CE 815 – Secure Software Systems

Causal Analysis (Benchmark)

Mehdi Kharrazi
Department of Computer Engineering
Sharif University of Technology

Acknowledgments: Some of the slides are fully or partially obtained from other sources. A reference is noted on the bottom of each slide, when the content is fully obtained from another source. Otherwise a full list of references is provided on the last slide.

Are we there yet? An Industrial Viewpoint on Provenance-based Endpoint Detection and Response Tools, F. Dong, S. Li, P. Jiang, D. Li, H. Wang, L. Huang, X. Xiao, J. Chen, X. Luo, Y. Guo, CCS 2023.

EDR Systems

- EDR systems are cybersecurity tools designed for continuous monitoring of endpoints.
- They detect, investigate, and respond to security threats across workstations, servers, and mobile devices.
- They collect extensive data from endpoints, including process activities, network connections, and file changes.
- Data analysis involves behavioral analysis, machine learning, and integration of threat intelligence.
- Aimed at early detection of potential security incidents and anomalies.

P-EDR

- P-EDR as a next-generation system for APT attack defense by using a provenance graph for modeling dependencies between system activities.
- Superiority over conventional EDR systems in detection accuracy and interpretability.
- Rapid growth of P-EDR research and industry adoption noted in recent years.
- Study objectives: Assessing effectiveness and limitations of P-EDR systems.
- The paper's study includes interviews, questionnaires, literature surveys, and measurement studies to evaluate P-EDR systems.

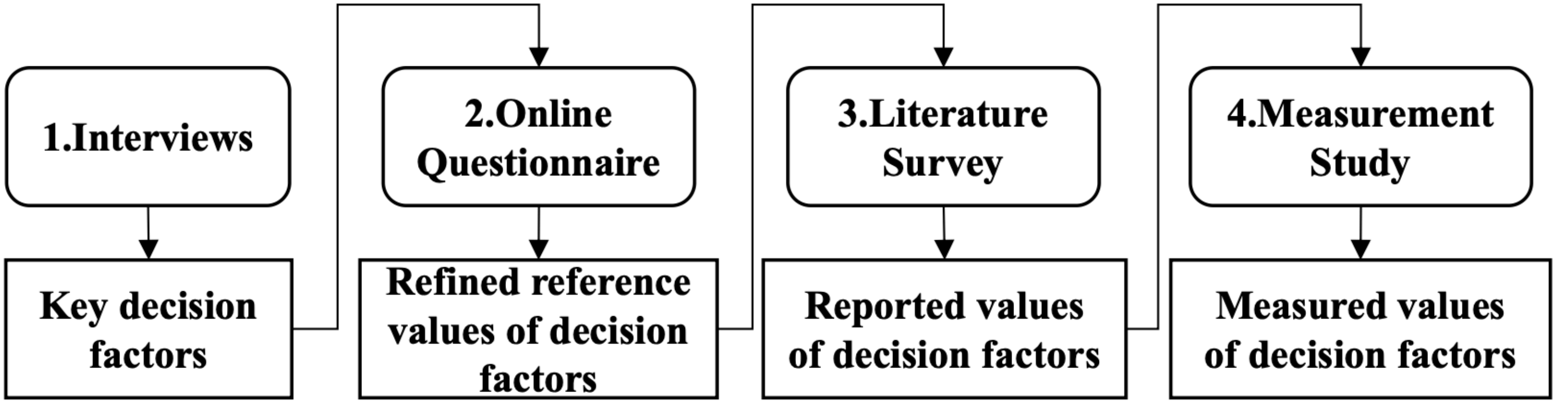
Research Questions

- RQ1: How does the industry compare the effectiveness of P-EDR and conventional EDR?
- RQ2: What are the bottlenecks for the industry to adopt EDR Systems?
- RQ3: How well can existing P-EDR systems proposed in academia meet the expectations of the industry?

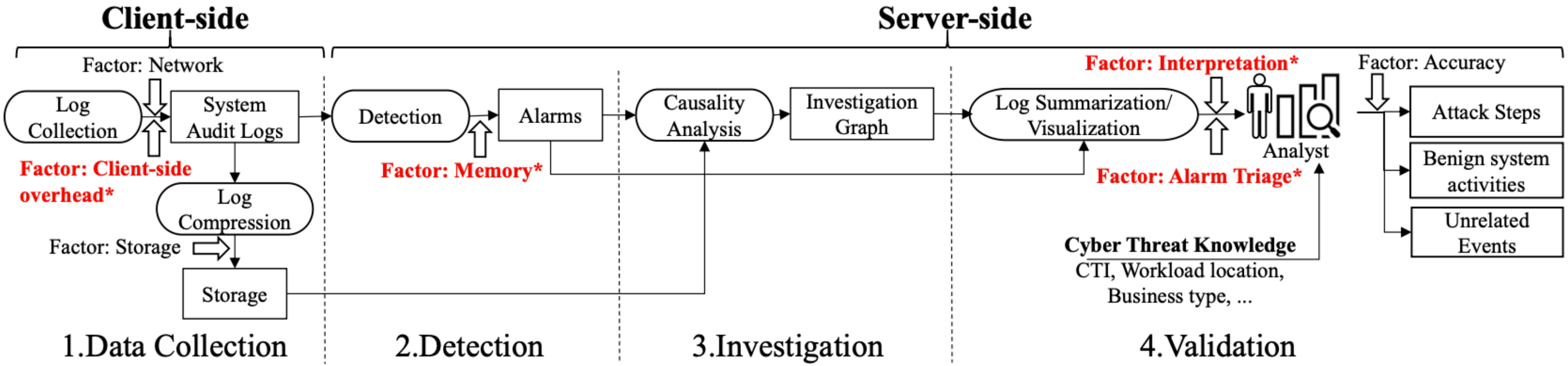
Methodology of Industrial Viewpoint Study on P-EDR

- One-to-one interviews with technical managers from top IT companies.
- Online questionnaire for feedback from a broader scope of security engineers.
- Literature survey of recent publications on P-EDR systems.
- Focus on evaluating effectiveness, limitations, and decision factors for P-EDR adoption.

Overview


1.Interviews
Key decision factors
2.Online Questionnaire
Refined reference values of decision factors
3.Literature Survey
Reported values of decision factors
4.Measurement Study
Measured values of decision factors

Overview of P-EDR systems



One-to-one interview

- 5 EDR vendors from top-tier endpoint security companies
- 5 consumers of EDR systems from diverse kinds of organizations
- participants are experienced in security: on average, 10.5 years of experience

ID	Role	Company Name	Industry Area	Job Title	Years of Exp.	Team Size	Adopt P-EDR
E1	Consumer	ByteDance	Technology	Head of Server Security	6	20~25	Yes
E2		MeiTuan	Technology	Cloud Workload Security Leader	5	20~25	Yes
E3		Peking University	Education	Director of Network Security Office	19	10~15	No
E4		S.F. Express	Transportation	Endpoint Security Manager	10	20~25	No
E5		FiberHome	Manufacturing	Endpoint Security Manager	8	5~10	No
E6	Vendor	Tencent Security	Security	Director of EDR	10	10~15	Yes
E7		Trend Micro	Security	Detection Engine Architect of EDR	9	20~25	Yes
E8		Sangfor	Security	Director of Workload Protection Product	8	65~70	No
E9		Rising	Security	EDR Architect	21	50~55	No
E10		NSFOCUS	Security	EDR Product Manager	9	30~35	No

Interview Feedback

Answers	Participants
Limitations of EDR/P-EDR	
High Client-Side Overhead	E1, E2, E3, E4, E5, E6, E7, E8, E9, E10
Too Many False Alarms	E1, E2, E4, E5, E6, E7, E8
Incomplete Rule Set	E1, E2, E4, E5, E7, E9, E10
Data Privacy	E3
Effectiveness of P-EDR	
P-EDR Already Deployed	E1, E2, E6, E7
P-EDR Better Than EDR	E1, E2, E3, E4, E5, E6, E7, E8, E9, E10

Seven key factors

- Average number of nodes of provenance graphs of alarms as the metric for the interpretation cost.
- Preferred to use the average number of alarms per monitored host per day to evaluate the performance rather than using precision

Factor	Description
Computing Cost	
CC1: Client-Side Overhead	how much an EDR system slows down the protected hosts
CC2: Network	bandwidth occupied by transmitting system audit logs to the server
CC3: Storage	hard-disk used to store the system logs
CC4: Memory	server memory size required to analyze the collected logs
Labor Cost	
LC1: Alarm Triage	man-hour required to detect false alarms
LC2: Interpretation	man-hour required to interpret attack results
Performance	
Accuracy	attack detection accuracy

Interview results for key decision factors

	Computing Cost				Labor Cost		Performance
ID	**Network**	**Storage**	**Memory***	**Client-Side Overhead***	**Interpretation Cost***	**Alarm Triage Cost***	**Accuracy**
E1	None	None	3, ServerMem*: 30MB/host	2, ClientMem*: 100MB/host, RT OH*:1%	4, Number of nodes*: 100	1, Alarms*: 0.001/day/host	None
E2	None	None	3, ServerMem*: 50MB/host	1, ClientMem*: 150MB/host, RT OH*:5%	4, Number of nodes*: 10	2, Alarms*: 0.001/day/host	5, Precision, > 0.85
E3	None	3, Disk: 60MB/day/host	2, ServerMem*: 30MB/host,	1, ClientMem*: 100MB/host, RT OH*:5%	None	None	5, Precision, > 0.9
E4	None	None	3, ServerMem*: 50MB/host,	1, ClientMem*: 200MB/host, RT OH*:10%	None	2, Alarms*: 0.004/day/host	None
E5	None	None	3, ServerMem*: 30MB/host,	1, ClientMem*: 100MB/host, RT OH*:5%	None	2, Alarms*: 0.02/day/host	None
E6	5, Net: 100MB/day/host	6, Disk: 15MB/day/host	3, ServerMem*: 30MB/host,	1, ClientMem*: 200MB/host, RT OH*:1%	4, Number of nodes*: 100	2, Alarms*: 0.1/day/host	None
E7	5, Net: 10MB/day/host	6, Disk: 70MB/day/host	3, ServerMem*: 20MB/host,	1, ClientMem*: 50MB/host, RT OH*:5%	4, Number of nodes*: 100	2, Alarms*: 0.1/day/host	None
E8	5, Net: 42MB/day/host	4, Disk: 100MB/day/host	3, ServerMem*: 26MB/host,	2, ClientMem*: 250MB/host, RT OH*:5%	None	1, Alarms*: 0.05/day/host	None
E9	4, Net: 1MB/day/host	3, Disk: 15MB/day/host	2, ServerMem*: 10MB/host,	1, ClientMem*: 150MB/host, RT OH*:10%	None	None	None
E10	4, Net: 100MB/day/host	5, Disk: 35MB/day/host	3, ServerMem*: 30MB/host,	1, ClientMem*: 100MB/host, RT OH*:5%	None	2, Alarms*: 0.1/day/host	None
Reference Range	1~100MB /day/host	15~100MB day/host	10~50MB/host	50~250MB/host, 1~10%	10~100	0.001~0.1 /day/host	> 0.85

Online questionnaire (37 responses)

- Design the questionnaire based on the results from the interview
- Four must-meet factors: Memory, Client-Side Overhead, Interpretation, and Triage.
- Divide the reference range obtained in the interviews into five equal-sized

Must-meet Factors	Summarized Result
Memory	< 20 MB/host
Client-side Overhead (RT OH)	< 3 %
Client-side Overhead (ClientMem)	< 100 MB/host
Interpretation	< 50 nodes
Alarm Triage	< 0.1 alarms/day/host

Literature Survey

- Selected 20 papers on P-EDR systems 2017-2022
  - Rule-based approaches
  - Anomaly-based approaches
  - Investigation approaches
- Look into whether they have been evaluated against the decision factors

Summarization of Literature Survey

Type	Tool Name	Client-side Overhead			Storage (/MB/host/day)	Memory (MB/host)	Alarm Triage (#Alarm/host/day)	Interpretation (#Node, #Edge)	Precision	Recall	Accuracy
		Agent	RT OH(%)	ClientMem (MB)							
Detection	SLEUTH [35]	Auditd	-	-	362.87	81.93	-	(52, -)	-	-	-
	MORSE [37]	Auditd, DTrace	-	-	1266.67	230.4	-	(283, -)	≈ 0	1.00	-
	HOLEMS [60]	Auditd, Dtrace, ETW	-	-	179.23	104.76	-	(-, 400)	1.00	1.00	1.00
	RapSheet [31]	Symantec EDR	-	-	358.00	-	-	(12, 39)	0.26	1.00	0.75 - 0.95
	Pagoda [83]	Karma [19], PASS [62]	-	-	1126.40	-	-	(13315, 10964)	0.92-1.00	1.00	0.75 - 0.95
	StreamSpot [56]	SystemTap [41]	-	-	-	-	-	(8315,173857)	0.50-1.00	-	0.50 - 0.80
	UNICORN [29]	CamFlow [65]	-	-	24917.33	-	-	$(1.76 \times 10^5, 2.82 \times 10^6)$	0.80 - 0.99	0.88 - 1.00	0.84 - 0.99
	ProvDetector [81]	-	-	-	-	-	-	(-, -)	0.96	0.99	-
	ZePro [75]	-	-	-	266.67	57.14	-	(1853, 2249)	-	-	-
	P-Gaussian [84]	-	-	-	864	152.5	-	(1949, 3045)	-	0.66 - 0.94	0.65 - 0.95
	Poirot [59]	Auditd, Dtrace, ETW	-	-	6500.55	122.39	-	(-, -)	1.00	1.00	1.00
	SHADEWATCHER [89]	Auditd	-	-	59112.73	4194.30	-	(-, -)	0.86 - 1.00	0.95 - 1.00	0.98 - 1.00
Investigation	RTAG [43]	RAIN	4.84	-	1536 - 4096	-	-	(164.67, 3200)	-	-	1.00
	MCI [46]	Auditd, Dtrace, ETW	-	-	-	-	-	(34.56, 62.87)	0.92- 1.00	0.95 - 1.00	-
	PrioTracker [52]	Auditd, ETW	-	-	998.64	-	-	(-, 1219)	-	-	-
	NoDoze [33]	Auditd, ETW	-	-	428.90	-	-	(14, 14)	0.50	1.00	0.86
	ATLAS [15]	-	-	-	2286.93	-	-	(-, -)	0.91	0.97	0.99
	DEPCOMM [85]	Sysdig	-	-	-	-	-	(289, -)	-	-	-
	DEPIMPACT [26]	Sysdig	-	-	-	-	-	(-, 234.27)	0.79 - 0.85	1.00	-
	RAPID [51]	Auditd, Dtrace, ETW	-	-	4743.40	30.04	-	(-, -)	-	-	-

Summarization of Literature Survey

- Alarm Triage: None of the papers provide evaluation. Thus, even though they can achieve high accuracy the triage costs are usually not acceptable in practice.
- Rule-based systems, can generate smaller provenance graphs for alarms than anomaly-based systems
- Memory: reported values are much higher than the reference values (< 20MB/host)
- Only a small set of papers provide evaluations for part of the four factors & fail to satisfy the reference values

Summarization of Literature Survey

- None of the existing provenance collectors can satisfy the reference value of runtime overhead (< 3%).

	Platform	Owner	Affect	RT OH (%)	Mem (MB)
Sysdig [17]	Linux	Sysdig.Inc	[26, 85]	NA	NA
Auditd [71]	Linux	Linux Foundation	[33, 35, 37, 46, 51, 52, 59, 60, 89]	NA	NA
DTrace [18, 82]	Linux	Sun Microsystems	[37, 46, 51, 59, 60]	3.2	NA
Camflow [66]	Linux	University of Cambridge	[29]	9.7	NA
LTTng [23]	Linux	EfficiOS	NA	NA	NA
ETW [24]	Windows	Microsoft	[33, 46, 51, 52, 59, 60]	NA	NA
KennyLoggings [64]	Linux	UIUC	NA	4.6	NA
Hardlog [13]	Linux	Microsoft	NA	6.3	NA
Quicklog [34]	Linux	Florida State University	NA	5.3	NA
SystemTap [25, 41]	Linux	Linux Foundation	[56]	NA	NA
RAIN [42]	Linux	Georgia Institute of Technology	[42, 43]	NA	NA
Karma [19, 74]	Linux	Indiana University	[83]	NA	NA
PASS [62]	Linux	Harvard University	[83]	10.5	NA

Data Collecter Measurement Study

- Three most widely used industrial open-source collectors,
  - Sysdig, LTTng, and Auditd,
- Seven representative applications used in the surveyed papers
  - I/O-intensive applications :Nginx, Redis, Postmark, Django ,http
  - CPU-intensive applications : OpenSSL,7-ZIP.

Physical Machine	C1	C2	C3	C4
	1CPU + 2GB	4CPU + 8GB	16 CPU + 32GB	32 CPU + 64GB
Virtual Machine	C5	C6	C7	C8
	1CPU + 2GB	4CPU + 8GB	16 CPU + 32GB	32 CPU + 64GB

Client-Side Measurement Study

Application	Collector	C1	C2	C3	C4	C5	C6	C7	C8	Avg
Nginx	Auditd	597.30	101.30	34.60	34.80	821.10	186.30	23.70	10.90	226.25
	Sysdig	70.20	26.10	14.60	15.60	68.10	21.20	9.50	7.20	29.06
	LTTng	24.80	10.70	10.00	11.70	26.30	25.80	7.00	1.40	14.71
Redis	Auditd	457.00	58.10	41.70	50.20	512.00	53.20	46.00	43.20	157.67
	Sysdig	17.90	20.00	17.20	16.20	21.00	16.40	15.60	5.70	16.25
	LTTng	8.30	8.40	10.00	5.10	13.60	6.90	1.40	2.70	7.05
Postmark	Auditd	406.00	81.80	84.30	78.40	658.00	149.40	157.20	116.20	216.41
	Sysdig	88.80	19.20	18.00	22.00	98.80	23.20	16.50	7.50	36.75
	LTTng	10.30	9.40	12.30	18.10	12.90	10.30	10.90	11.60	11.98
Django (Python)	Auditd	2.50	0.70	2.10	2.30	1.20	0.50	1.50	2.10	1.62
	Sysdig	1.00	1.00	0.40	1.10	1.10	1.40	0.10	0.30	0.80
	LTTng	1.70	2.10	1.70	1.00	1.20	0.30	0.80	1.10	1.24
http (Golang)	Auditd	341.00	97.30	31.20	11.30	516.00	91.60	35.30	15.50	142.40
	Sysdig	60.70	13.90	10.60	2.80	76.70	11.90	4.10	2.20	22.86
	LTTng	13.80	6.50	4.20	4.10	13.40	6.20	5.80	4.20	7.28
OpenSSL	Auditd	2.90	1.80	1.20	1.00	6.90	0.10	1.70	0.20	1.98
	Sysdig	0.50	0.80	0.40	0.10	0.50	1.40	0.30	0.10	0.51
	LTTng	2.50	0.50	0.10	0.10	0.20	0.20	1.70	0.60	0.74
7-ZIP	Auditd	17.40	10.90	5.40	3.70	16.90	5.60	2.40	2.00	8.04
	Sysdig	1.50	1.30	1.10	1.10	1.20	1.00	0.80	0.70	1.08
	LTTng	2.40	1.80	0.90	0.80	4.70	2.30	0.10	0.10	1.64

Client-Side Measurement Study

- Memory consumption of provenance data collectors

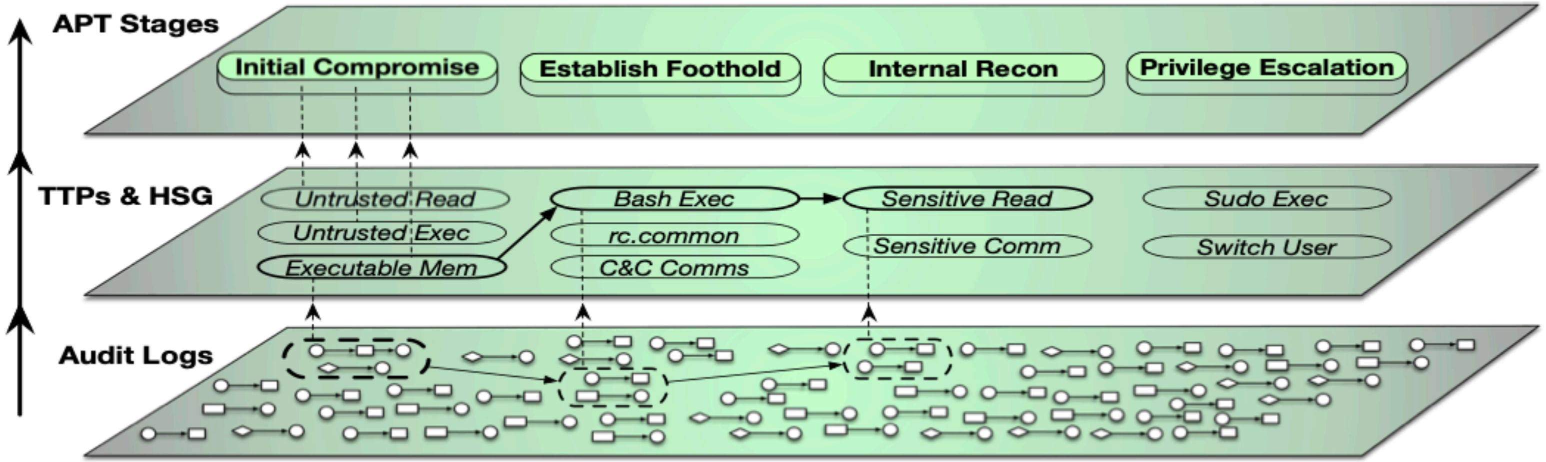
Agent	C1/C5	C2/C6	C3/C7	C4/C8
Auditd	65.9M	65.9M	65.9M	65.9M
Sysdig	38M	62M	158M	286M
LTTng	17.9M	23.9M	47.9M	79.9M

Server-Side Measurement Study

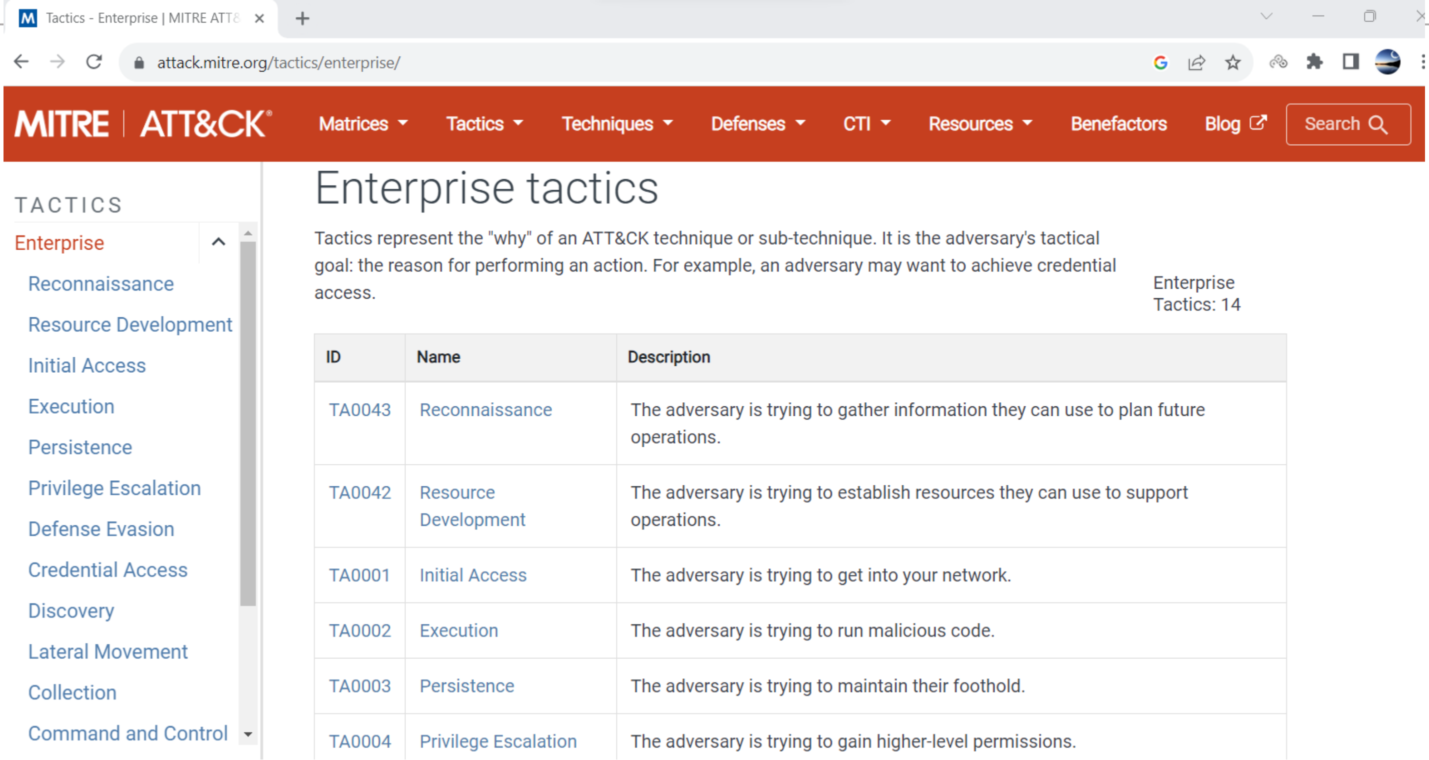
- Evaluated ProvDetector, UNICORN , and HOLMES use five datasets :
- DARPA-Cadets, DARPA-Theia, and DARPA-Trace
- Production dataset: real auditing data collected from a security company
- Simulation dataset: is an in-lab dataset created for attack simulation

Dataset	Host Num	Days	Data Size	Event Num	Event Rate	Event Size
DARPA-Cadets	1	11	14 GB	15 M	16.87 eps	1013 Byte
DARPA-Theia	1	11	7.5 GB	10 M	11.25 eps	810 Byte
DARPA-Trace	1	11	62 GB	72 M	75.76 eps	925 Byte
Simulation	5	12	23 GB	50 M	48.23 eps	483 Byte
Production	300+	5	16.85 GB	17 M	39.35 eps	1064 Byte

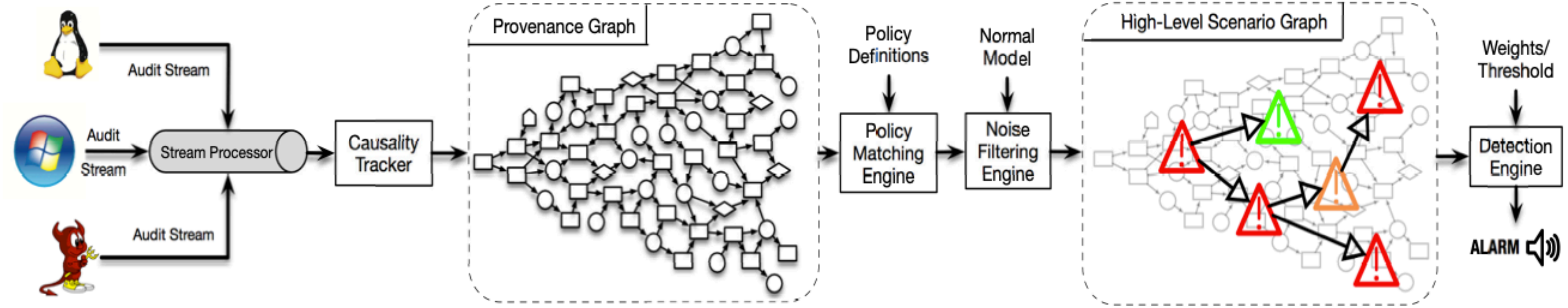
Holmes


APT Stages
Initial Compromise
Establish Foothold
Internal Recon
Privilege Escalation
TTPs & HSG
Untrusted Read
Untrusted Exec
Executable Mem
Bash Exec
rc.common
C&C Comms
Sensitive Read
Sensitive Comm
Sudo Exec
Switch User
Audit Logs

Introduction to MITRE



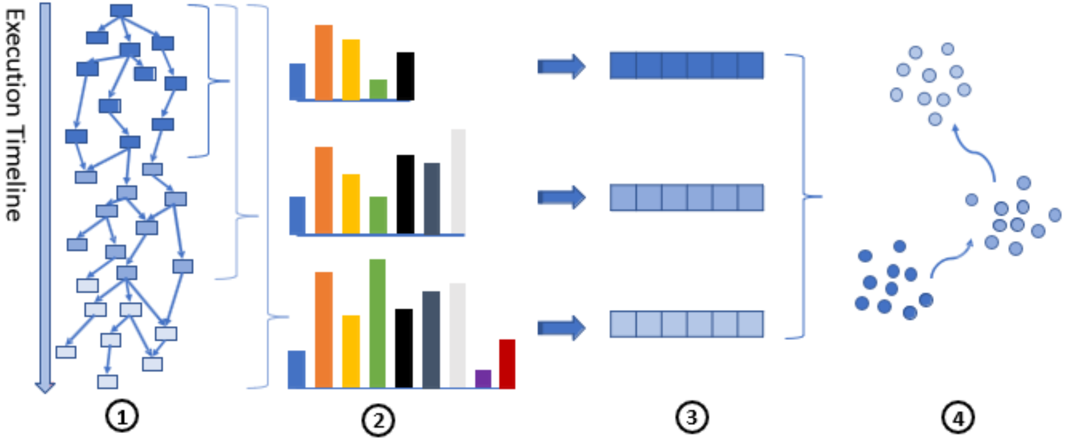
Holmes



Audit Stream
Audit
Stream
Audit Stream
Stream Processor
Causality Tracker
Provenance Graph
Policy Definitions
Policy Matching Engine
Normal Model
Noise Filtering Engine
High-Level Scenario Graph
Weights/ Threshold
Detection Engine
ALARM

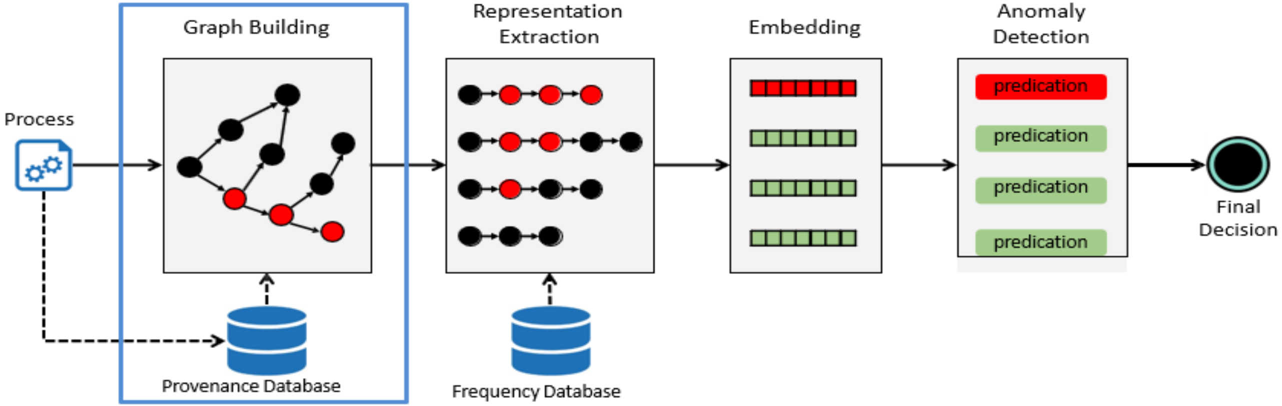
Unicorn

1. Takes as input a labeled, streaming provenance graph
2. Builds at runtime an in-memory graph histogram
3. Computes a fixed-size graph sketch periodically
4. Clusters sketches into a system model



ProveDetector

- Uses path instead of node to find anomaly



Server-Side Measurement Study

- Memory consumption results
  - HOLMES and ProvDetector was positively correlated with the data volume both exceeded the reference value (<20MB/host)
  - For UNICORN, stable memory consumption -> Parallel Sliding Windows it exceeded the reference value by 11.9 times.
- Therefore, none of these systems meet the requirement for the Memory

Dataset	# of Graph Nodes	Memory (MB/host)		
		HOLMES	ProvDetector	UNICORN
DARPA-Cadets	280W+	5683	10240	274
DARPA-Theia	125W+	3870	6574	242
DARPA-Trace	325W+	9605	-	242
Simulation	3W+	73	195	213
Production	5W+	84	240	219

Server-Side Measurement Study

- Interpretation:
- ProvDetecor satisfy the reference value (< 50 nodes)
- HOLMES generates alarms within ten times larger than the reference value.
- UNICORN generates too coarse-grained provenance graphs -> is not practical in industry.

Dataset	HOLMES	ProvDetector	UNICORN
DARPA-Cadets	173	15	154730
DARPA-Theia	73	8	522735
DARPA-Trace	450	-	1454033
Simulation	566	7	11587
Production	81	5	17853

Server-Side Measurement Study

- Alarm Triage
  - UNICORN can roughly satisfy the reference value (<0.1 alarms/host/day).
  - HOLMES and ProvDetector will need to improve their precision significantly.

Dataset	HOLMES	ProvDetector	UNICORN
DARPA-Cadets	21	90	0.3
DARPA-Theia	36.7	90	0.1
DARPA-Trace	13.9	-	0.45
Simulation	2.3	23	0.09
Production	12.1	56.3	0.13

FINDINGS OF STUDY

- RQ1: How does the industry compare the effectiveness of P-EDR and conventional EDR?

- The industry acknowledges that P-EDR systems are superior to conventional EDR systems due to better interpretability. Experienced security analysts can easily understand basic provenance graphs that consist of low-level system audit events, and companies have designed training sessions in provenance analysis for training novice analysts.

FINDINGS OF STUDY

- RQ2: What are the bottlenecks for the industry to adopt EDR Systems?
- The operating cost, which consists of the four-must factors: Memory, Client-Side Overhead, Interpretation, and Alarm Triage, is the primary bottleneck for the industry to adopt an EDR/P-EDR system.

FINDINGS OF STUDY

- RQ3: How well can existing P-EDR systems proposed in academia meet the expectations of the industry?

- There exist three important gaps (overlooking client-side over-head, the imbalance between alarm triage cost and interpretation cost, and excessive server-side memory consumption) between the academic research and the industry expectations.

Acknowledgments

- [Dong] Are we there yet? An Industrial Viewpoint on Provenance-based Endpoint Detection and Response Tools, F. Dong, S. Li, P. Jiang, D. Li, H. Wang, L. Huang, X. Xiao, J. Chen, X. Luo, Y. Guo, CCS 2023.
- [Wang] You Are What You Do: Hunting Stealthy Malware via Data Provenance Analysis, Q. Wang, W.U. Hassan, D. Li, K. Jee, X. Yu, K. Zou, J. Rhee, Z. Chen, W. Cheng, C.A. Gunter, H. Chen, NDSS 2020.
- [Unicorn] UNICORN: Runtime Provenance-Based Detector for Advanced Persistent Threats, X. Han, T. Pasquier, A. Bates, J. Mickens, and M. Seltzer, NDSS 2020.
- [Holmes] HOLMES: Real-Time APT Detection through Correlation of Suspicious Information Flows, S. Momeni Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, V. N. Venkatakrishnan, IEEE Symposium on Security and Privacy 2019.