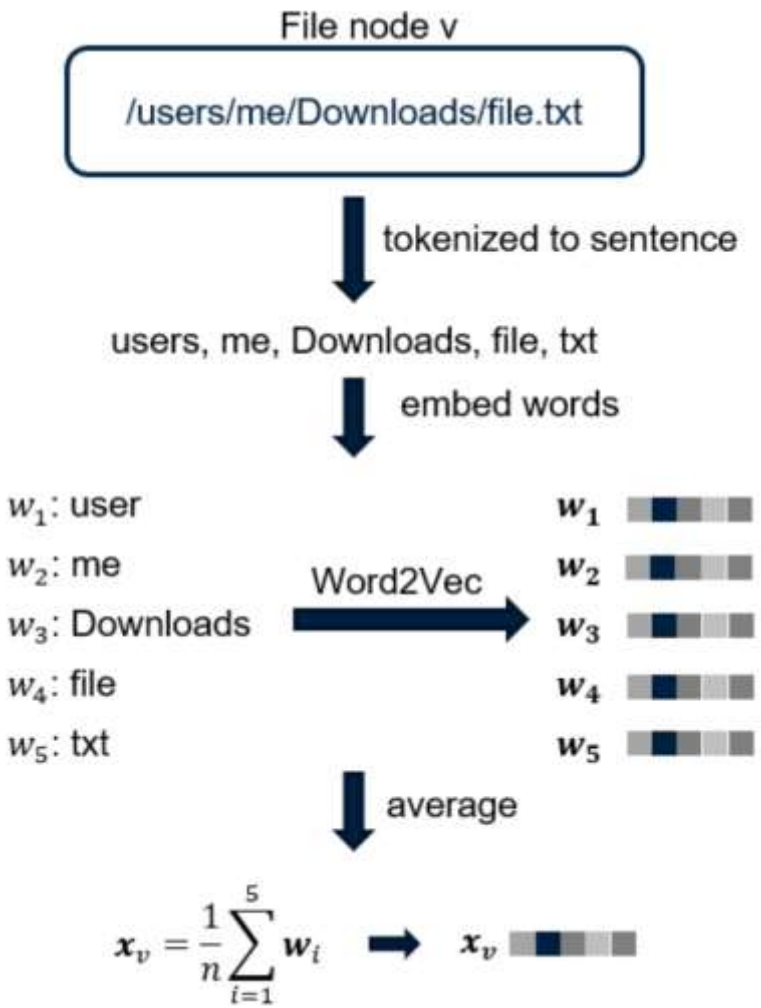


Considered System Events



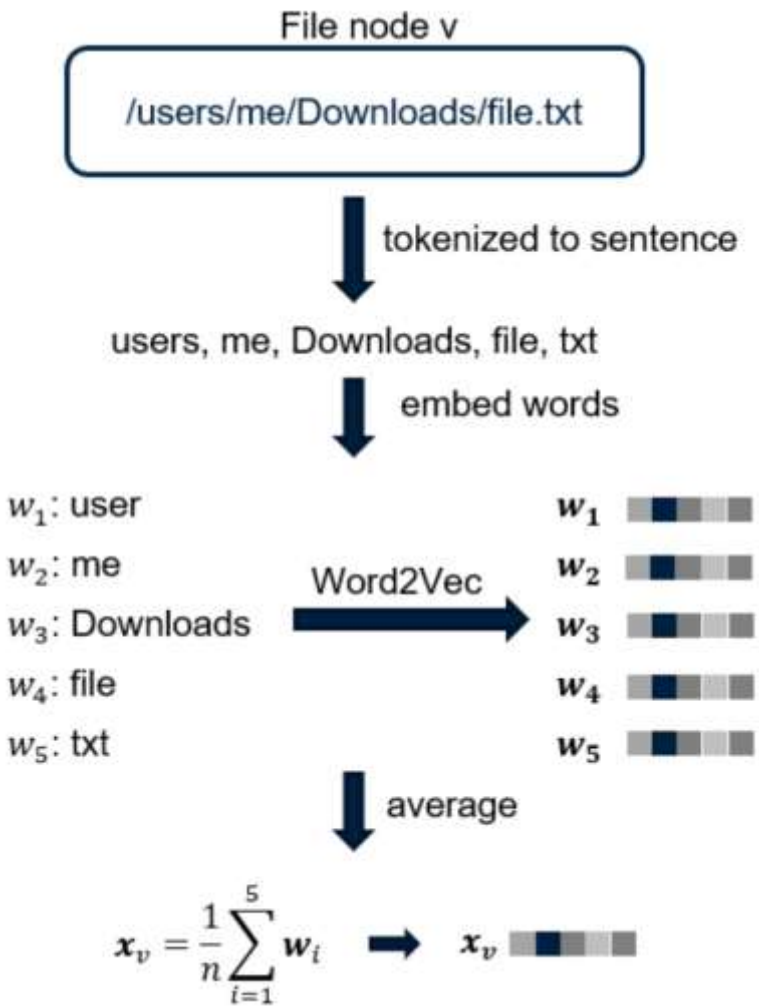
Design: Edge Featurization

1) Encode Node Attribute

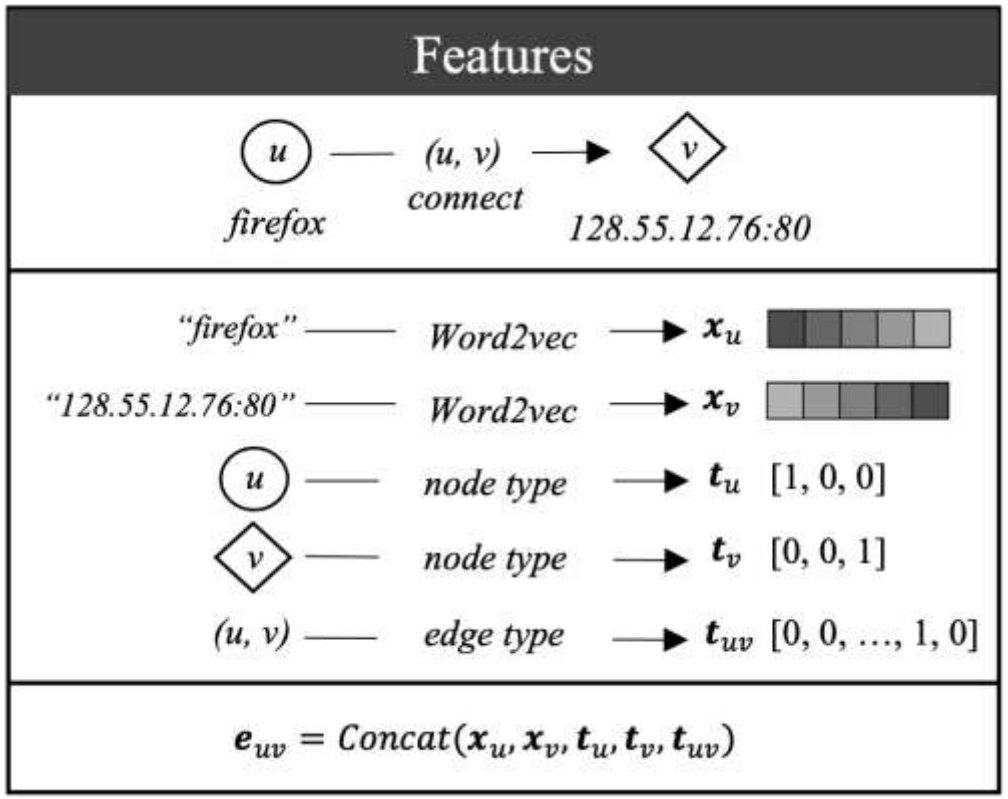


Design: Edge Featurization

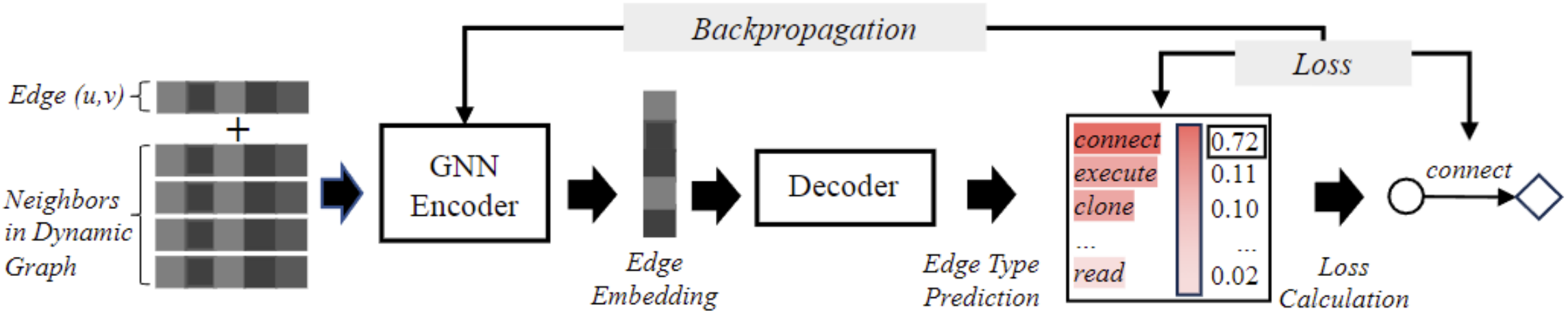
1) Encode Node Attribute



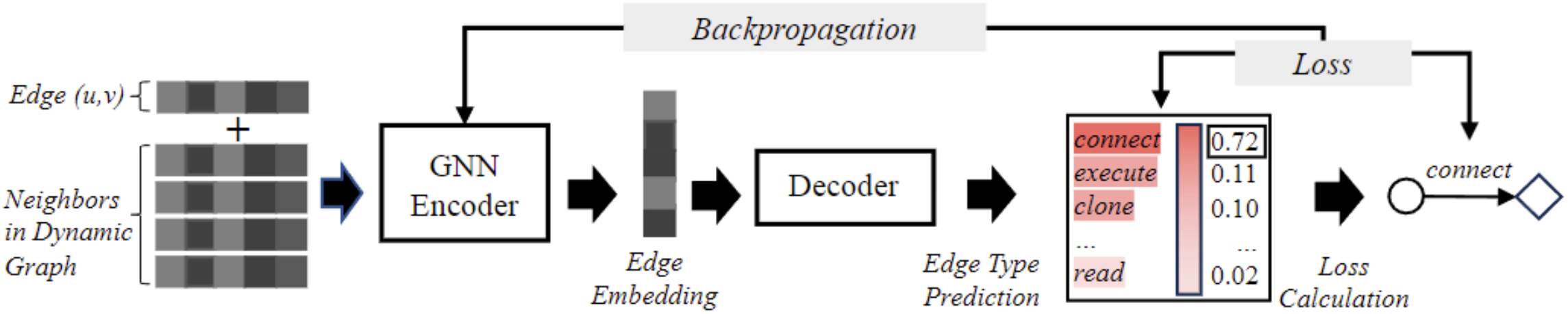
2) Compute Edge Feature Vector



Design: Graph Learning



Design: Graph Learning



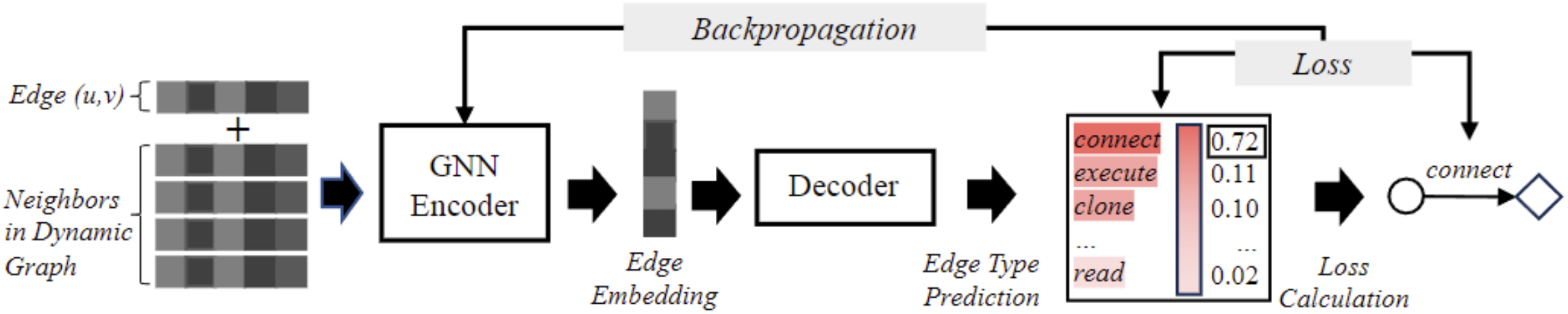
Encoder:

- For each edge (u, v), N most recent events connecting to v are sampled from previous events $S_t(v)$ for information aggregation

$$S_t(v) = \{(u, v) \in E | t_{uv} < t, t_{uv} \in T\}$$

$$S_N(v) = SAMPLE(S_t(v), N, T, t)$$

Design: Graph Learning

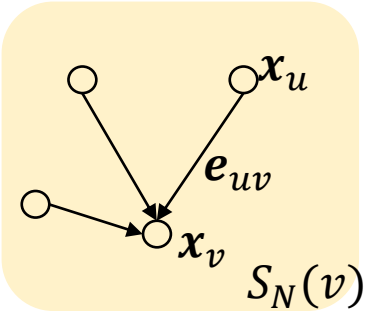


Encoder:

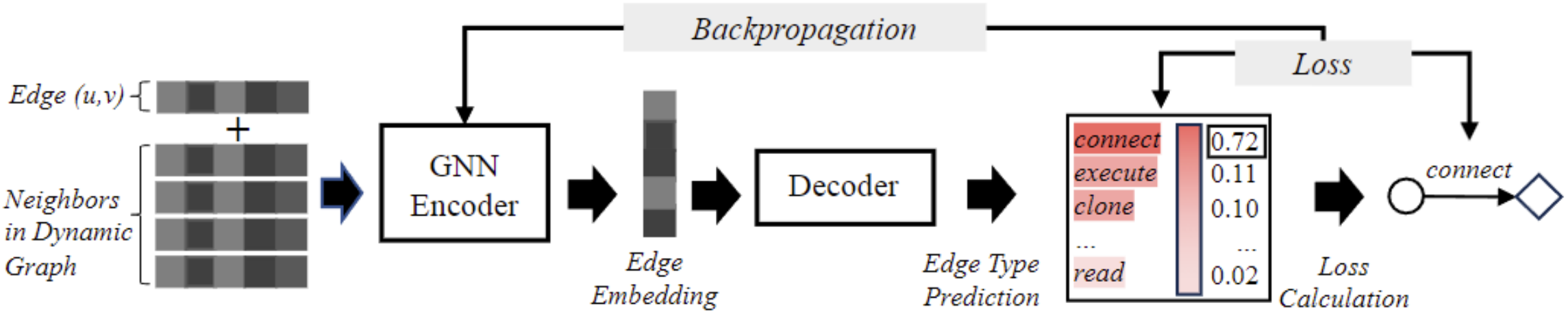
- Attention-based GNN aggregates information from sample events to target node v
- Attention coefficients are calculated between v and each sampled neighbor u

$$\boldsymbol{\alpha}_{u,v} = softmax(\frac{(\boldsymbol{W}_3\boldsymbol{x}_v)^T(\boldsymbol{W}_4\boldsymbol{x}_u + \boldsymbol{W}_5\boldsymbol{e}_{uv})}{\sqrt{d}})$$

$$\boldsymbol{h}_v = \boldsymbol{W}_1\boldsymbol{x}_v + \sum_{(u,v)\in S_N} \boldsymbol{\alpha}_{u,v}(\boldsymbol{W}_2\boldsymbol{x}_u + \boldsymbol{W}_3\boldsymbol{e}_{uv})$$



Design: Graph Learning

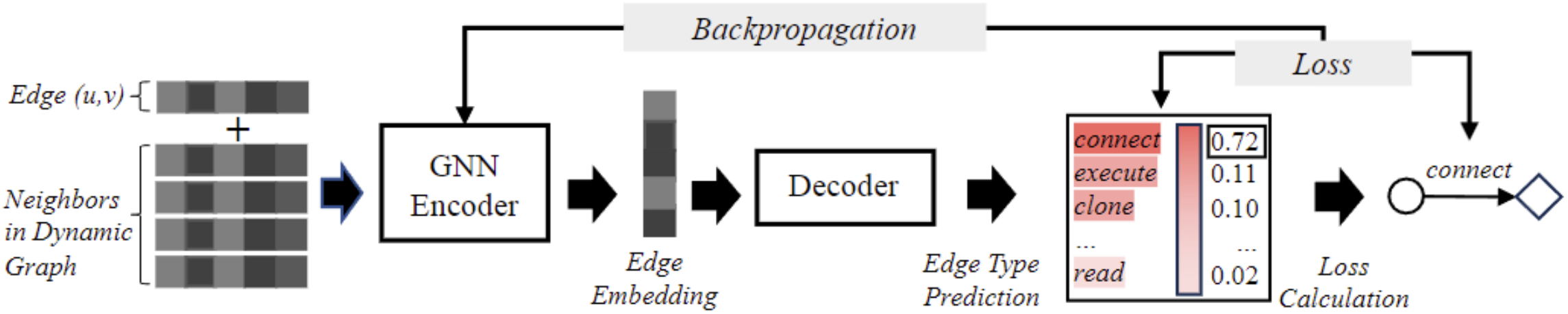


Decoder:

- Predicts edge type $\hat{y}_{uv}$ based on embeddings of end nodes $\boldsymbol{h}_u$ and $\boldsymbol{h}_v$

$$\hat{y}_{uv} = \sigma(\boldsymbol{W}_g \cdot Concat(\boldsymbol{W}_s \boldsymbol{h}_u, \boldsymbol{W}_d \boldsymbol{h}_v))$$

Design: Graph Learning



Decoder:

- Predicts edge type $\hat{y}_{uv}$ based on embeddings of end nodes $\boldsymbol{h}_u$ and $\boldsymbol{h}_v$

$$\hat{y}_{uv} = \sigma(\boldsymbol{W}_g \cdot Concat(\boldsymbol{W}_s \boldsymbol{h}_u, \boldsymbol{W}_d \boldsymbol{h}_v))$$

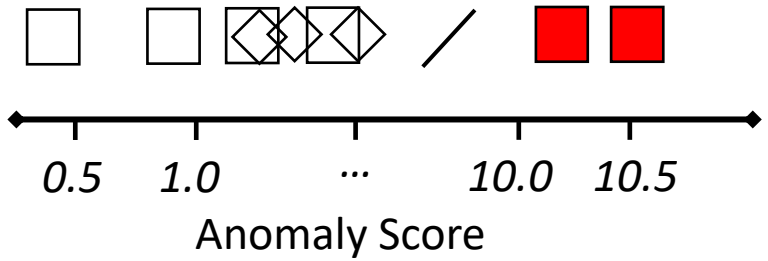
- The reconstruction error is set as the Cross-Entropy loss across edge types to prediction

$$L_{uv} = CE(\hat{y}_{uv}, y_{uv})$$

Design: Anomaly Detection

1) Automatic Thresholding:

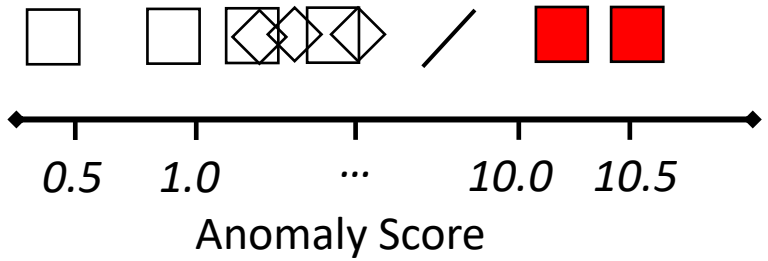
- Max anomaly score in validation set as the threshold
- Indicates the largest deviation among benign activities



Design: Anomaly Detection

1) Automatic Thresholding:

- Max anomaly score in validation set as the threshold
- Indicates the largest deviation among benign activities

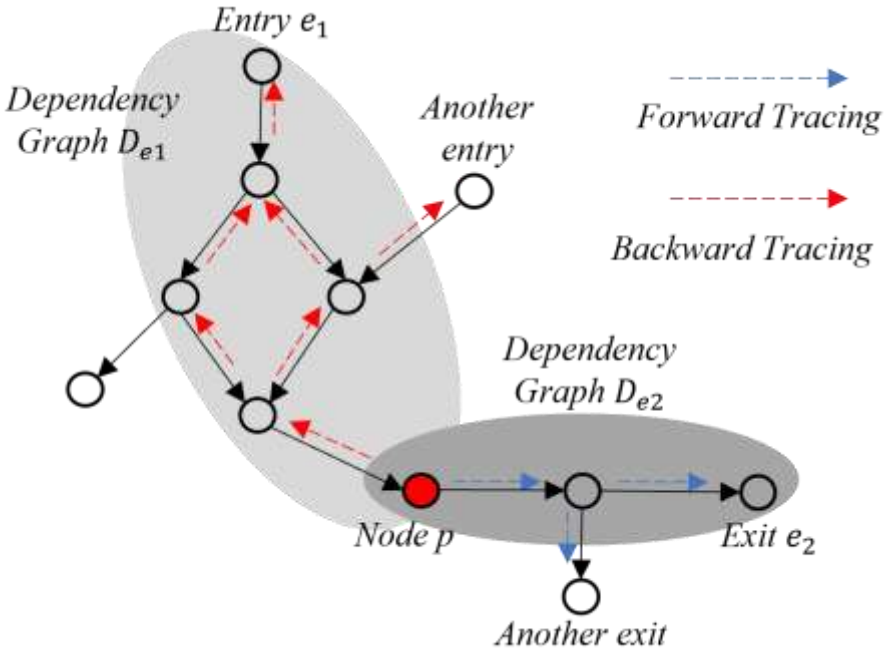


2) Outlier Clustering:

- K-means is used to divide over-threshold nodes into 2 clusters
- The more abnormal cluster is reported as malicious, the other is filtered as benign malicious

Design: Attack Reconstruction

- Conduct causality analysis to identify potential attack *entry* and *exit* nodes
- A *dependency graph* is defined as the subgraph between an entry/exit node and p
- Calculate critical scores for dependency graphs to identify the most critical dependency graph as reconstruction of the attack



Evaluation Questions

- RQ1: Is Orthrus able to detect all attacks?
- RQ2: What is the quality of attribution?
- RQ3: Is Orthrus computationally efficient?
- RQ4: How do hyperparameters influence performance?
- RQ5: How the different Orthrus components contribute to overall performance?

Evaluation: Datasets

- Benchmark datasets published by DARPA's Transparent Computing (TC) programs.
- TC organized several adversarial engagements that simulated real-world APTs on enterprise networks.
  - Simulation Duration: two weeks
  - Benign activities: browse website, check emails, SSH connection, etc.
  - Attack activities: browser vulnerability exploitation, malicious process execution, sensitive data leakage

Datasets	Training	Validation	Test	Total	Neigh.	Batch	Source	Ours	Prevalence
E3-CADETS	449,325	40,581	268,153	758,059	12,852	4,929	2,062	68	2.5×10^{-4}
E3-THEIA	410,023	34,365	699,295	1,143,683	25,362	51,098	35,794	118	1.7×10^{-4}
E3-CLEARSCOPE	132,121	797	111,394	244,312	32,451	8,727	2,750	41	3.7×10^{-4}
E5-CADETS	3,275,875	1,245,539	3,111,378	7,632,792	20,524	717,783	401,065	123	4.0×10^{-5}
E5-THEIA	745,773	234,896	747,452	1,728,121	162,714	61,368	9,374	69	9.2×10^{-5}
E5-CLEARSCOPE	171,771	3,842	150,725	326,338	48,488	8,636	1,020	51	3.4×10^{-4}

Evaluation: Baselines

- 5 state-of-the-art works are taken for comparison: Flash, MAGIC, Kairos, ThreaTrace and SIGL

	2016	2017	2020		2021	2022		2024		
	StreamSpot [44]	Frappuccino [28]	Unicorn [27]	ProvDetector [65]	Prov-Gem [37]	SIGL [29]	ThreaTrace [66]	Kairos [18]	MAGIC [36]	Flash [57]
ThreaTrace	✓		✓	✓						
SIGL	✓	✓								
Kairos			✓				✓			
MAGIC	✓		✓		✓		✓			
Flash			✓				✓			
Orthrus						✓	✓	✓	✓	✓

RQ1: Is Orthrus able to detect all attacks?

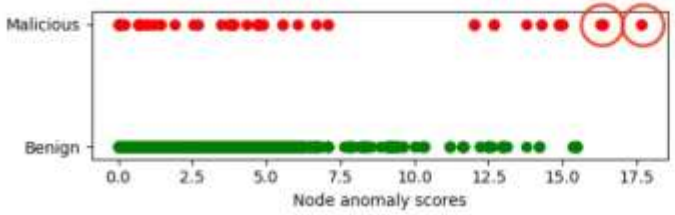
Dataset	System	E3	E5
CADETS	ORTHRUS	✓ 3/3	✓ 2/2
	Kairos	✗ 0/3	✗ 0/2
	ThreaTrace	✓ 3/3	✓ 2/2
	SIGL	✗ 0/3	✗ 0/2
	MAGIC	✓ 3/3	✓ 2/2
	Flash	✓ 3/3	✓ 2/2
THEIA	ORTHRUS	✓ 2/2	✓ 1/1
	Kairos	~ 1/2	✗ 0/1
	ThreaTrace	✓ 2/2	✓ 1/1
	SIGL	~ 1/2	✗ 0/1
	MAGIC	✓ 2/2	✓ 1/1
	Flash	✓ 2/2	✓ 1/1
CLEARSCOPE	ORTHRUS	✓ 1/1	✓ 3/3
	Kairos	✗ 0/1	~ 1/3
	ThreaTrace	✓ 1/1	✓ 3/3
	SIGL	✓ 1/1	~ 2/3
	MAGIC	✓ 1/1	✓ 3/3
	Flash	✗ 0/1	✓ 3/3

- An attack is considered detected if the system flags any node directly involved in the attack as malicious

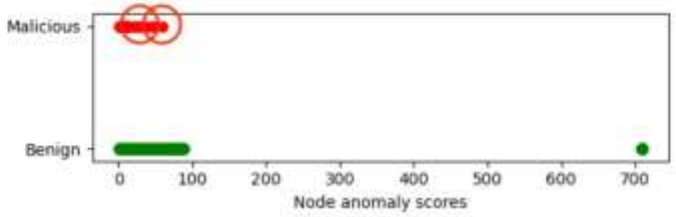
RQ2: What is the quality of attribution?

Dataset	System	TP	FP	TN	FN	Precision	MCC	Training Time	GPU Memory
E3-CADETS	ORTHRUS-full	25	23	268,062	43	0.52	**0.44**	**4min40**	**3.82GB**
	ORTHRUS-ano	10	0	268,085	58	**1.00**	0.38		
	Kairos	0	9	268,076	68	0.00	0.00	22min49	3.93GB
	Threatrace	61	252,117	15,968	7	0.00	0.00	28min28	5.22GB
	SIGL	0	80	268,005	68	0.00	0.00	4h48	10.07GB
	MAGIC	63	79,766	188,319	5	0.00	0.02	13h18	4.22GB
	Flash	13	2,381	265,704	55	0.01	0.03	10h33	19.18GB
E3-THEIA	ORTHRUS-full	48	11	699,166	70	0.81	**0.57**	**3min58**	**2.03GB**
	ORTHRUS-ano	8	0	699,177	110	**1.00**	0.26		
	Kairos	4	0	699,177	114	**1.00**	0.18	24min21	2.53GB
	Threatrace	88	671,883	27,294	30	0.00	-0.01	10min19	4.51GB
	SIGL	1	29	699,148	117	0.03	0.02	14h07	10.44GB
	MAGIC	115	394,906	304,271	3	0.00	0.01	11h39	5.35GB
	Flash	22	32,082	667,095	96	0.00	0.01	6h51	36.93GB
E3-CLEARSCOPE	ORTHRUS-full	2	6	111,347	39	0.25	**0.11**	**2min50**	**0.65GB**
	ORTHRUS-ano	1	1	111,352	40	**0.50**	**0.11**		
	Kairos	0	7	111,346	41	0.00	0.00	9min52	0.74GB
	Threatrace	41	87,501	23,852	0	0.00	0.01	3min55	4.90GB
	SIGL	1	11,372	99,981	40	0.00	0.00	1h01	9.71GB
	MAGIC	40	101,737	9,616	1	0.00	0.00	1h37	9.75GB
	Flash	0	15,137	96,216	41	0.00	-0.01	19h01	11.60GB

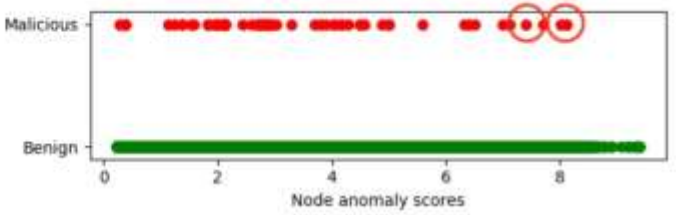
RQ2: What is the quality of attribution?



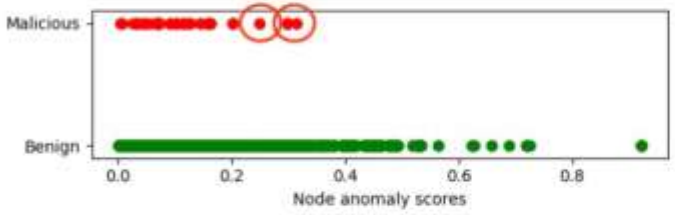
(a) ORTHRUS.



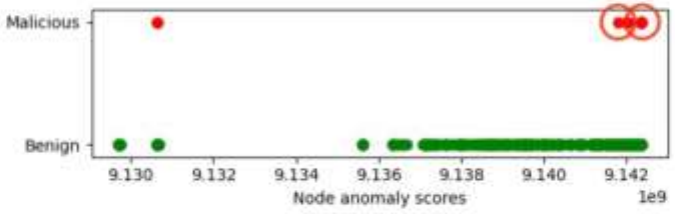
(b) Threatrace.



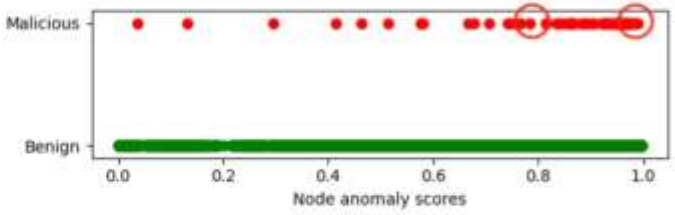
(c) Kairos.



(d) SIGL.



(e) Magic.



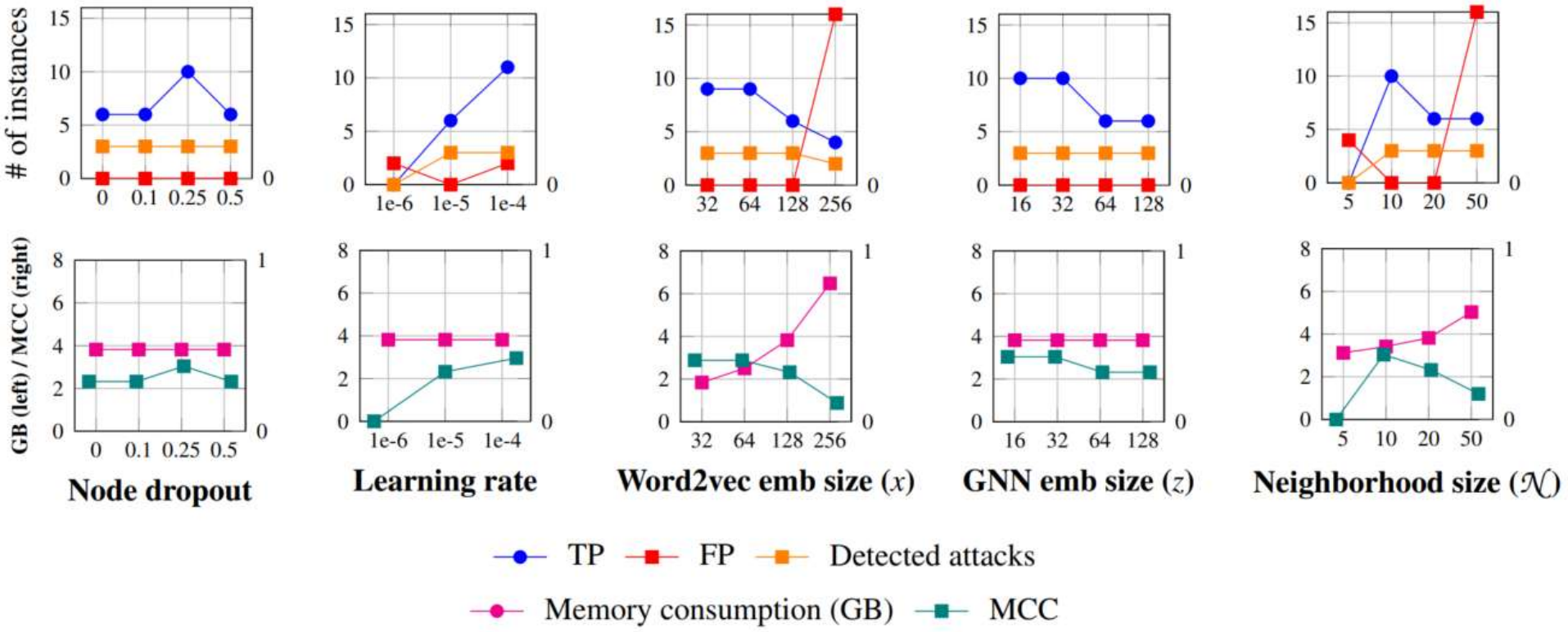
(f) Flash.

- Orthrus differentiates malicious and benign nodes with a larger margin.

RQ3: Is Orthrus computationally efficient?

Dataset	System	TP	FP	TN	FN	Precision	MCC	Training Time	GPU Memory
E5-CADETS	ORTHRUS-full	2	10	3,111,245	121	**0.17**	**0.05**	**42min35**	21.10GB
	ORTHRUS-ano	1	5	3,111,250	122	**0.17**	0.04		
	Kairos	0	6	3,111,249	123	0.00	0.00	4h03	23.85GB
	Threatrace	91	3,104,018	7,237	32	0.00	-0.03	5h45	**17.31GB**
	SIGL	0	66	3,111,189	123	0.00	0.00	38h00	22.72GB
	MAGIC	123	3,110,714	541	0	0.00	0.00	77h13	79.36GB
	Flash	45	33,941	3,077,314	78	0.00	0.08	101h26	80.19GB
E5-THEIA	ORTHRUS-full	13	2	747,381	56	0.87	0.4	**14min30**	4.23GB
	ORTHRUS-ano	2	0	747,383	67	**1.00**	**0.17**		
	Kairos	0	2	747,381	69	0.00	0.00	1h02	**4.16GB**
	Threatrace	66	739,322	8,061	3	0.00	0.00	2h51	11.59GB
	SIGL	0	23	747,360	69	0.00	0.00	40h20	24.44GB
	MAGIC	1	296,554	450,829	68	0.00	-0.01	13h21	16.95GB
	Flash	43	295,729	451,654	26	0.00	0.00	47h50	80.18GB
E5-CLEARSCOPE	ORTHRUS-full	4	8	150,666	47	**0.33**	**0.16**	**22min19**	**1.72GB**
	ORTHRUS-ano	2	7	150,667	49	0.22	0.09		
	Kairos	1	3	150,671	50	0.25	0.07	1h06	2.26GB
	Threatrace	41	142,487	8,187	10	0.00	-0.01	44min53	5.94GB
	SIGL	10	63	150,610	41	0.14	0.16	82h50	16.38GB
	MAGIC	51	139,385	11,289	0	0.00	0.01	11h39	48.24GB
	Flash	15	4,552	146,122	36	0.00	0.03	25h34	11.60GB

RQ4: How do hyperparameters influence performance?



- Orthrus can detect all 3 attacks in most cases, even if the parameter change a lot
- Results demonstrate the robustness of Orthrus

RQ5: Contribution of components to overall performance?

Dataset	Featurization	Encoding	Clustering	Reconstruction	TP	FP	Precision	Memory
E3-THEIA	×	✓	✓	✓	51	13	0.79	2.03GB
	✓	×	✓	✓	41	772	0.05	5.75GB
	✓	✓	×	✓	48	11	0.81	2.03GB
	✓	✓	✓	×	8	0	1.00	2.03GB
	✓	✓	✓	✓	48	11	0.81	2.03GB
E5-THEIA	×	✓	✓	✓	0	155	0.00	4.23GB
	✓	×	✓	✓	13	53	0.20	11.10GB
	✓	✓	×	✓	20	11,420	0.00	4.23GB
	✓	✓	✓	×	2	0	1.00	4.23GB
	✓	✓	✓	✓	13	2	0.87	4.23GB

- **Ablation Study**: we replace or remove one component at a time

Summary

1. We identify the main factors leading to low attribution quality of SOTA PIDS
2. We propose Orthrus, the first PIDS capable of providing high-attribution-quality attack reports
3. We generate and open-source a comprehensive and solid ground truth without any improper evaluation strategies
4. We open-source our system and evaluate it using publicly available datasets

Thanks for your listening!