KCon

名侦探的下午茶
Hunting with Provenance

张润滋

+1 进阶，
护航未来

KCon

+1

CONTENTS

目录

KCon

+1

威胁狩猎与溯源数据

Hunting with Provenance

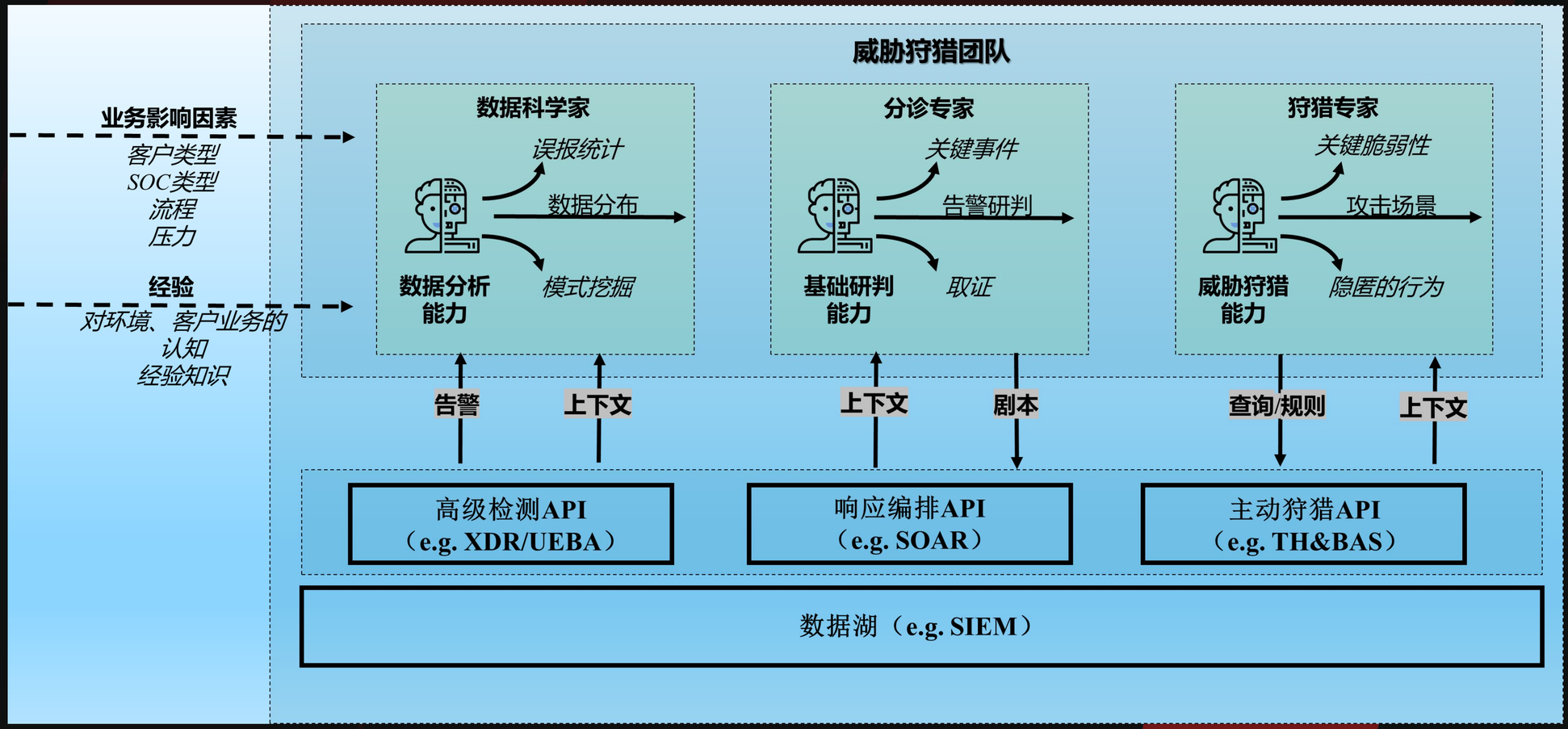
KCon Threat Hunting 威胁狩猎

"The process of *proactively* and *iteratively* searching through networks to detect and isolate *advanced threats that evade existing security solutions*. " – Wikipedia

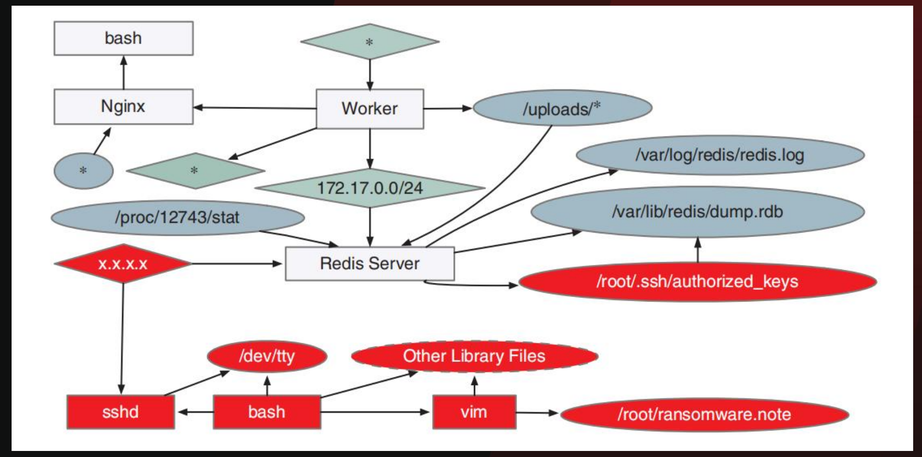
"You should be able to leverage your tools or prod ucts to *detect high fidelity alerts which you identifi ed previously*. " – Carbon Black Researcher

经验与数据驱动的狩猎

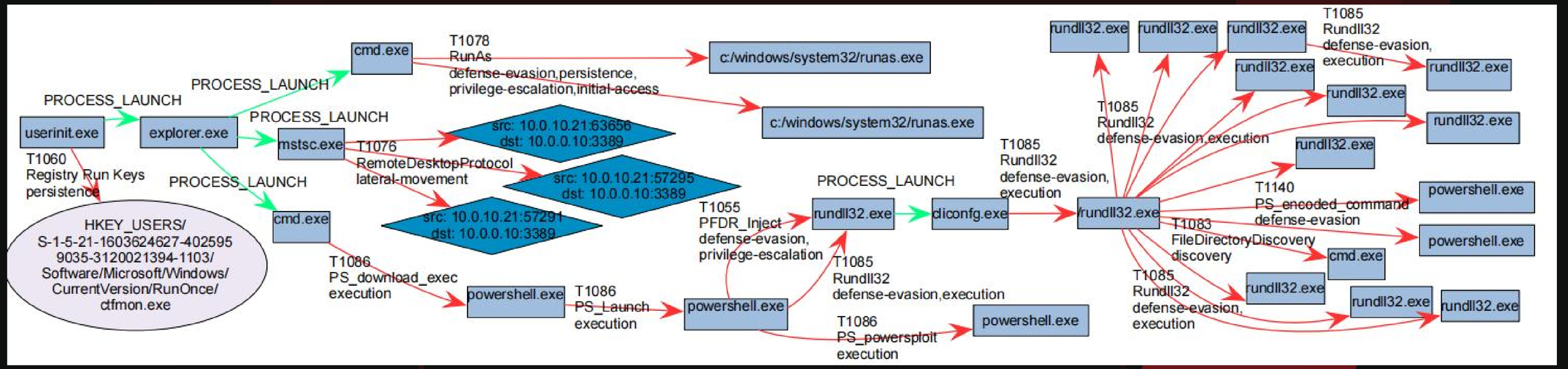
- 数据平台已成为威胁狩猎的基础设施
- 经验已不限于传统攻防方面，数据科学成为狩猎团队不可获取的关键能力组成



Provenance（溯源数据）



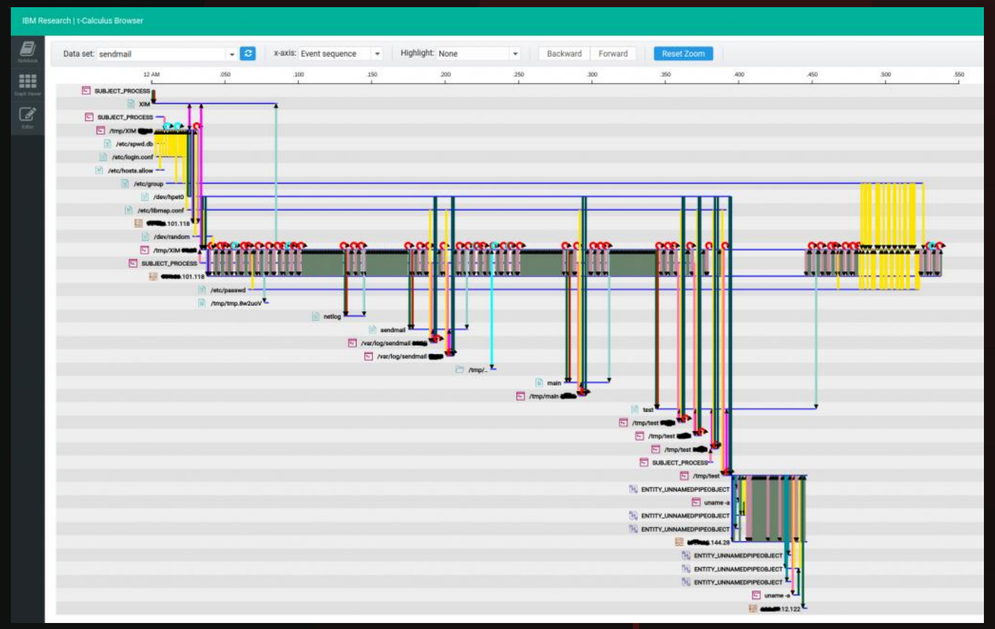
- 溯源数据是终端侧可观测性的基础
- 溯源数据的粒度是技术级的
- 溯源数据自然构成具有因果关联的图结构（Provenance Graph）

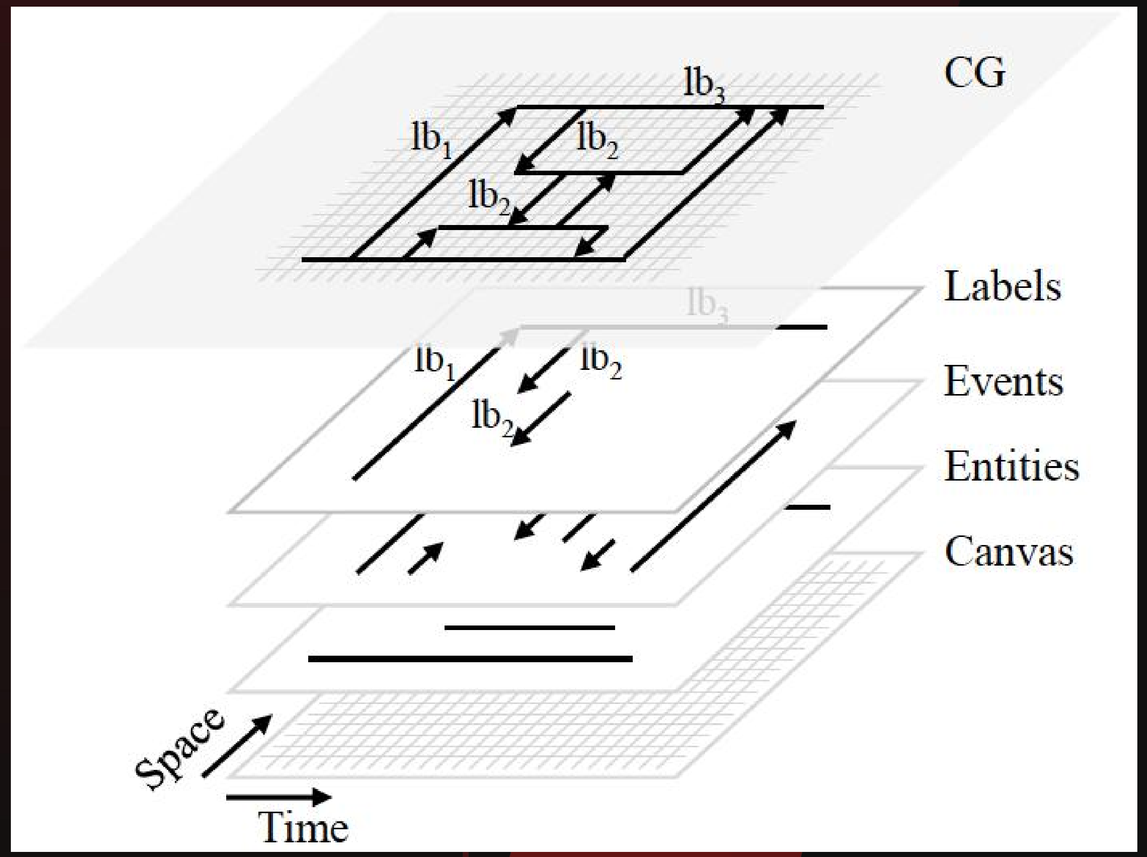


KCon Hunting with Provenance Graph

+1

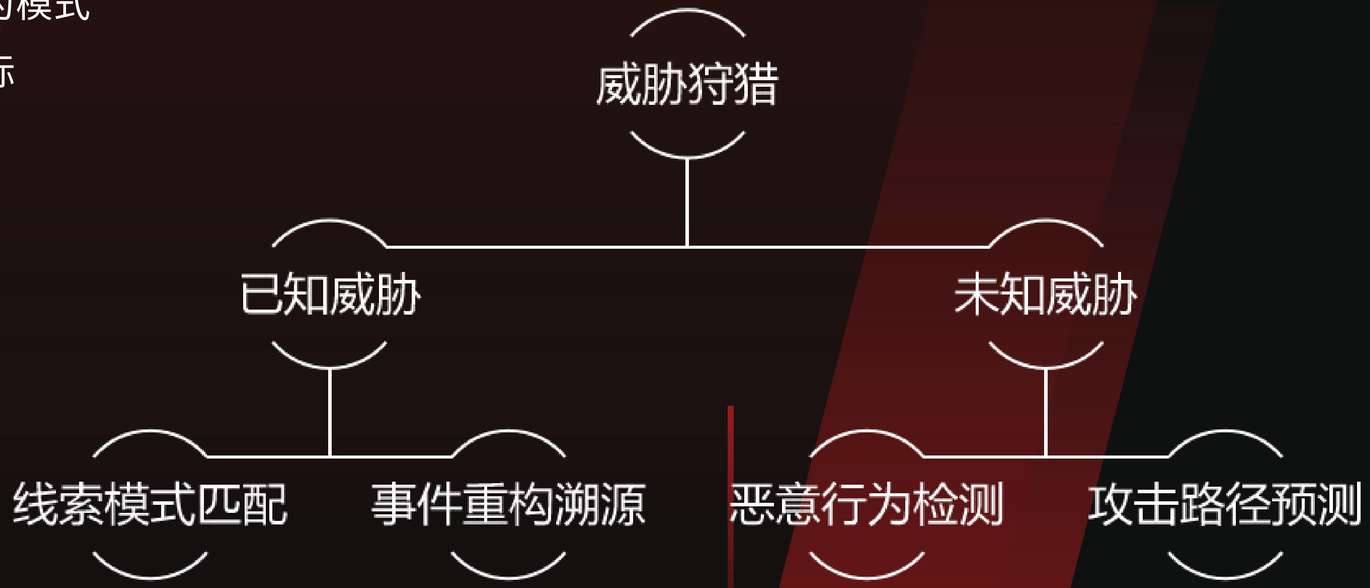
- 溯源数据是上下文构建的基本数据
- 溯源数据是战术意图抽取的推理基础
- 图分析技术自然的可应用与溯源数据分析





Hunting with Provenance Graph (Cont.)

- **针对已知威胁，**
  - 如情报等，在溯源图中匹配查询
  - 针对已确认攻击事件，进行攻击路径重构溯源
- **针对未知威胁**
  - 通过异常检测、语义分析，挖掘潜在恶意行为模式
  - 串联攻击步骤，分析攻击意图，预测攻击目标

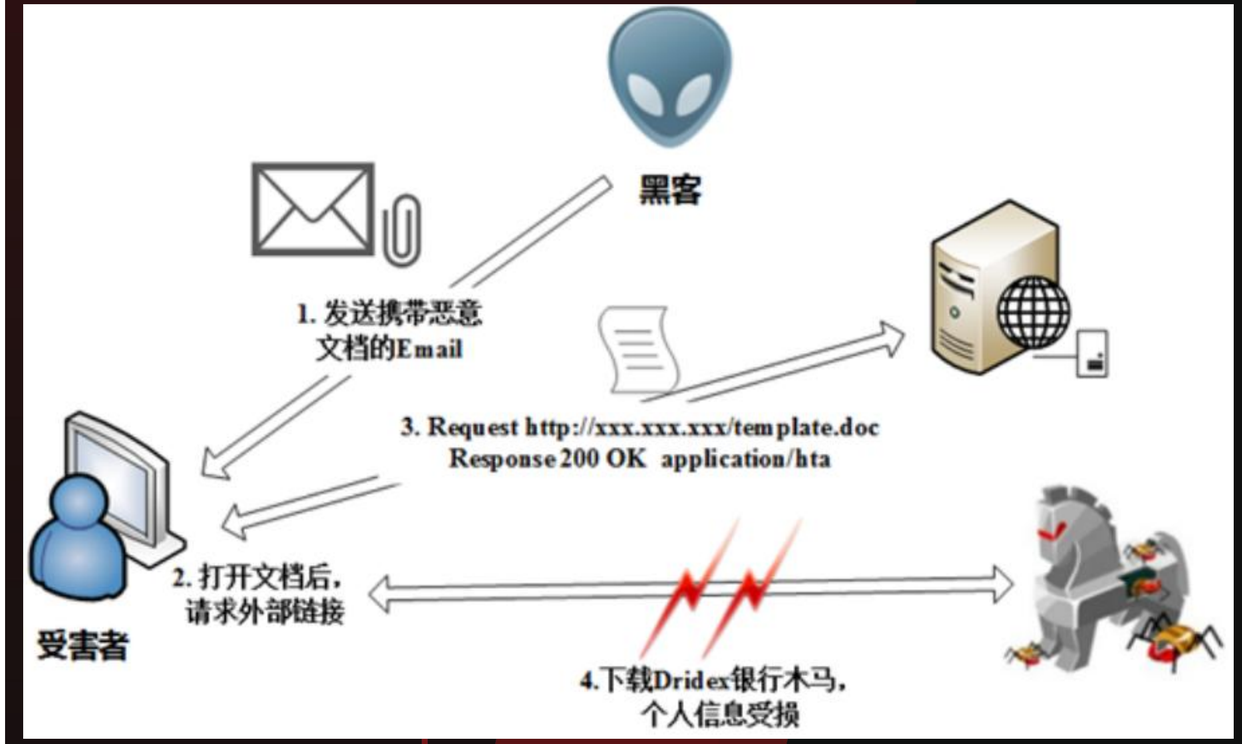


KCon 一个实例：CVE-2017-0199

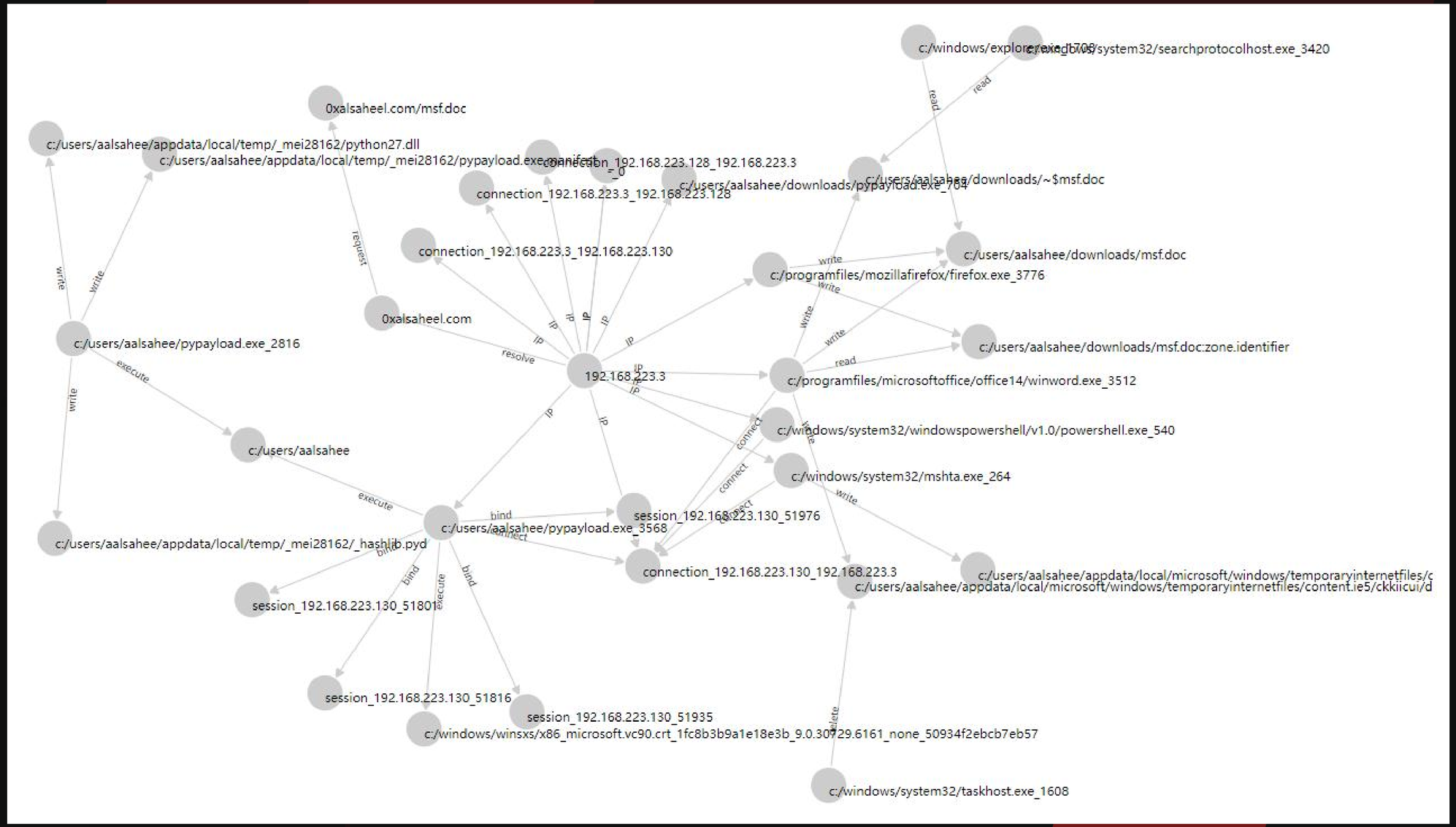
+1

Table 2: Additional Routinely Exploited Vulnerabilities in 2021

CVE	Vendor and Product	Type
CVE-2021-42237	Sitecore XP	RCE
CVE-2021-35464	ForgeRock OpenAM server	RCE
CVE-2021-27104	Accellion FTA	OS command execution
CVE-2021-27103	Accellion FTA	Server-side request forgery
CVE-2021-27102	Accellion FTA	OS command execution
CVE-2021-27101	Accellion FTA	SQL injection
CVE-2021-21985	VMware vCenter Server	RCE
CVE-2021-20038	SonicWall Secure Mobile Access (SMA)	RCE
CVE-2021-40444	Microsoft MSHTML	RCE
CVE-2021-34527	Microsoft Windows Print Spooler	RCE
CVE-2021-3156	Sudo	Privilege escalation
CVE-2021-27852	Checkbox Survey	Remote arbitrary code execution
CVE-2021-22893	Pulse Secure Pulse Connect Secure	Remote arbitrary code execution
CVE-2021-20016	SonicWall SSLVPN SMA100	Improper SQL command neutralization, allowing for credential access
CVE-2021-1675	Windows Print Spooler	RCE
CVE-2020-2509	QNAP QTS and QuTS hero	Remote arbitrary code execution
CVE-2019-19781	Citrix Application Delivery Controller (ADC) and Gateway	Arbitrary code execution
CVE-2019-18935	Progress Telerik UI for ASP.NET AJAX	Code execution
CVE-2018-0171	Cisco IOS Software and IOS XE Software	Remote arbitrary code execution
CVE-2017-11882	Microsoft Office	RCE
CVE-2017-0199	Microsoft Office	RCE



一个示例：CVE-2017-0199（Cont.）



KCon

+1

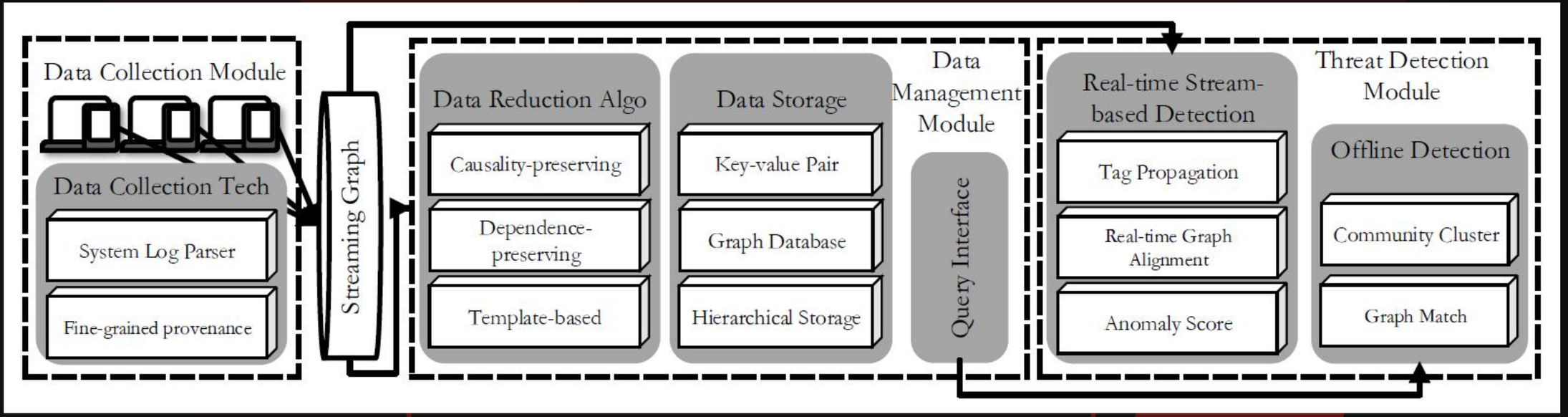
02

溯源数据挖掘的挑战

Open Challenges

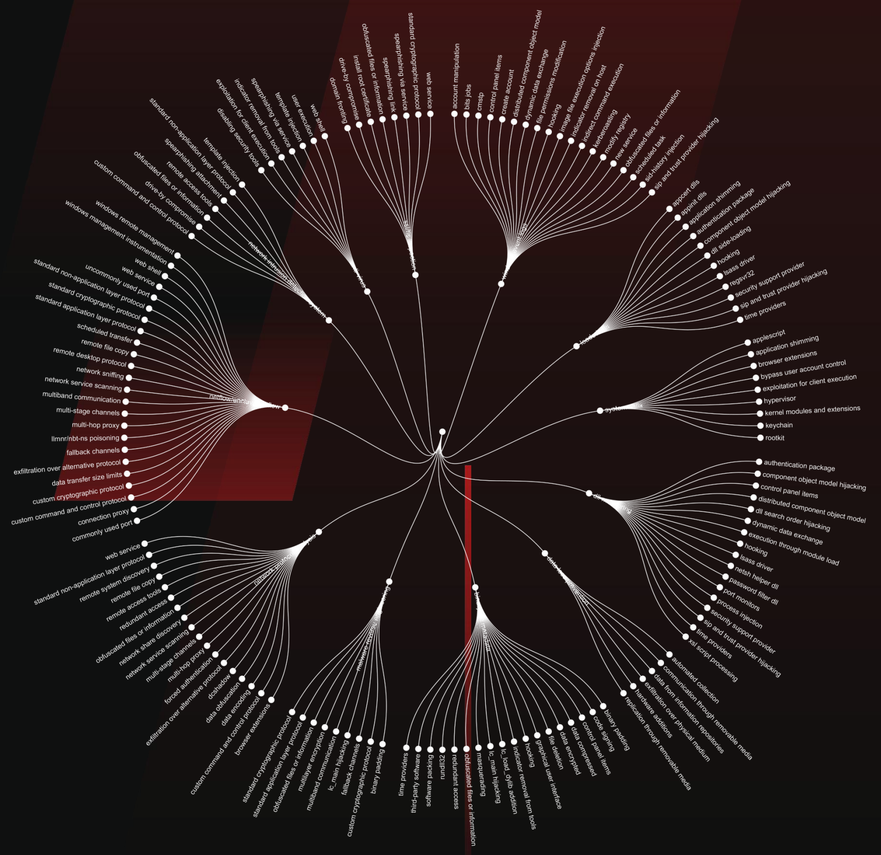
挑战1：采集与分析系统瓶颈（观测性挑战）

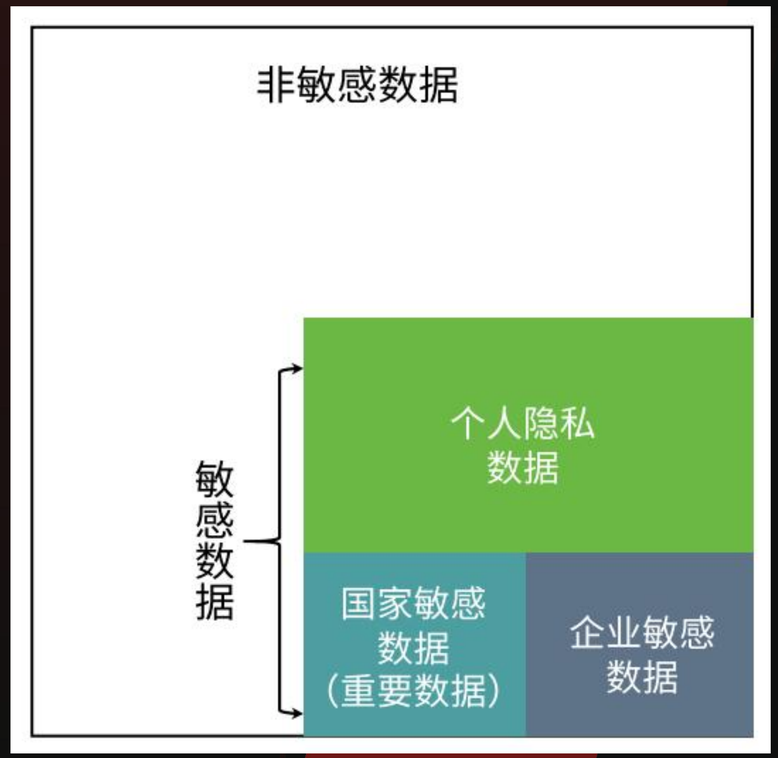
- **理想：**Provenance全面收集系统实体的可观测数据，支撑威胁研判与溯源
- **现实：**高级威胁低频且具有隐匿性与企业和组织需要持续进行风险管控的矛盾，导致大规模数据采集开销、传输开销、存储开销、分析开销大幅提升



挑战2：数据隐私风险（观测性挑战）

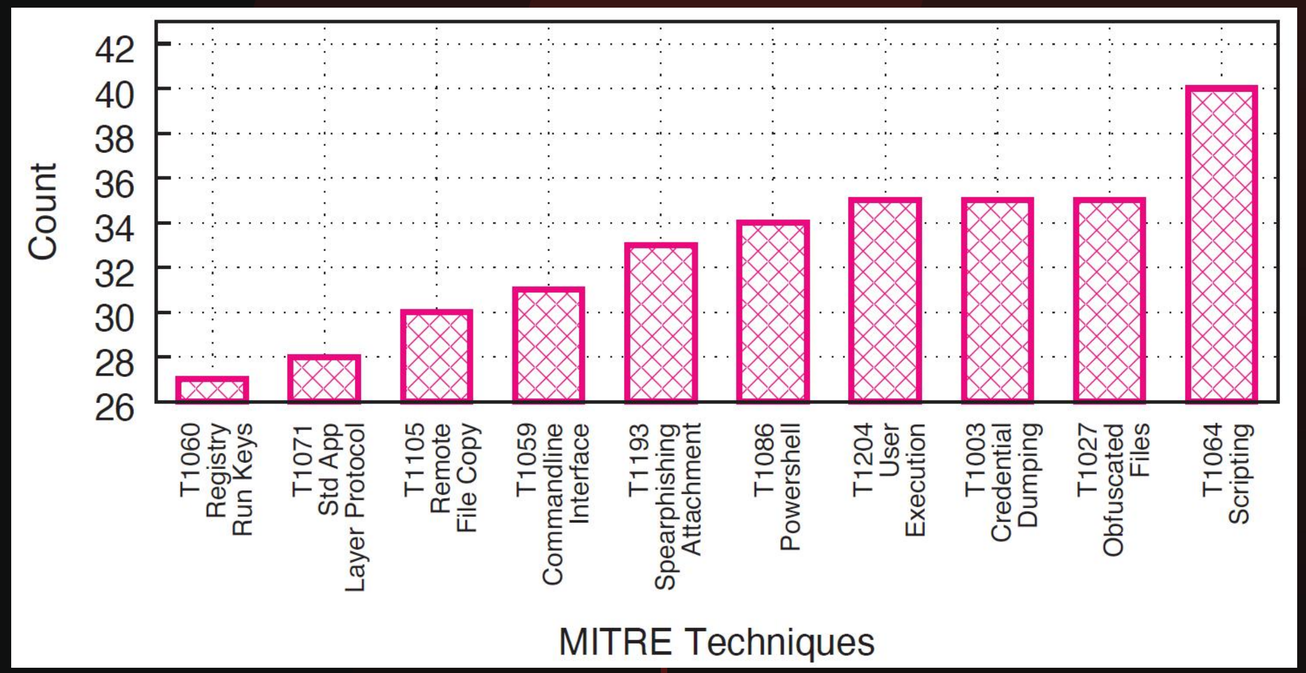
- **理想**：主机全覆盖，用户全覆盖，行为全覆盖
- **现实**：关键数据资产与个人隐私保护需求上升与数据分析可用性之间的矛盾





挑战3：大海捞针与高误报率（表象挑战）

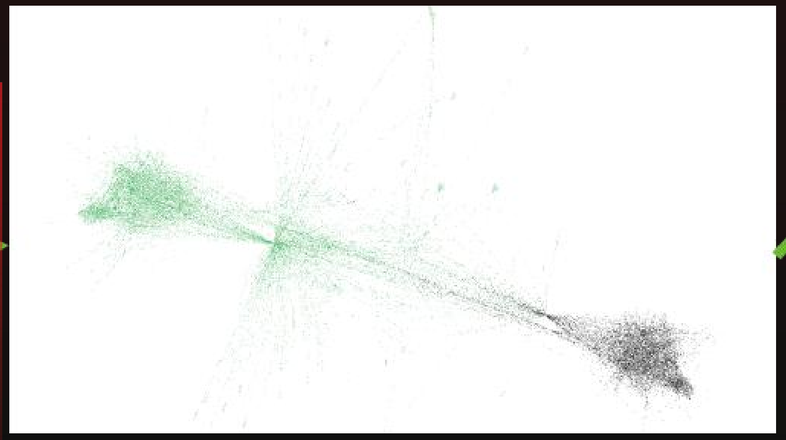
- **理想：**基于溯源图精准定位威胁、还原攻击路径
- **现实：**溯源图分析是典型的复杂图、小样本、低频分析问题，规则驱动的方式耗时耗力、覆盖率低，基本的统计分析方法误报率高

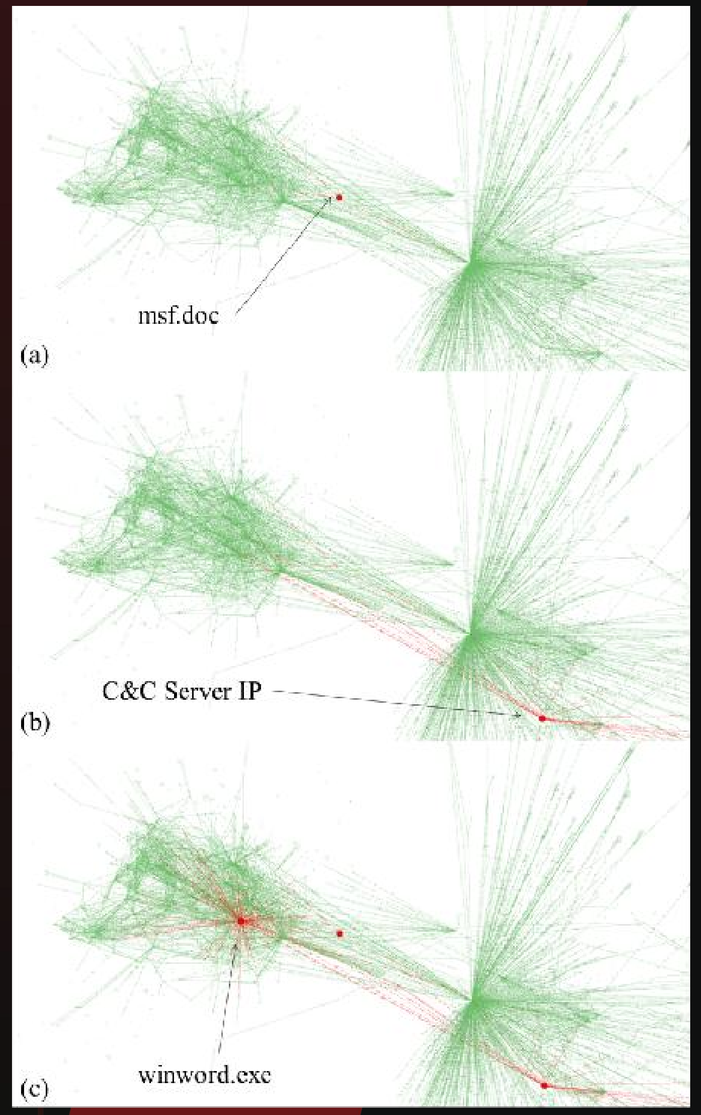


一项针对某著名国外安全企业终端告警的分析表明，由34台机器触发的58096条告警中，与检测目标APT29行为相关真实告警只有1104条，告警的精度只有1.9%。

挑战4：信息流依赖爆炸（本质挑战）

- **理想：**根据溯源图路径，快速还原攻击路径，进行检测、溯源、取证
- **现实：**现阶段的溯源数据采集在一定的资源限制下，难以精细刻画信息传递流，特别是在跨网络、终端数据场景下
  - 在没有先验知识或基线数据的情况下，任何恶意信息流经过中间实体的中转，将以一定的概率向上下游实体扩散。

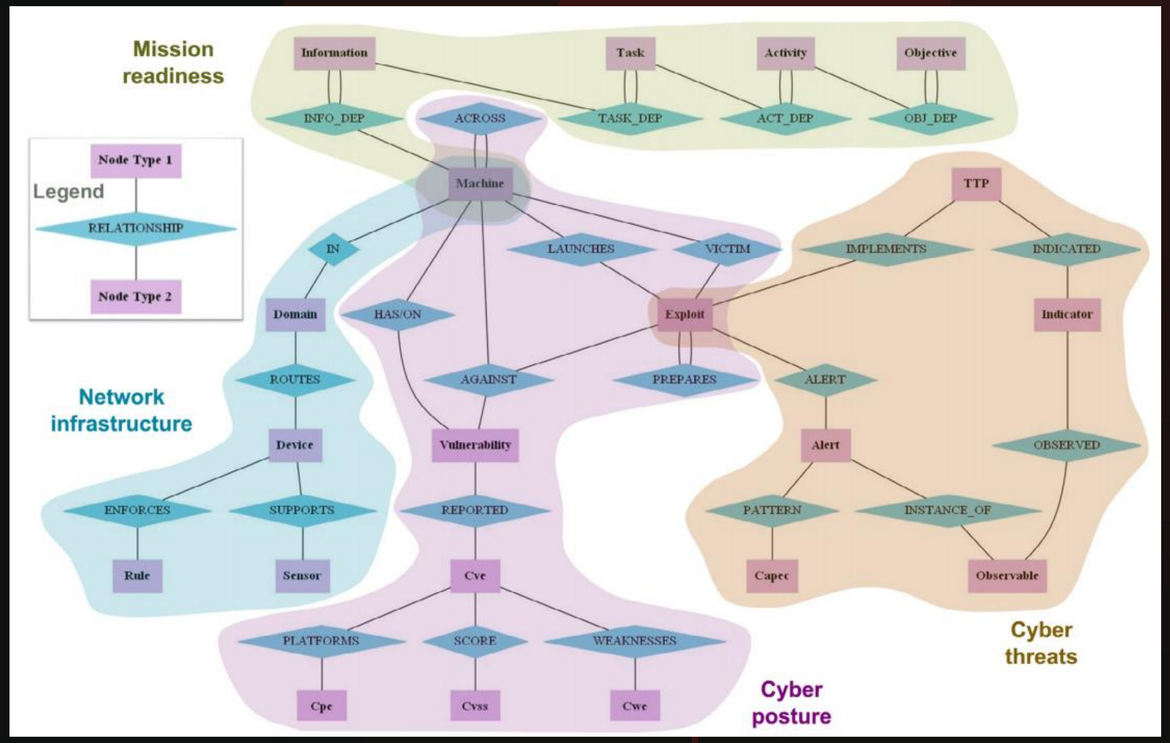


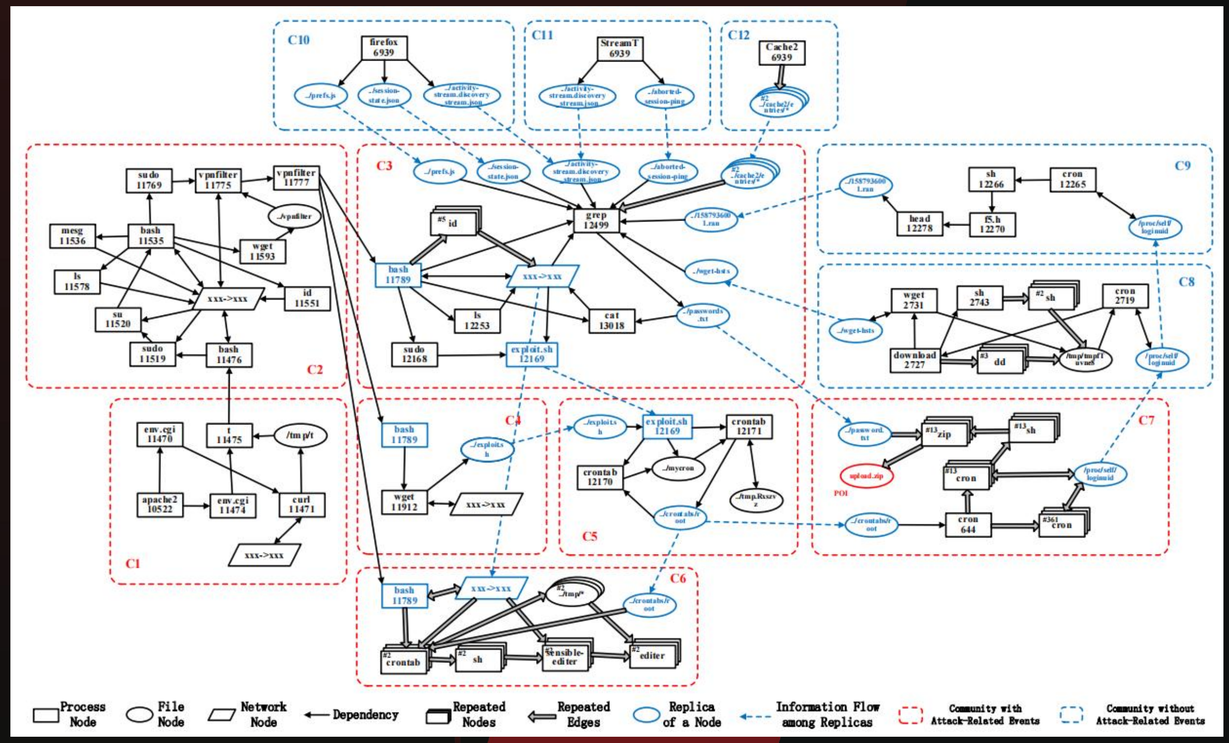


KCon 挑战5：技战术、环境动态博弈（本质挑战）

+1

- **理想：**根据ATT&CK矩阵，逐一落地技战术识别方法
- **现实：**威胁狩猎是环境分析、任务推测、脆弱性暴露及威胁分析多维度的复杂任务，相同技战术在不同的主机环境下、采集机制下、攻击者实现模式下，有着不同的数据表现方式，导致经验的持续失效，给模式挖掘带来挑战





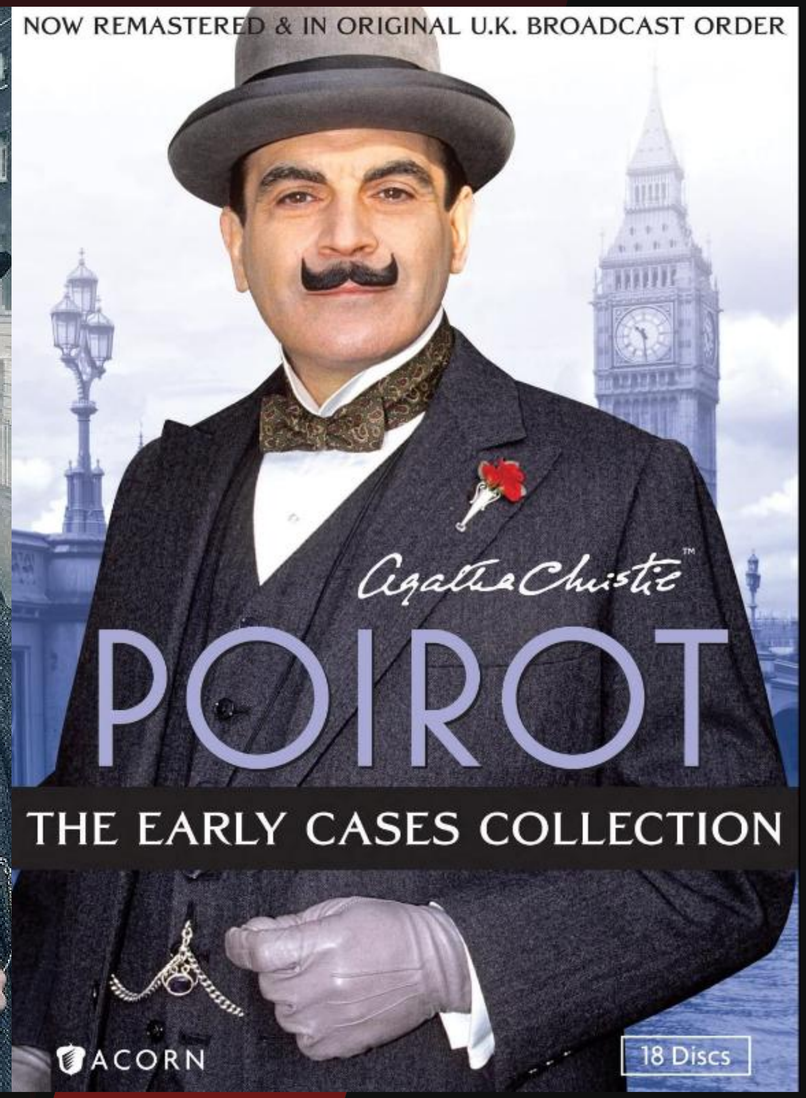
KCon

+1

溯源数据狩猎方法总结

State of the Art

KCon 名侦探的下午茶 +1

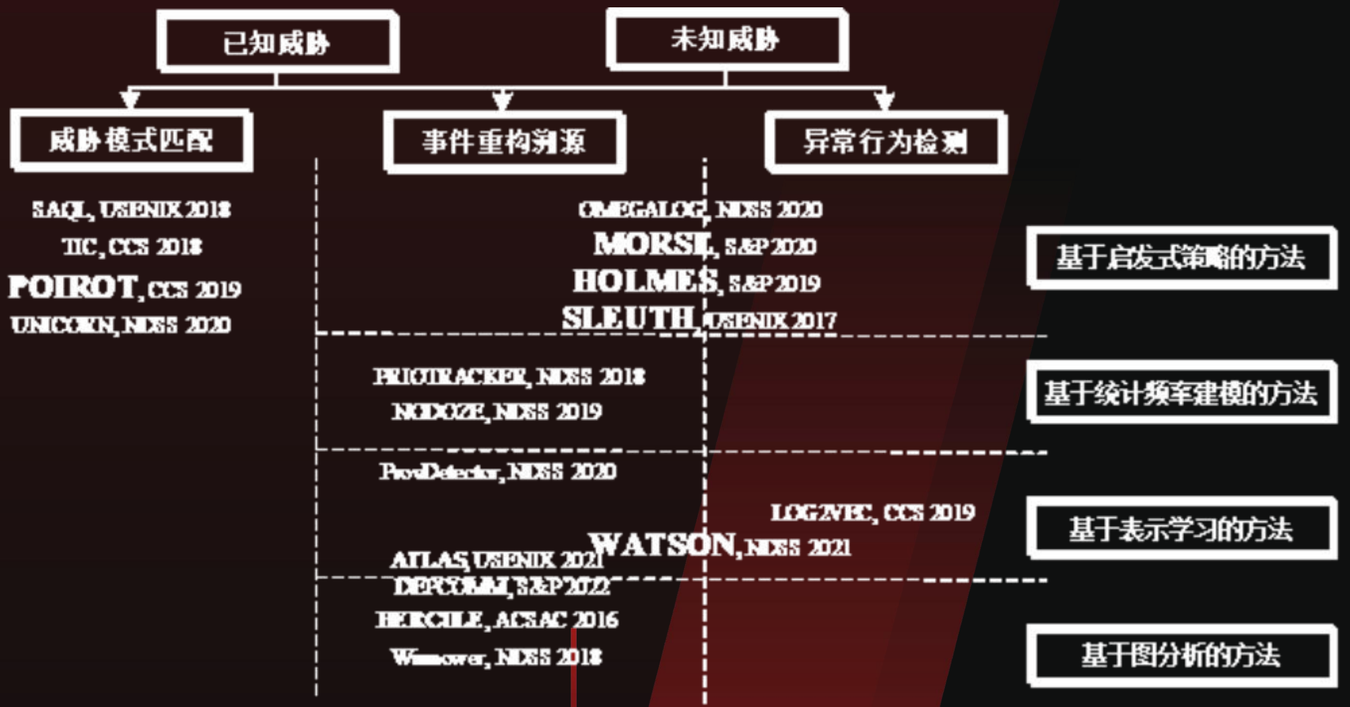


KCon

名侦探的下午茶

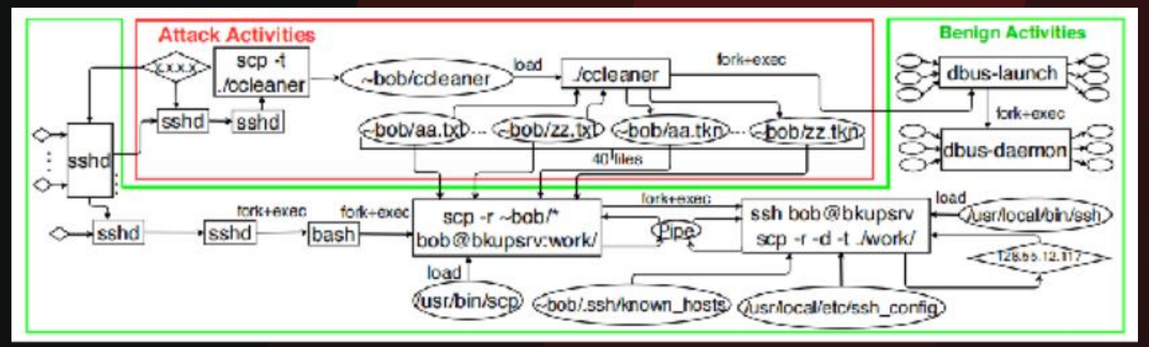
+1

- 通过溯源数据实现APT等高级威胁分析，已成为TOP安全研究机构的军备竞赛
- 基于大规模溯源线索进行威胁狩猎，堪比侦探与犯罪对抗的“猫鼠游戏”



KCon 语义分析学派

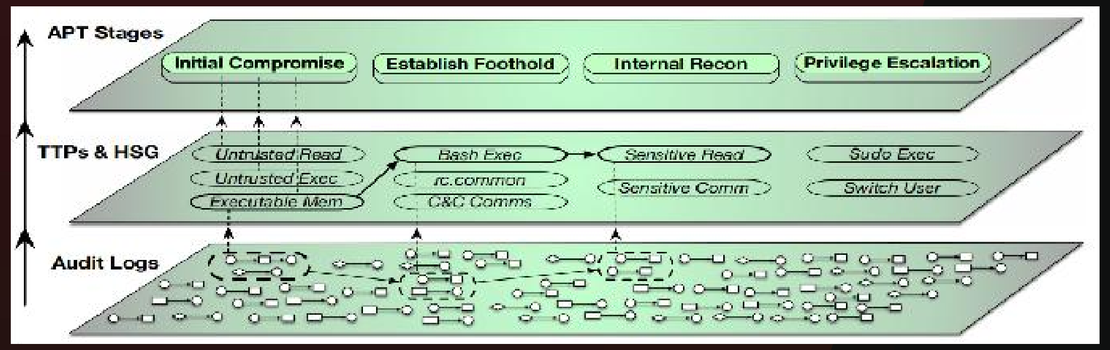
+1



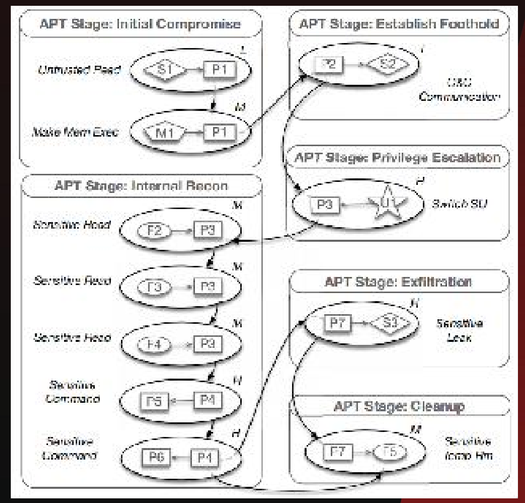
Event	Tag to update	New tag value for different subject types		
		benign	suspect	suspect environment
create(s,x)	$x.itag$	$s.itag$		
read(s,x)	$s.dtag$	$min(s.dtag, x.dtag)$		
write(s,x)	$x.itag$	$min(s.dtag + u_b, x.itag)$	$min(s.dtag, x.itag)$	$min(s.itag + a_c, x.itag)$
periodically:	$s.dtag$	$max(s.dtag,\ d_b * s.dtag \mid (1 \quad d_b) \in T_{pb})$	no change	$max(s.dtag,\ d_c * s.dtag \mid (1 \quad d_c) \in T_{pc})$

Event	Tag to update	New tag value for different subject types		
		benign	suspect	suspect environment
load(s,x)	$s.stag$	$min(s.stag, x.itag)$		
	$s.dtag$	$min(s.dtag, x.dtag)$		
exec(s,x)	$s.stag$	$x.itag$	$min(x.itag, susp\_env)$	$x.itag$
	$s.dtag$	$\langle 1.0, 1.0 \rangle$	$min(s.dtag, x.dtag)$	$min(s.dtag, x.dtag)$
inject(s,s')	$s'.stag$	$min(s'.stag, s.itag)$		
	$s'.dtag$	$min(s.dtag, s'.dtag)$		

语义标签定义与传播



APT Stage	TTP	Event Family	Events	Severity	Prerequisites
Initial_Compromise(P)	*Untrusted_Read*(S,P)	READ	FileIoRead (Windows), read/pread/readv/preadv (Linux,BSD)	L	$S.ip \notin$ {Trusted_IP_Addresses}
	Make_Mem_Exec(P,M)	MPROTECT	VirtualAlloc (Windows), mprotect (Linux,BSD)	M	$\text{SPRIT\_EXECS} \in M.flags \wedge \exists\ Untrusted\_Read(T,P') : path\_factor(P',P) <= path\_thres$
Establish Foothold(P)	*Shell Exec*(F,P)	EXEC	ProcessStart (Windows), execve/fexecve (Linux,BSD)	M	$F.path \in$ {Command_Line_Utilities} $\wedge \exists\ Initial\_Compromise(P') : path\_factor(P',P) <= path\_thres$

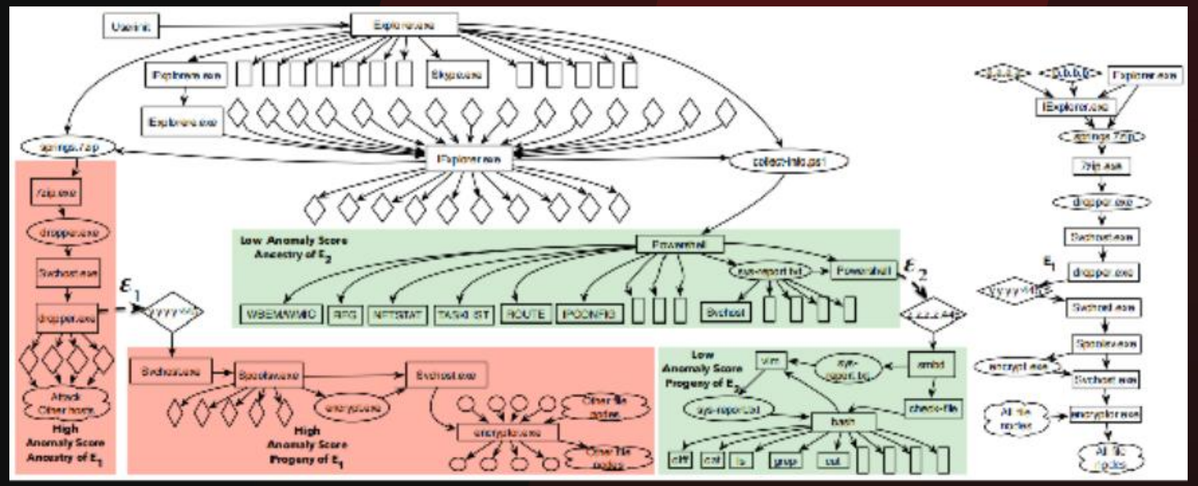


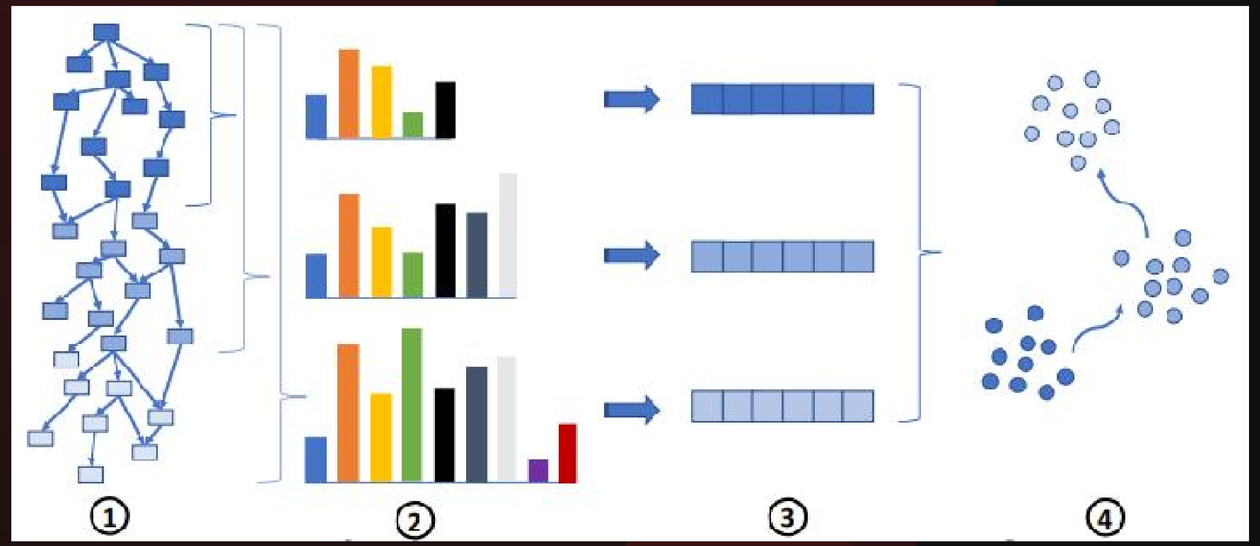
技战术模板定义与匹配

KCon

统计分析学派

+1



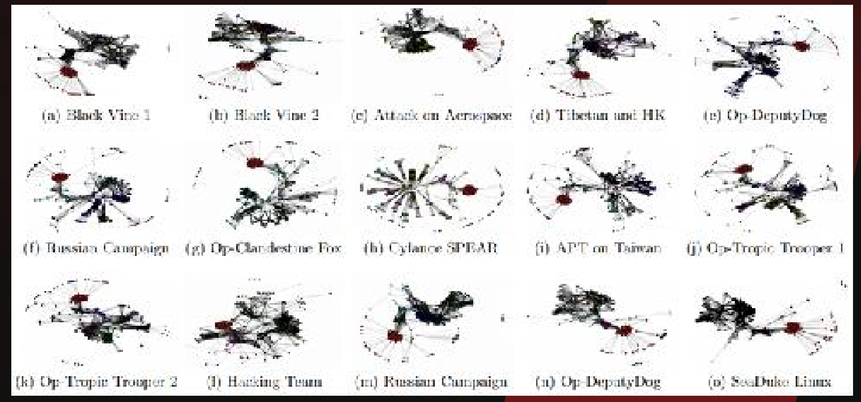


Algorithm 1: GETPATHANOMALYSCORE

Inputs : Alert Event $\mathcal{E}_\alpha$:
Max Path Length Threshold τ_l
Output: List $L_{<P,AS>}$ of dependency path and score pairs.

G_α = GETDEPENDENCYGRAPH($\mathcal{E}_\alpha$)
$V_{src} \leftarrow$ GETSRCVERTEX($\mathcal{E}_\alpha$)
$V_{dst} \leftarrow$ GETDSTVERTEX($\mathcal{E}_\alpha$)
$L_b \leftarrow$ DFSTRAVERSALBACKWARD($G_\alpha, V_{src}, \tau_l$)
$L_f \leftarrow$ DFSTRAVERSALFORWARD($G_\alpha, V_{dst}, \tau_l$)
/* *Combine Backward and Forward Dependency Paths* */
$L_p \leftarrow$ COMBINEPATHS(L_b, L_f)
/* *Generate a transition matrix of an input graph using Eq. 1* */
M = GETTRANSITIONMATRIX(G)
foreach $P \in L_p$ **do**
  /* *Calculate Path anomaly score using Eq. 2 and Eq. 3* */
  $AS \leftarrow$ CALCULATESCORE(P, M)
  /* *Append path and its anomaly score to a list* */
  $L_{<P,AS>} \leftarrow L_{<P,AS>} \cup < P, AS >$
end
return $L_{<P,AS>}$

统计低频与降噪剪枝



图结构信息统计与社区挖掘

KCon 表示学习学派 +1

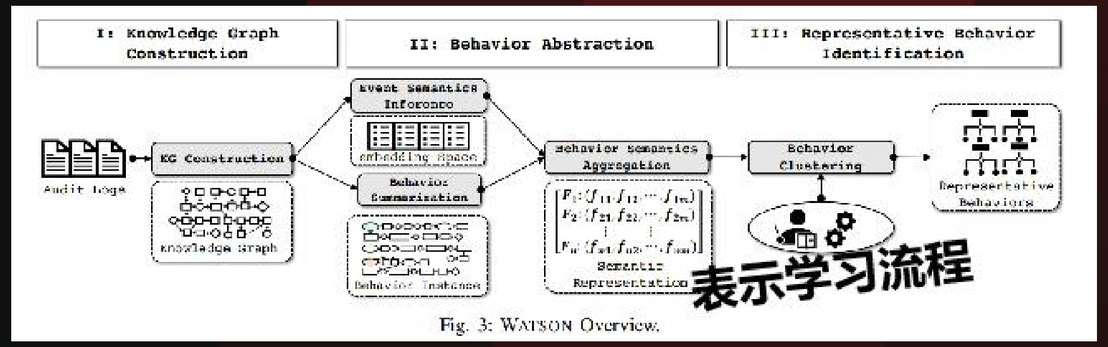
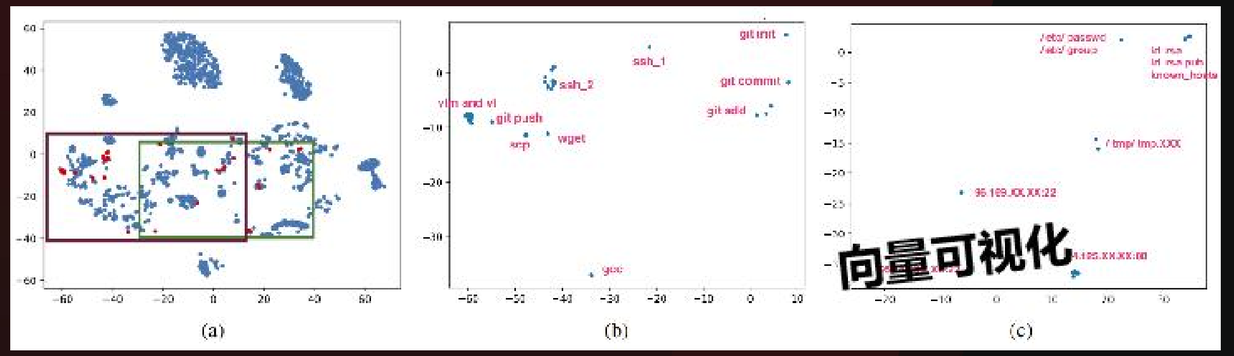
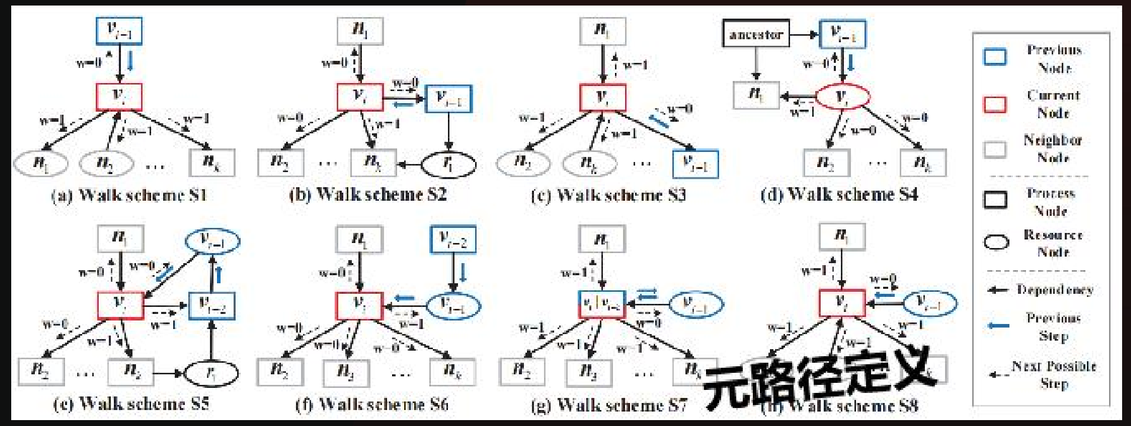
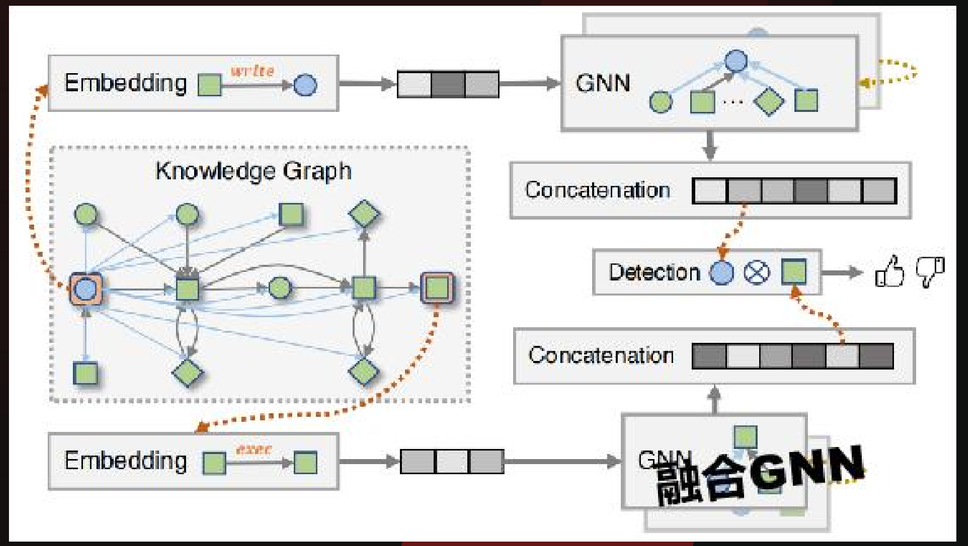


Fig. 3: WATSON Overview.

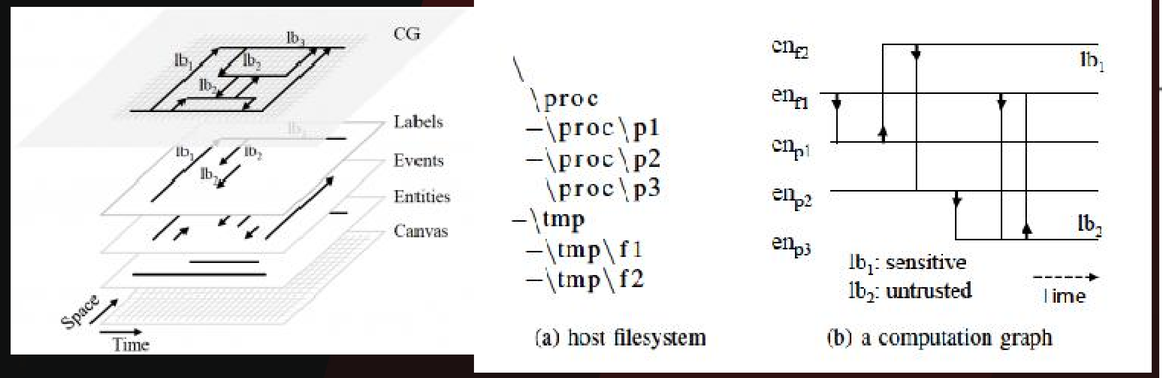




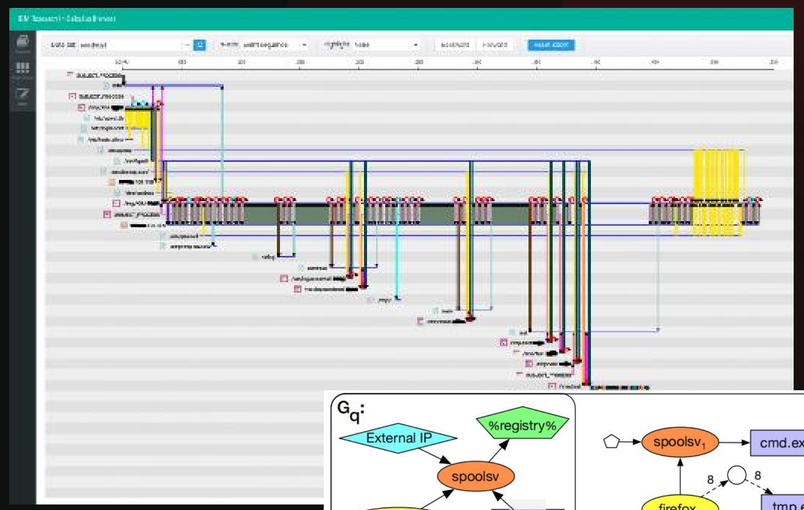


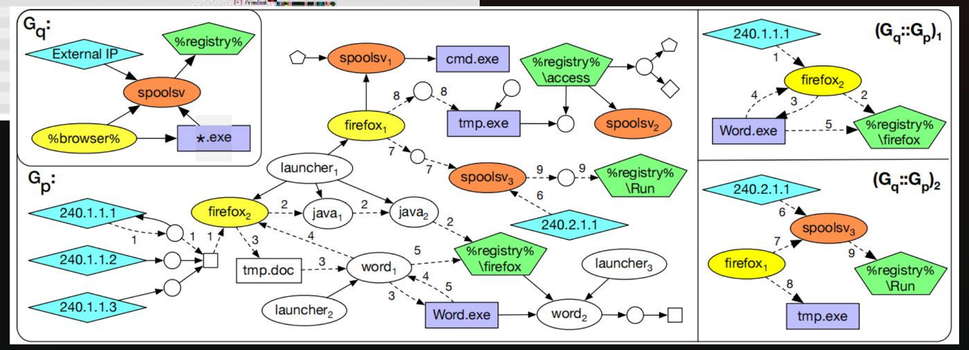
KCon 语言模型学派

+1



名称	数据类型				支持的模式类型						实时性		特性
	行为	环境	情报	知识	单点	集合	序列	静态图	时序图	模型	流式	批式	
Kestrel [8]	√	√	√		√	√		√		√		√	IBM Threat Hunting Language开源项目 基于Jupyter notebook交互式实现
EQL [9]	√				√	√	√				√		Event Query Language 无结构约束（Schema-independent）
CyQL [10]	√	√	√	√				√				√	Elastic/Endgame终端分析开源实现 CyGraph Query Language 基于MITRE CyGraph多源异构图架构
π-calculus [11]	√	√	√		√	√	√	√	√			√	基于IBM Threat Intelligence Computing 时序图分析引擎
Sigma [12]	√				√	√						√	Generic Signature Formats for SIEM Systems 应用层指纹识别格式语言
Flink CEP [13]	√	√	√		√	√	√				√	√	通用流式分析引擎 灵活的序列模式支持
TBQL [14]	√		√		√			√				√	Threat Behavior Query Language 面向情报要素提取与匹配
SAQL [15]	√				√	√				√	√		Stream based Anomaly Query Language 异常检测专用语言，支持多种类型异常分析
AIQL [16]	√				√	√				√		√	Attack Investigation Query Language 与SAQL同体系架构



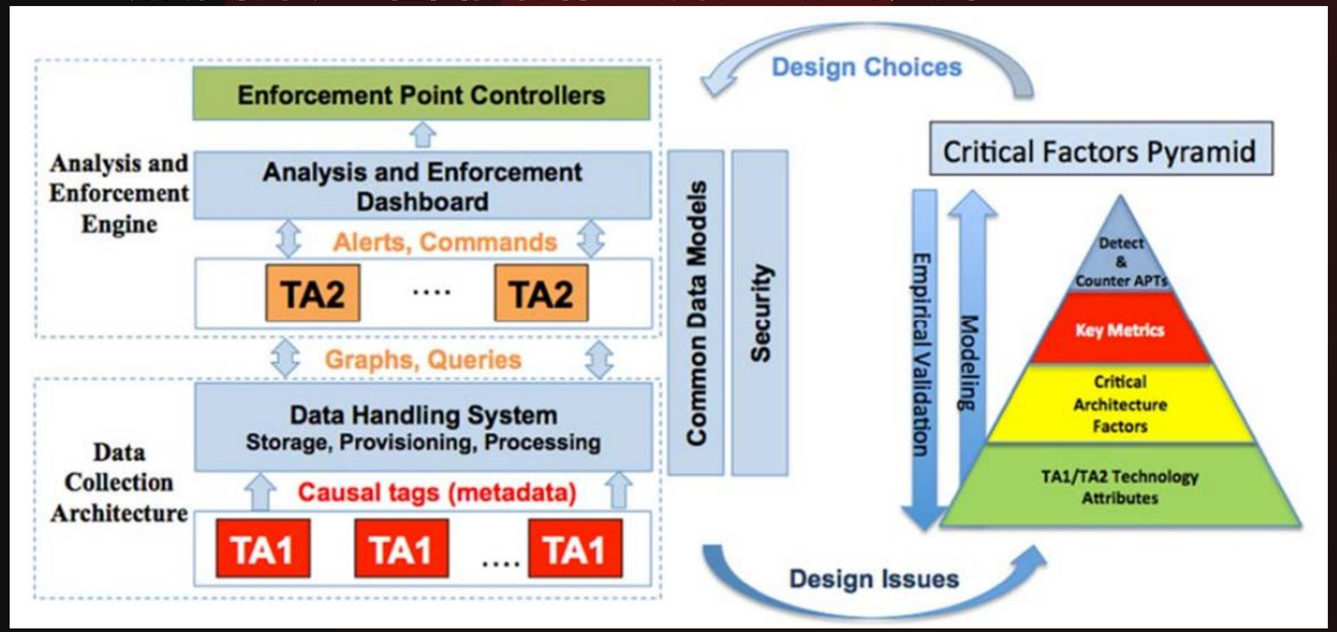


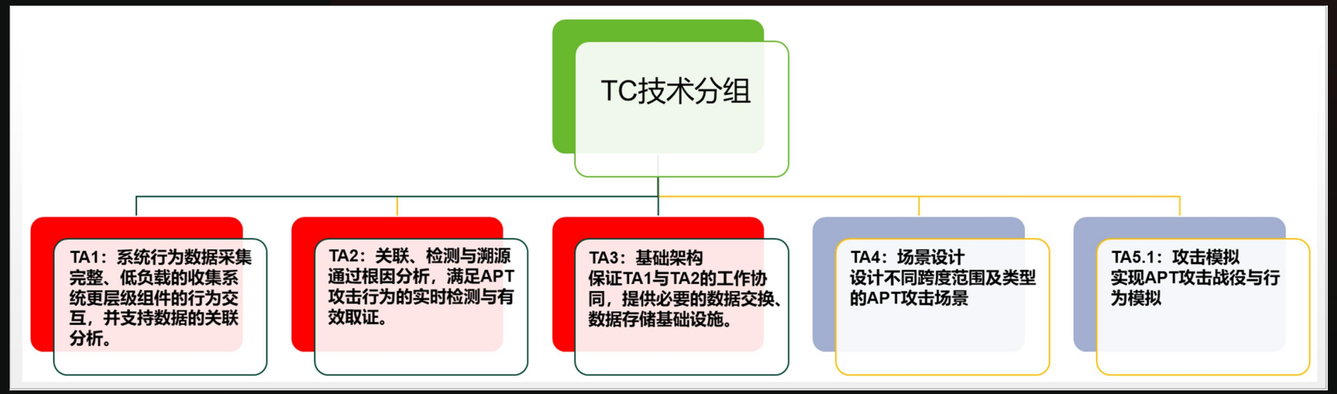
- 面向时序、图、集合、统计的搜索语言井喷式出现
- 溯源图数据搜索平台是溯源、取证的基础设施

KCon 名侦探的召集令： DARPA TC

+1

- DARPA透明计算项目在团队协作、场景上的顶层设计

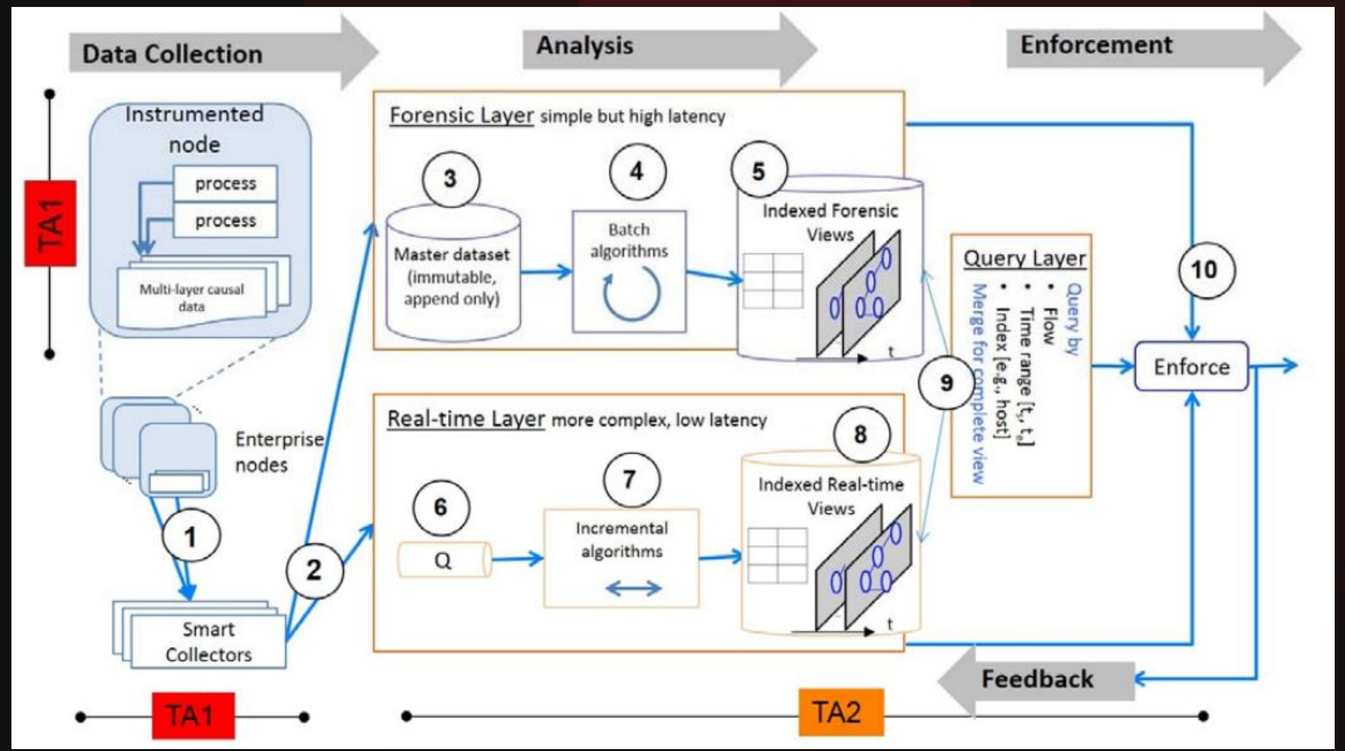


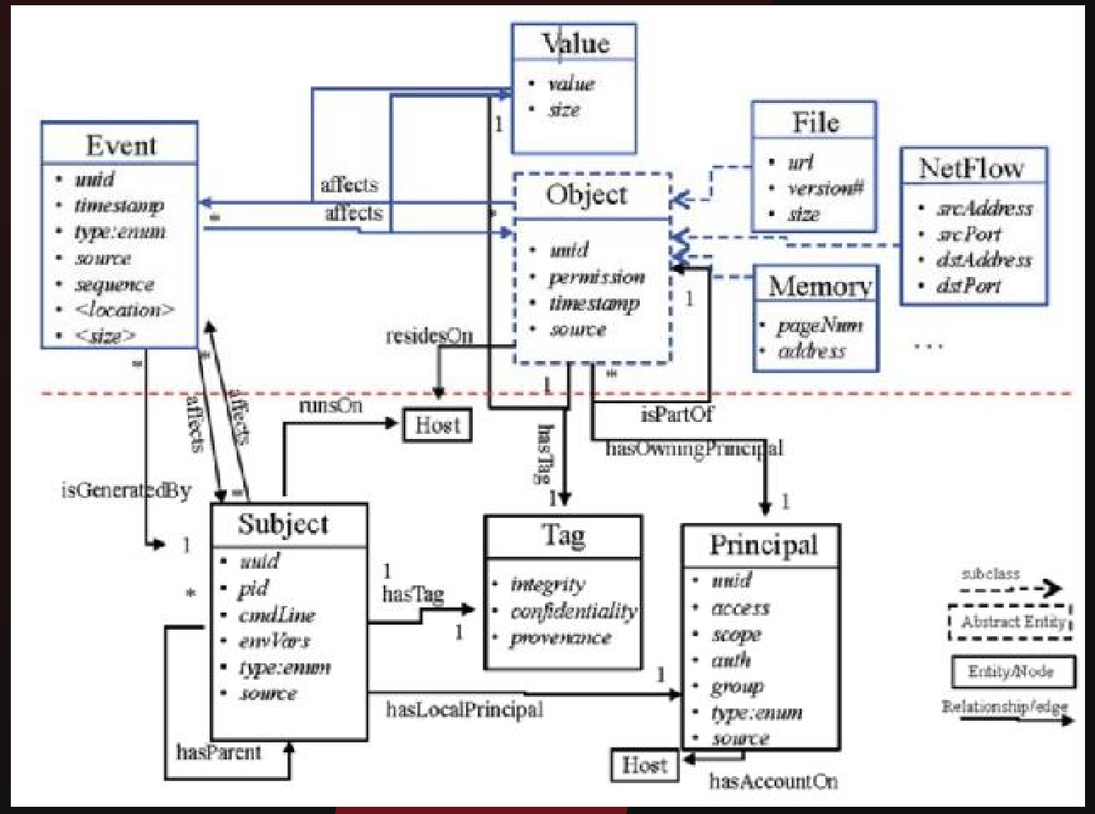


数据集名	数据周期	总事件数量	攻击名称	攻击描述
L-3	263h5m	714M	Firefox backdoor	Firefox backdoor w/ Drakon in-memory: Firefox is exploited by a malicious web site to execute an in-memory payload. This provides a remote console for the attacker.
			Browser extension	Browser extension w/ Drakon dropper: Exploit the victim system using a preexisting malicious Firefox browser extension, drop and execute a malicious file on disk.
			Executable attachment	Phishing e-mail w/ executable attachment: A malicious executable file was sent as an emailattachment, which, after opening, established a connection to the attacker's machine.
F-3	263h28m	21M	Malicious HTTP request	Nginx backdoor w/ Drakon in-memory (4 instances): Attacker exploits Nginx server using a malicious HTTP request. Nginx then downloads and executes several malicious files
L-4	15h28m	36.5M	User-level rootkit	Azazel: Using a preexisting user-level rootkit, the attacker connected to the system using a remote shell and ran reconnaissance commands.
			Ccleaner ransomware	VNC attack: The red team made use of malware named ccleaner that was crafted to evade virus signatures, to organize a highly stealthy ransomware attack
			Recon w/ Metasploit	Metasploit:Malware was downloaded and executed usingMetasploit, giving the attacker remote access. Attacker ran various reconnaissance commands using this capability.
			Kernel malware	Firefox Drakon: In-memory exploit works with a preexisting malicious kernel module for privilege escalation. This allowed the attacker to compromise the sshd server
F-4	11h53m	37.2M	Dropbear Trojan	Dropbear SSH: Using a pre-installed Trojan ssh server, the attacker logged into the victim, ran multiple reconnaissance commands and exfiltrated the results.
			Recon w/ Rootkit	Micro APT: The attacker uploaded two rootkits using scp to the target systems separately, executed them, gained root privilege and ran multiple recon commands

KCon 名侦探的召集令： DARPA TC（Cont.）

- DARPA透明计算项目在分析架构、数据架构上的顶层设计

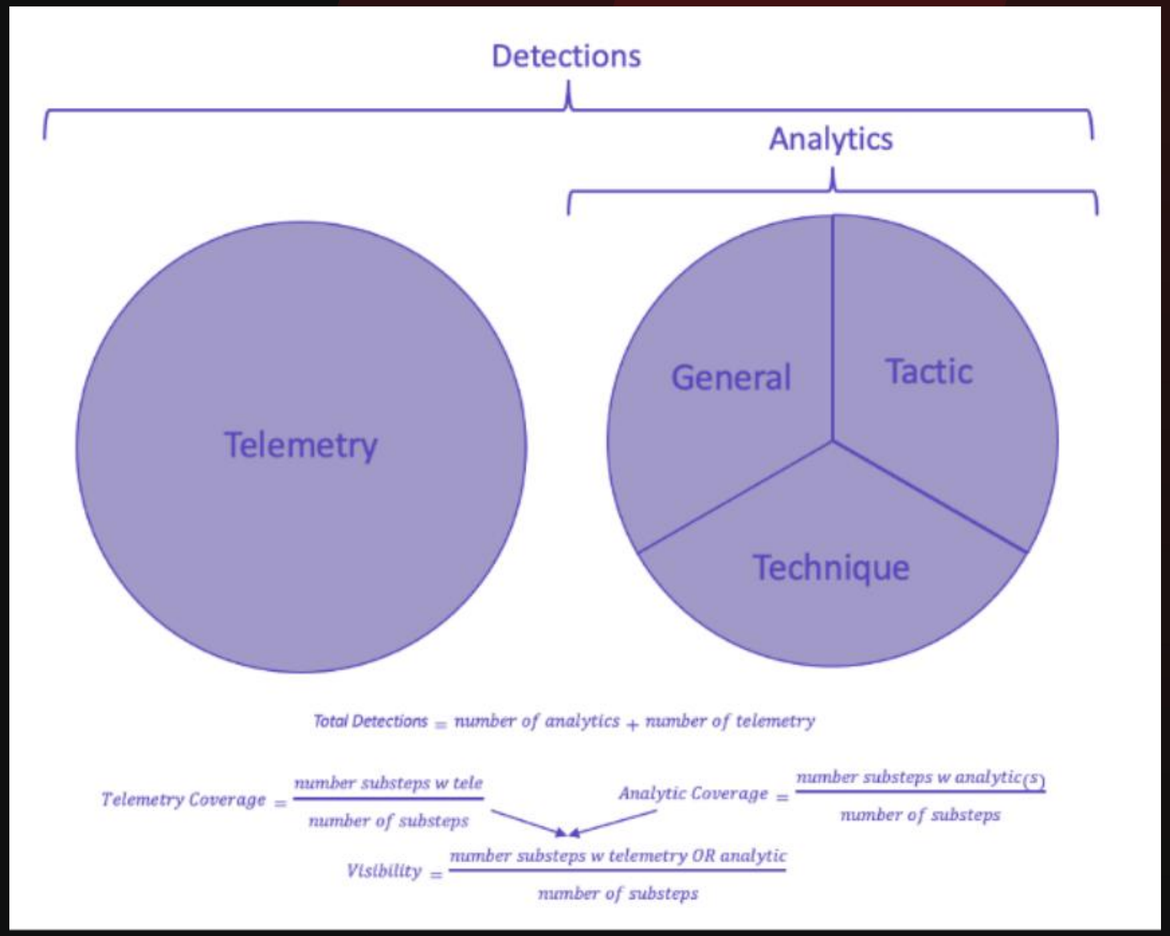


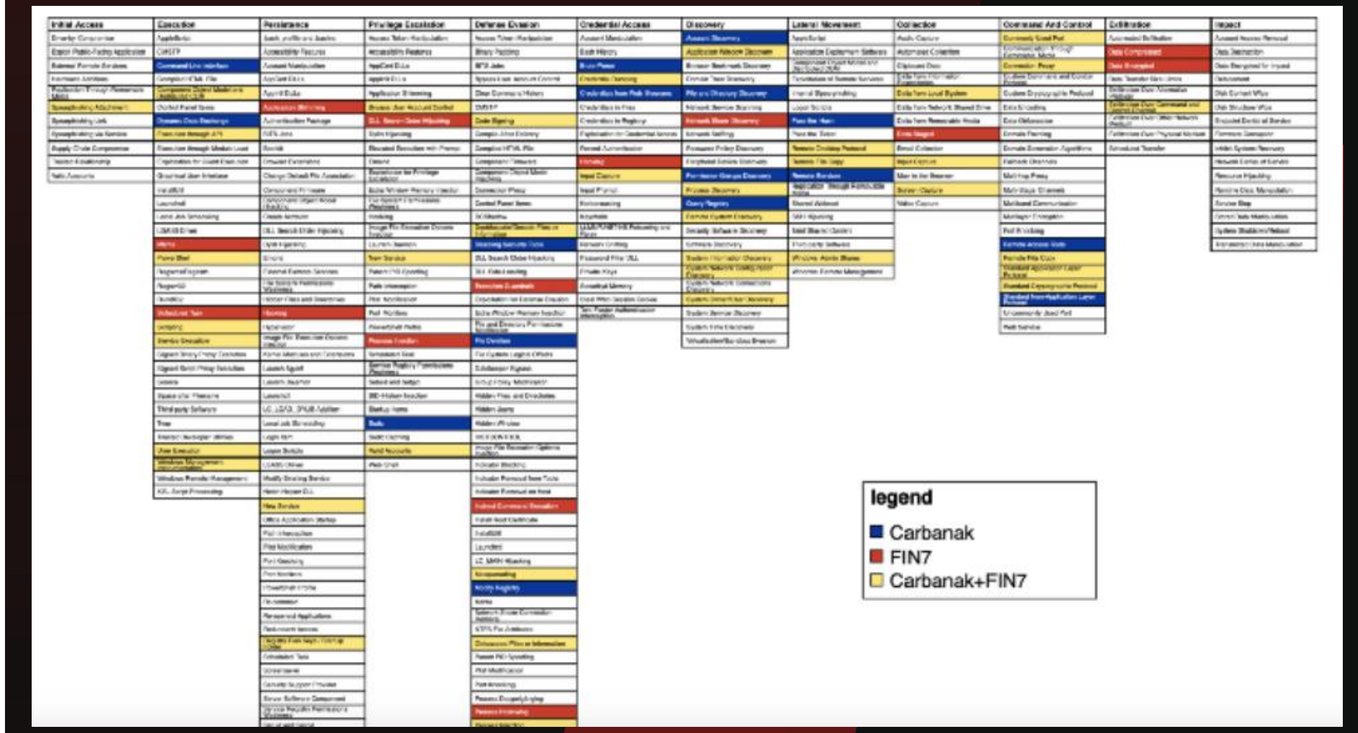


KCon 名侦探的召集令： MITRE ATT&CK Evaluation

+1

- MITRE在工业级应用功能的评估场景、方法、指标的设计





KCon

+1

溯源数据挖掘实践

Practical Thinks

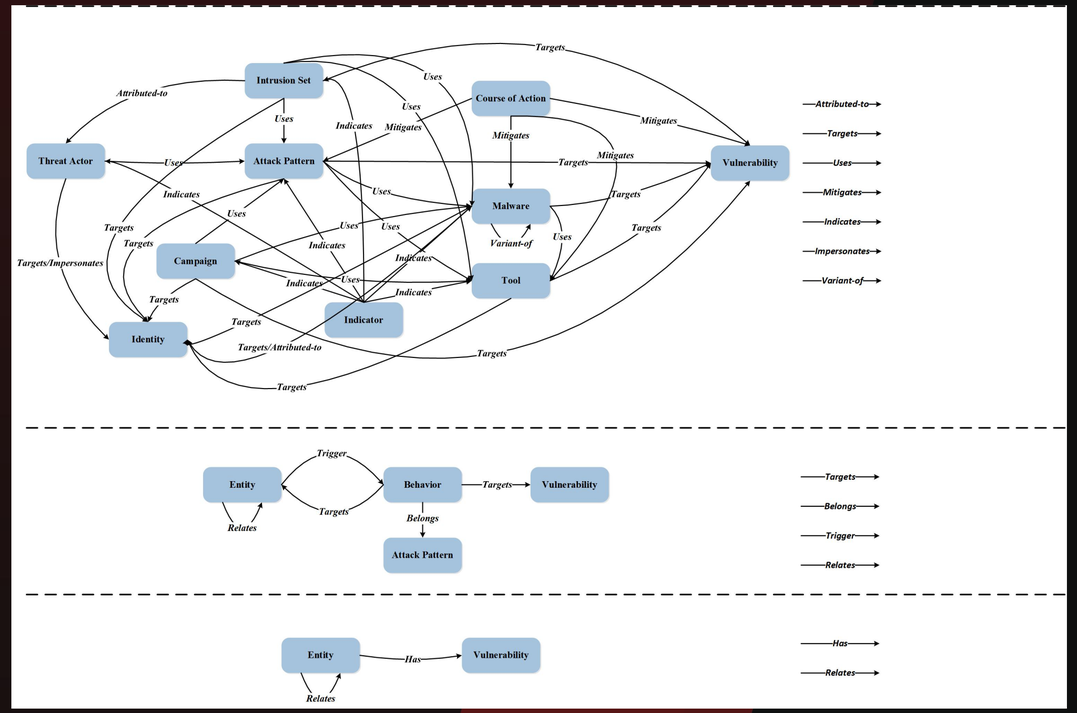
KCon

数据建模篇

+1

科学与艺术的交锋

- 统一处理溯源数据与其他多源数据，如资产、脆弱性、网络侧、知识库等。
- 合理定义处理跨域实体的命名一致性、粒度可分析性。
- 冗余数据的归并与留存，it is a problem.



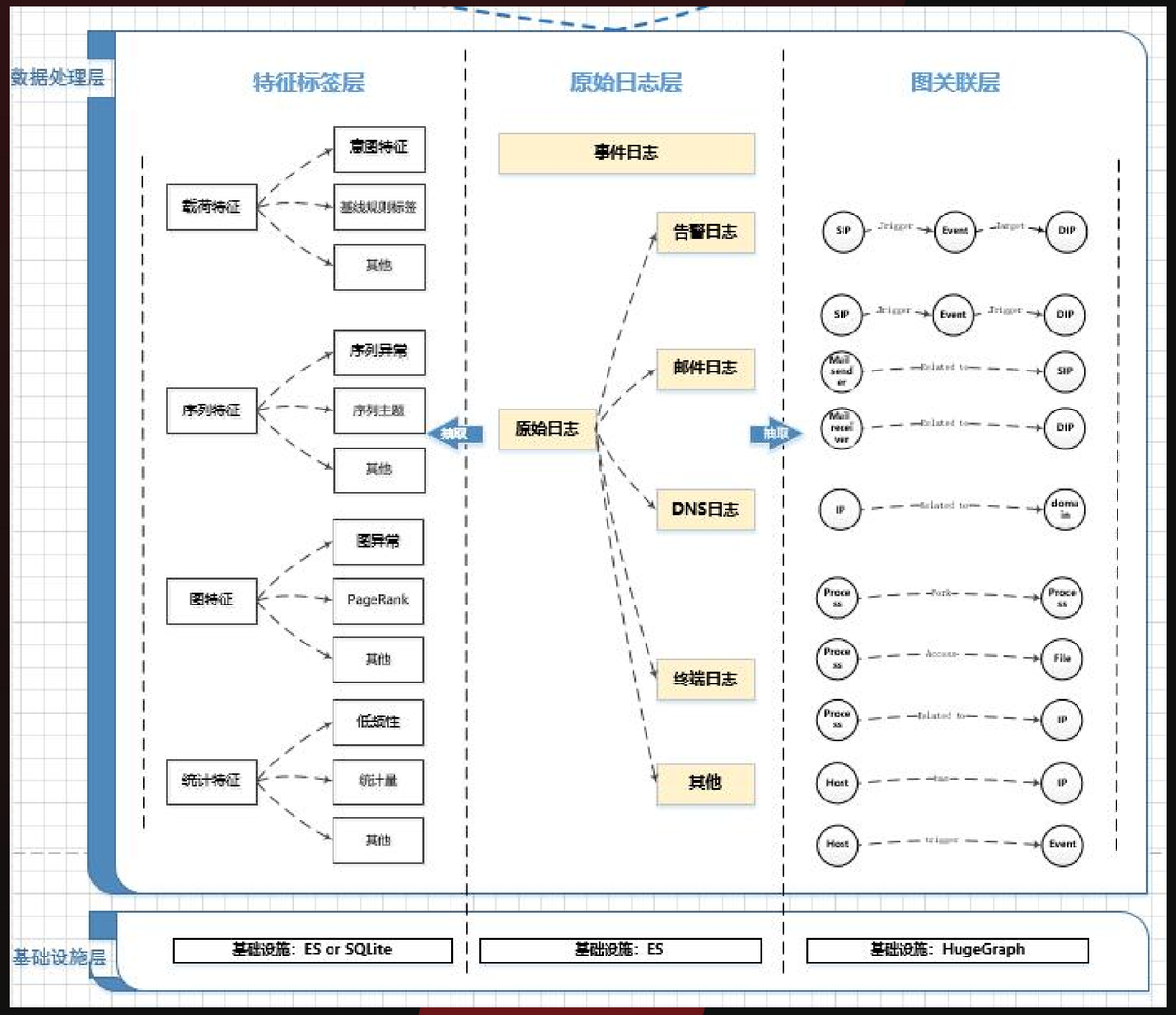
数据分析篇

图搜索基础设施：

- 语言：灵活多态的语言模型设计，提升模式的编排能力
- 存储：关系与属性解耦存储，图数据库与ES架构融合

图分析基础设施：

- 表示：表示学习统一上游数据与下游分析任务
- 检测：轻量级异常检测，支持威胁定位与路径发现，集成可解释方法，支持关键实体、属性定位辅助搜索研判

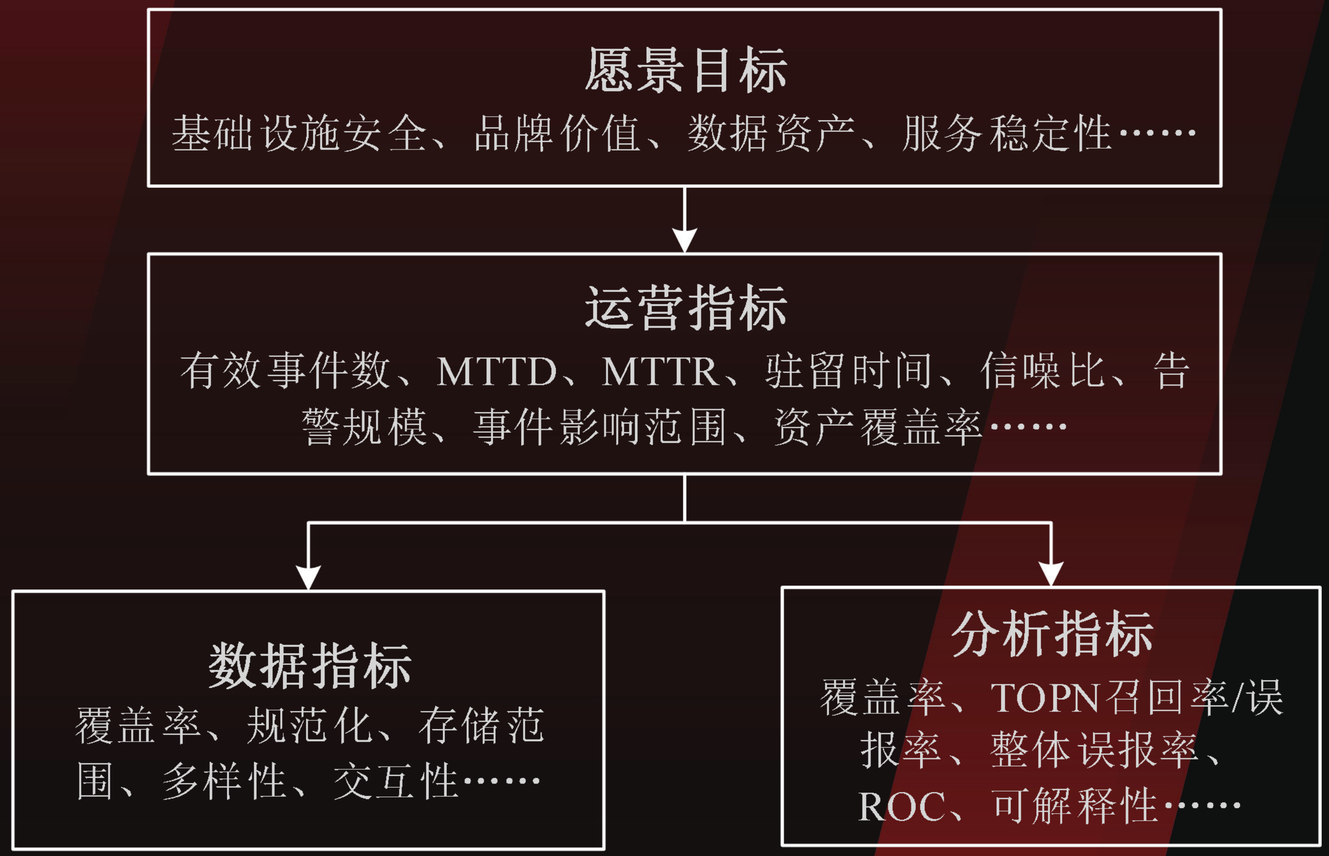


KCon 能力评估篇

+1

考察溯源数据分析技术的效率，几个实践指标

- 构建离线数据集，量化评估
  - 攻击事件的TOPN召回率
  - 关键攻击路径TOPN召回率
  - 攻击路径召回冗余度
- 定性评估类
  - 模型算法的可解释性
  - 算法的稳定性
  - 参数的可控性
  - 标签样本的依赖性



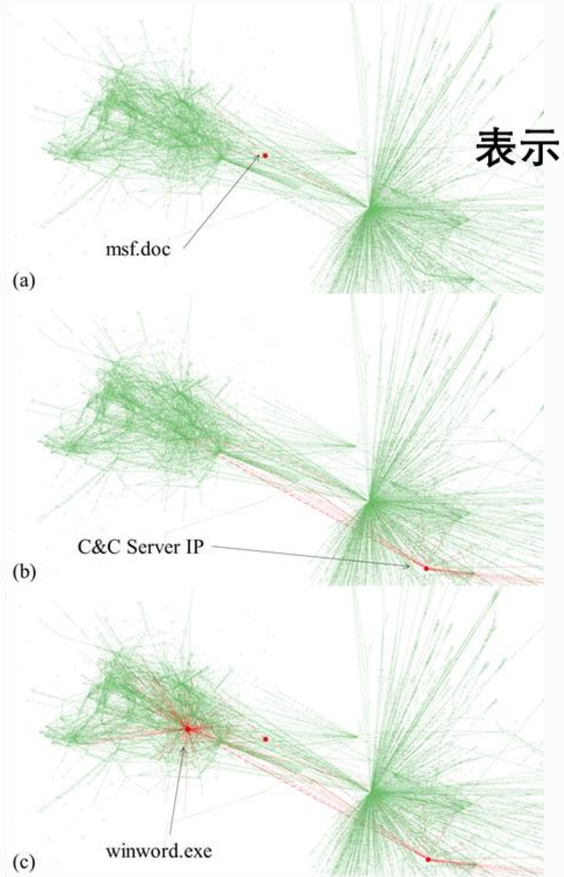
KCon

还是CVE-2017-0199

+1

图搜索基础设施：

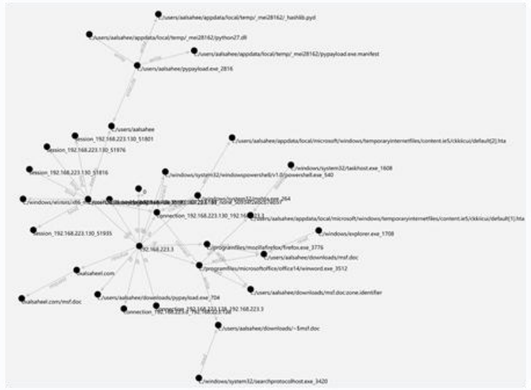
- 大规模溯源图中，表示学习+异常检测+解释模块，定位关键路径及节点的候选集合
- 通过时序、路径搜索，及结构图、时序图可视化，辅助专家研判真实攻击

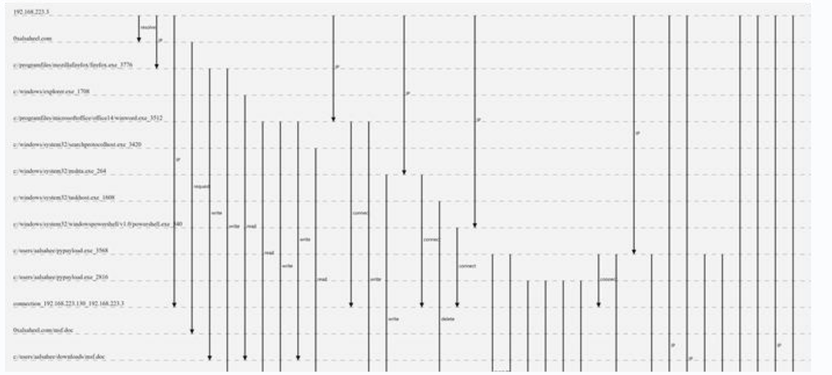


表示学习+可解释

```
malicious_labels:  ['192.168.223.3', 'c:/users/aalsahee/index.html']
[('connected_remote_ip', 0.2261269389720737), ('c:/users/aalsahee/payload.exe_1520', 0.18910913453192663), ('c:/windows/system32/searchprotocolhost.exe_3008', 0.1835993619484456), ('192.168.223.3', 0.17
malicious_labels:  ['192.168.223.3', 'c:/users/aalsahee/payload.exe']
[('c:/users/aalsahee/payload.exe_1520', 0.14556826481502405), ('connected_session', 0.10876634996967442), ('c:/users/aalsahee/appdata/local/temp/tardebb.tmp;c:/users/aalsahee/appdata/local/temp/cabdeba.
malicious_labels:  ['192.168.223.3', '0xalsaheel.com_9999']
[('connected_remote_ip', 0.26302729192801805), ('192.168.223.3', 0.21876128677370643), ('c:/users/aalsahee/payload.exe_1520', 0.20399195622402042), ('c:/programfiles/mozillafirefox/firefox.exe_3148', 0.
malicious_labels:  ['192.168.223.3', '0xalsaheel.com']
[('0xalsaheel.com_9999', 0.25919810762869766), ('connected_remote_ip', 0.2077480995677419), ('session_192.168.223.130_52521;52520', 0.18880184211236412), ('192.168.223.3', 0.18281957450941855), ('192.16
malicious_labels:  ['192.168.223.3', '0xalsaheel.com_9999/18f18fnc']
[('connected_remote_ip', 0.2252116088702944 8), ('c:/programfiles/mozillafirefox/firefox.exe_3148', 0.22482582537208118), ('c:/users/aalsahee/payload.exe_1520', 0.22106034094041552), ('c:/users/aalsahee/
malicious_labels:  ['192.168.223.3']
[('connected_remote_ip', 0.2882447881084565), ('c:/programfiles/mozillafirefox/plugin-container.exe_3396', 0.2357546821168424), ('c:/users/aalsahee/payload.exe_1520', 0.22964156678109876), ('connected_s
----------
entity:  ['192.168.223.3', 'connection_0.0.0.0_255.255.255.255']
[('connection_192.168.223.130_192.168.223.3', -0.15838476149256595), ('resolve', -0.15659524873131037), ('c:/users/aalsahee/payload.exe_1520', 0.15445657869924107), ('connected_remote_ip', 0.14937941680
entity:  ['192.168.223.3', 'connection_fe80__fd1b_d78f_dab1_8114_ff02__1_2']
[('resolve', -0.2229045971898737), ('c:/users/aalsahee/payload.exe_1520', 0.20867412572215097), ('connected_remote_ip', 0.19517420921394027), ('192.168.223.3', 0.19209429856646945), ('connected_remote_i
```

Graph Search





KCon

+1

05

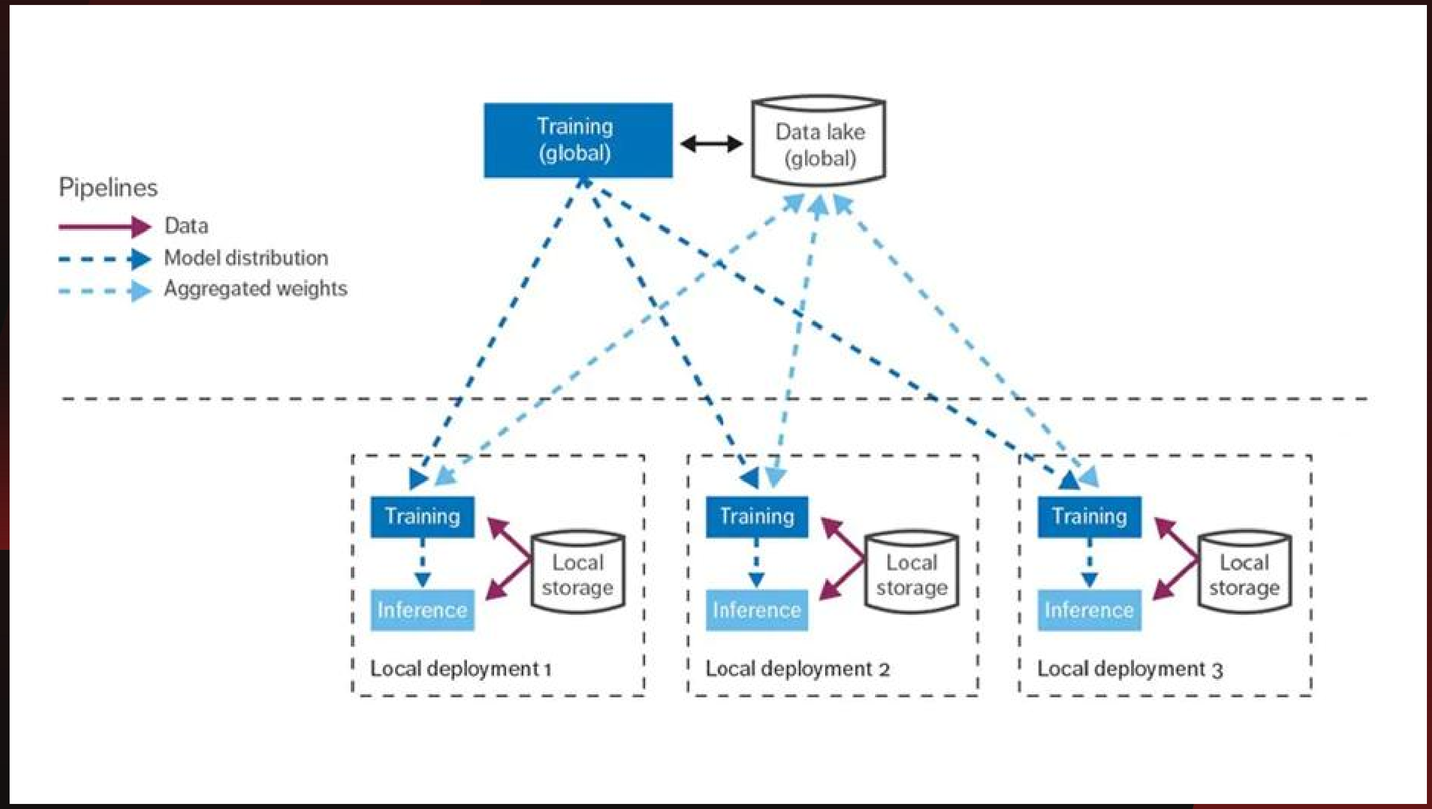
溯源数据挖掘技术发展趋势

Promising Directions

KCon 去中心化架构分析架构

+1

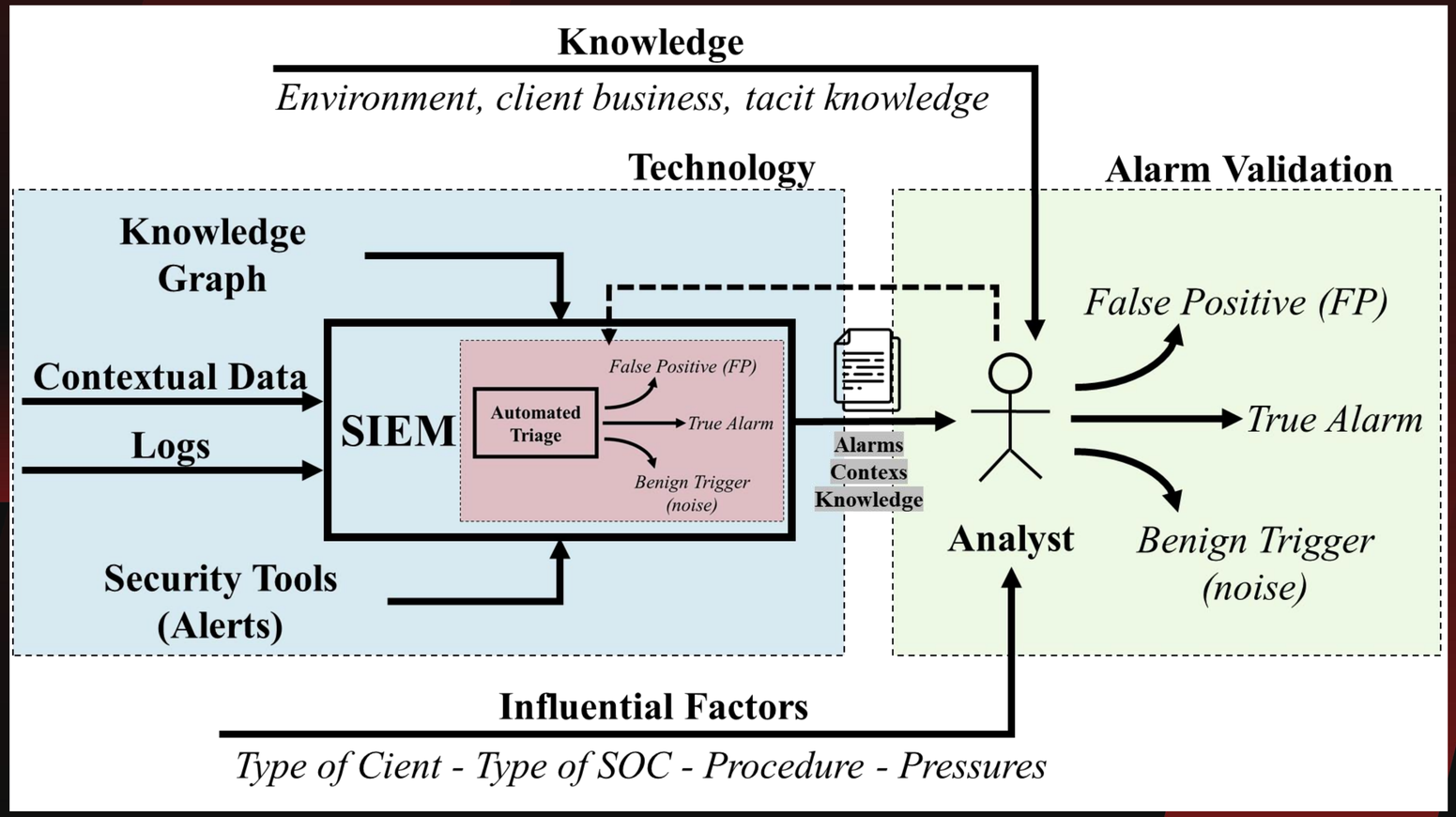
- 去中心化的边缘分析方案，融合分布式模式学习，缓解存储、传输、分析性能瓶颈，降低关键数据资产安全风险



图片来源：https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/privacy-aware-machine-learning

语义与统计结合的鲁棒分析方法

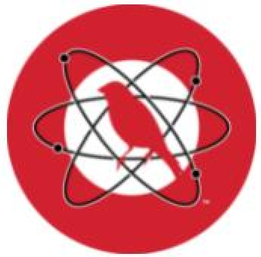
- 结合语义流派（模板与语言）与数据分析（统计与表示）流派的优势，构建能最大化数据价值、最小化经验依赖的分析技术栈



实战化自动化的攻防模拟数据构建

- 围绕ATT&CK等知识库，构建可沉淀的、可管理的场景框架
- 通过实战化、原子化、模板化攻防模拟习得数据与攻防行为之间的映射关系。
- 大型的、标准的、取得共识的统一评测，促进数据的标准化与兼容性提升。

SANS Webcast on MITRE ATT&CK® and Sigma



Atomic Red Team

CALDERA™

Full documentation, training and use-cases can be found here.

CALDERA™ is a cyber security framework designed to easily automate adversary emulation, assist manual red-teams, and automate incident response.

It is built on the MITRE ATT&CK™ framework and is an active research project at MITRE.

The framework consists of two components:

1. **The core system.** This is the framework code, consisting of what is available in this repository. Included is an

rules/

Rules within this folder are organized by solution or platform. The structure is flattened out, because nested file hierarchies are hard to navigate and find what you're looking for. Each directory contains several .toml files, and the primary ATT&CK tactic is included in the file name when it's relevant (i.e. `windows/execution_via_compiled_html_file.toml`)

KCon About Me

- 绿盟科技创新研究院/天枢实验室，负责数据科学+安全运营攻防的AISecOps技术团队
- 安全运营中的自动化与智能化技术前沿研究与落地应用攻关
- 目标定位是可实战的、可运营、可信任、有挑战的 安全数据科学技术领域

欢迎关注与交流：

绿盟科技研究通讯公众号：nsfocus_research

个人微信号：oasiszrz

网安数据科学与知识工程技术研讨

该二维码7天内(9月3日前)有效，重新进入将更新

KCon

主要参考文献

[1] RSA, Hypothesis in Threat Hunting, https://www.rsa.com/en-us/blog/2017-07/hypothesis-in-threat-hunting

[2] Hossain M N, Sheikhi S, Sekar R. Combating Dependence Explosion in Forensic Analysis Using Alternative Tag Propagation Semantics[J].

[3] Hassan W U, Guo S, Li D, et al. NoDoze: Combatting Threat Alert Fatigue with Automated Provenance Triage[C]. NDSS, 2019.

[4] Shu X, Araujo F, Schales D L, et al. Threat intelligence computing[C]. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018: 1883-1898.

[5] Milajerdi S M, Eshete B, Gjomemo R, et al. Poirot: Aligning attack behavior with kernel audit records for cyber threat hunting[C]. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019: 1795-1812.

[6] Milajerdi S M, Gjomemo R, Eshete B, et al. Holmes: real-time apt detection through correlation of suspicious information flows[C]. 2019 IEEE Symposium on Security and Privacy (SP), 2019: 1137-1152.

[7] Pei K, Gu Z, Saltaformaggio B, et al. Hercule: Attack story reconstruction via community discovery on correlated log graph[C]. Proceedings of the 32Nd Annual Conference on Computer Security Applications, 2016: 583-595.

[8] https://www.darpa.mil/program/transparent-computing

[9] Xu Z, Fang P, Liu C, et al. DEPCOMM: Graph Summarization on System Audit Logs for Attack Investigation[C]. IEEE Symposium on Security and Privacy (SP), San Francisco, CA, 2021: 22-26.

[10] Zeng J, Wang X, Liu J, et al. Shadewatcher: Recommendation-guided cyber threat analysis using system audit records[C]. 2022 IEEE Symposium on Security and Privacy (SP), 2022: 1567-1567.

[11] https://cybersecurityworks.com/blog/vulnerabilities/google-trends-most-searched-top-10-vulnerabilities-in-2020.html

[12] http://blog.nsfocus.net/microsoft-office-ole2link-exploits-technology-analysis/

[13] Alsaheel A, Nan Y, Ma S, et al. {ATLAS}: A Sequence-based Learning Approach for Attack Investigation[C]. 30th {USENIX} Security Symposium ({USENIX} Security 21), 2021.

[14] https://www.fireeye.com/blog/threat-research/2017/04/cve-2017-0199-hta-handler.html

[15] Zeng J , Zheng L C , Chen Y , et al. WATSON: Abstracting Behaviors from Audit Logs via Aggregation of Contextual Semantics[C]// Network and Distributed System Security Symposium. 2021.

KCon

+1

感谢您的观看

THANK YOU FOR YOUR WATCHING

KCon 2022 黑客大会