



基于数据挖掘的安全事件分析

张博锋

国防科技大学计算机学院
网络与信息安全研究所



- 背景
- 数据挖掘方法框架
- 安全事件的相似性度量
- 安全事件的聚类
- 攻击模式的提取
- 在线优化
- 大数据带来的挑战

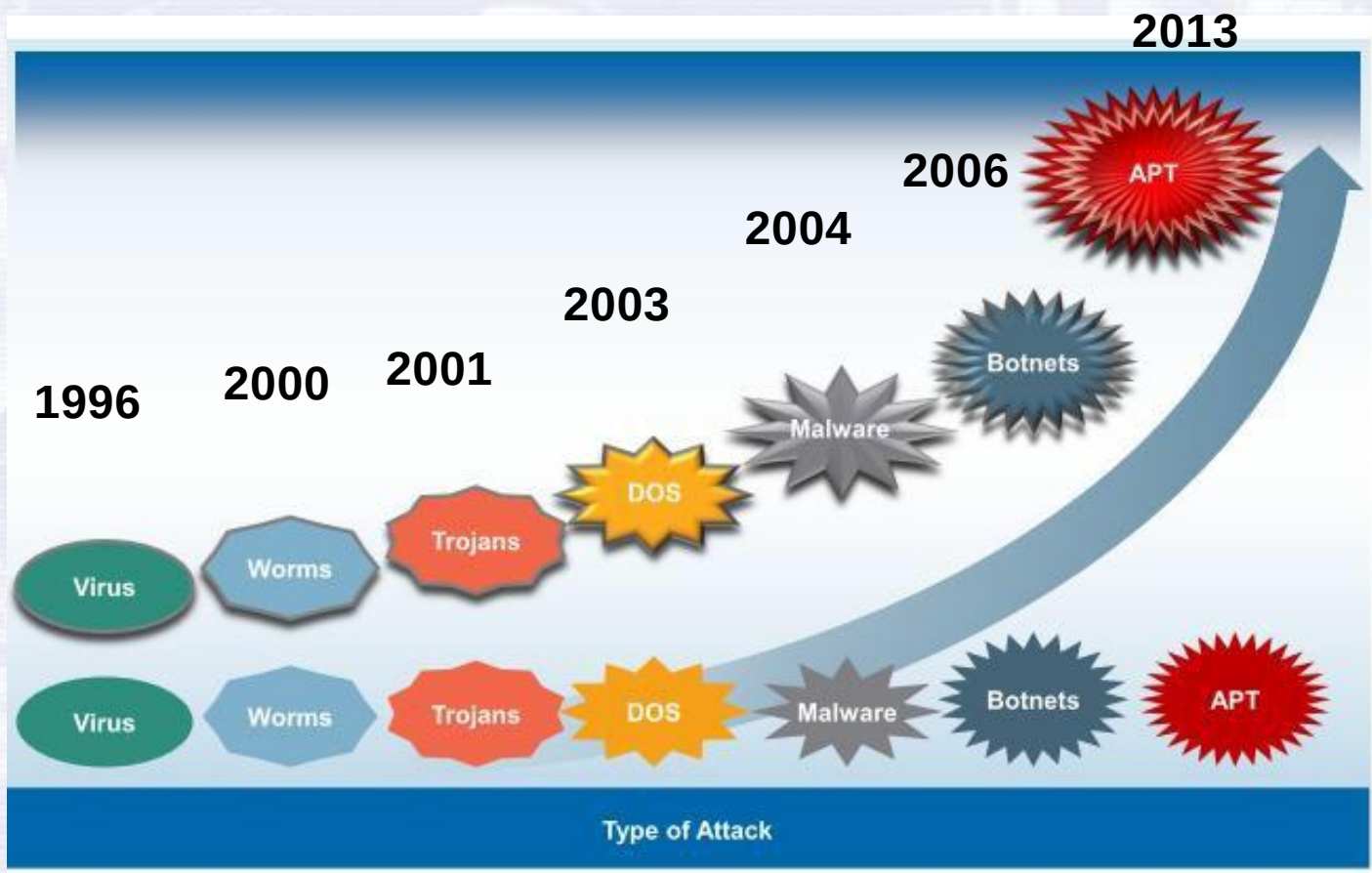


背景



网络威胁的改变

高
攻击手法的精细程度
低



资料来源：Juniper 2013



- APT是现今面临的最大威胁，其典型意图为从目标组织窃取智力成果，获取用户敏感数据，或者商业决策信息



2010：震网攻击伊朗核电站



2009：极光攻击



2011：窃取RSA令牌种子



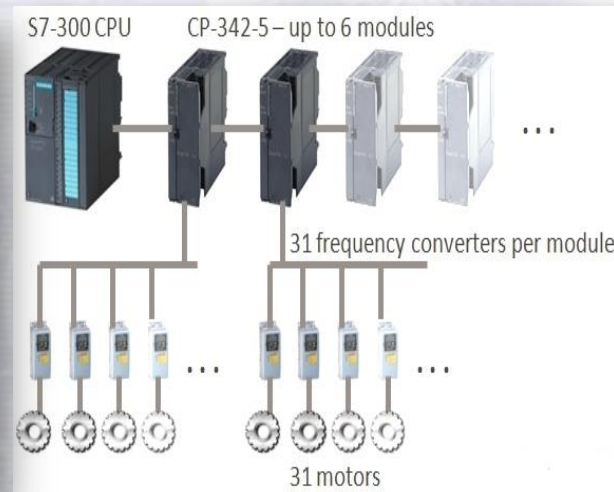
2012：窃取NASA资料



2012：大量攻击中东多年



❖ “震网”病毒(Stuxnet)对伊朗核设施的网络攻击



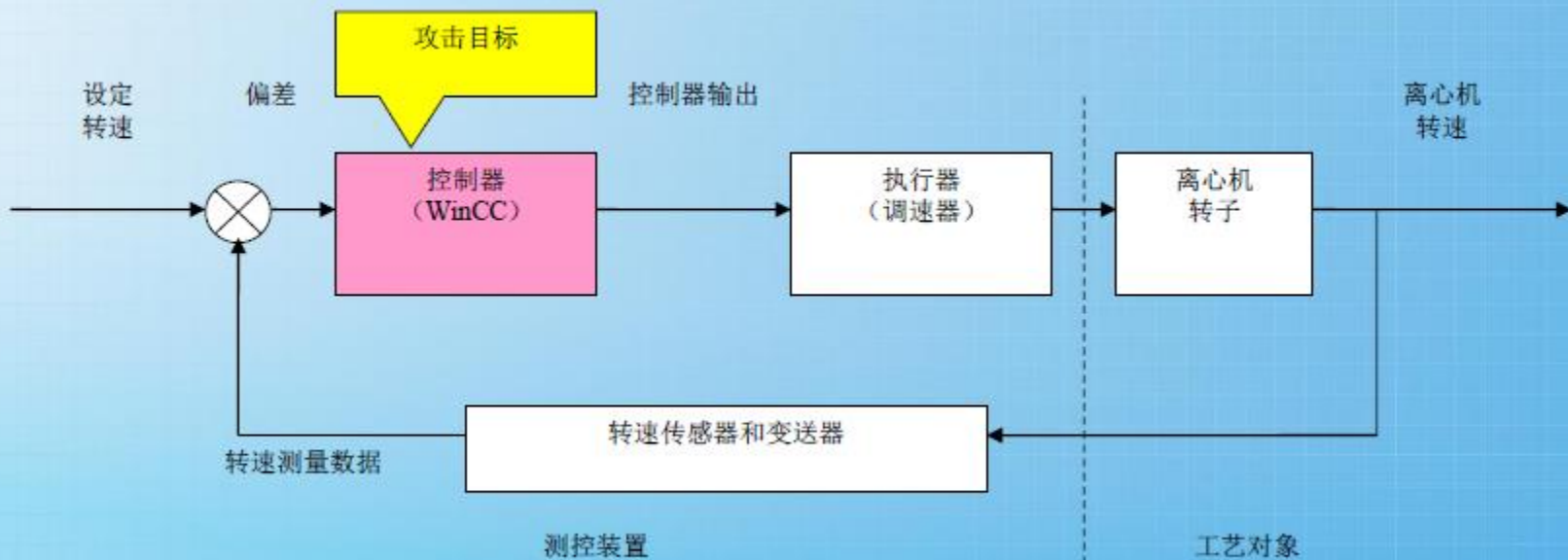
- ❑ 2010年，位于纳坦兹的铀浓缩工厂遭受到“震网”病毒的攻击
- ❑ 控制系统采用Windows操作系统和西门子SIMATIC WinCC控制软件
- ❑ 在遭受到“震网”蠕虫的攻击之后，伊朗大量用于提炼浓缩铀的离心机发生连串故障



复杂网络威胁

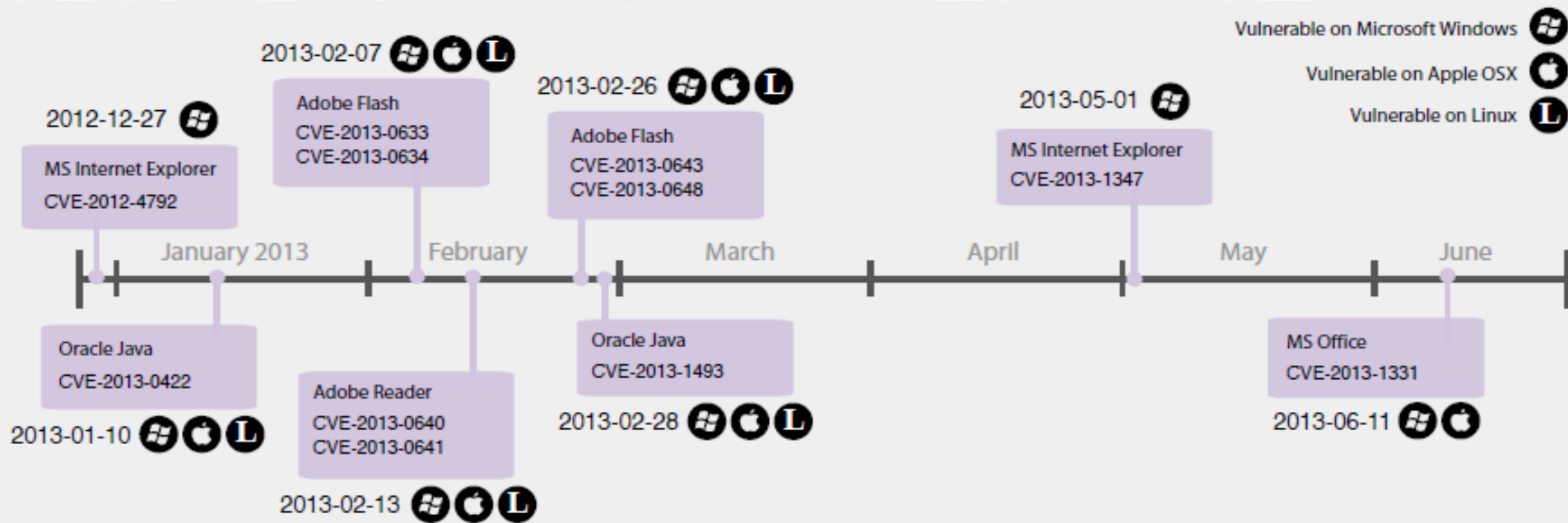
2010年7月，震网病毒(stuxnet)首先感染外部主机；然后感染U盘，利用微软操作系统快捷方式文件解析漏洞，传播到内部网络；在内网中，通过四个零日漏洞和一个已知漏洞，实现系统提权和联网主机之间的传播；最后抵达安装了西门子WinCC工业控制软件的主机，展开对离心机可编程控制器的攻击，修改离心机的转速。

资料来源：CNVD分析报告、互联网





- 传统的方法缺少相应的检测特征，特别是威胁行对0-day漏洞的利用，对发生的攻击行为易造成漏报



2013 H1 Zero-Day Timeline

(point in time when zero-day attack became publicly known)

资料来源: IBM X-Force 2013 Mid-Year Trend and Risk Report



- 传统的方法严重依赖于安全分析专家实施手工调查，而手工调查劳动强度高，通常是case by case的
- 异常检测方法普遍关注于显著的异常点，具有较高的误报率，难以应对复杂网络威胁的隐蔽攻击方式



- 网络数据成指数级增长，2009年18个月翻倍，而现在每11个月翻倍
- 麦肯锡全球学会（MGI）认为对数据的深入分析将为包括安全在内的各部门带来分析压力



美国医疗

- \$3千亿每年
- 每年增长~0.7%



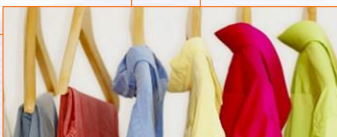
欧洲公共部门管理

- €2.5千亿每年
- 每年增长~0.5%



全球个人位置数据

- 服务商年收益逾\$1千亿
- 终端用户高达\$7千亿



美国零售业

- 可带来60%的销售增长
- 每年增长0.5-1.0%



制造业

- 减少50%产品开发成本
- 减少7%的工作支出



- 对数据的有效分析能够改善信息安全以及态势感知的能力
- 数据驱动的信息安全可回溯到银行欺诈和基于异常的入侵检测系统





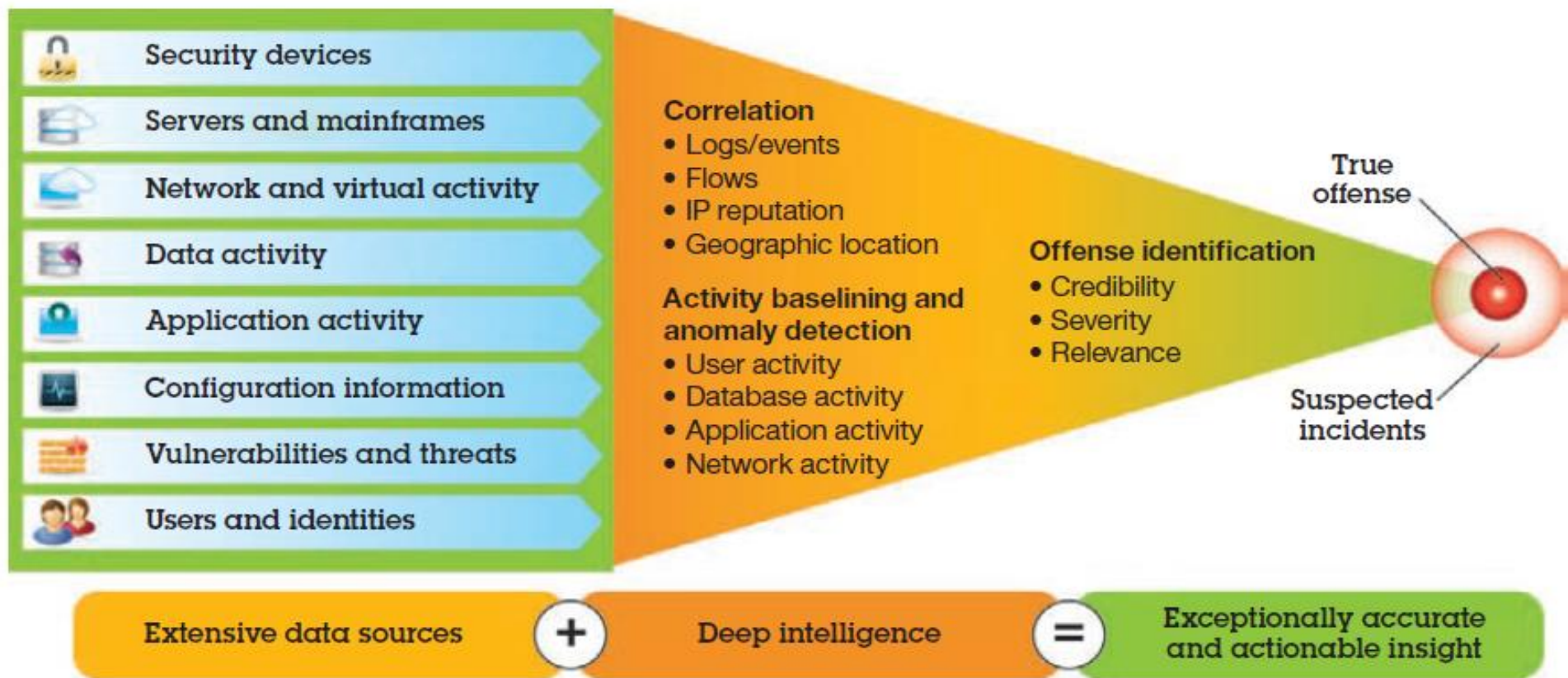
- 数据分析在入侵检测系统中经历3个阶段的发展：
 - 层次化安全的入侵检测系统。主要是被动安全保护和入侵响应机制
 - 安全信息和事件管理（SIEM）。管理不同探针的警报和规则，合并、过滤来自各个源的警报数据，将有用信息呈现给安全分析师
 - 大数据分析。大数据分析工具将显著改善安全事件信息的关联、合并和上下文分析处理，并在取证中有能力关联长期的历史数据。



- **RSA安全实验室指出无论攻击多么精巧，攻击者的行为还是会形成不同于正常模式的行为**
- **由于复杂网络攻击行为包含多个阶段，攻击者每次攻击行为都为检测提供了机会**
- **关联上述看似无关的事件有助于发现入侵证据，揭示传统方法无法识别的隐蔽攻**



- 关键是将非结构化数据和多个无关的数据集融合进一个统一的深度分析框架





数据挖掘分析框架



■ 研究的关键技术

- 面临海量安全事件数据，对长时间、全记录数据的深度分析
- 安全事件高层视图的构建
- 原始安全事件中的冗余数据的高效约减
- 网络威胁行为的还原与重建
- 网络威胁模式的挖掘提取
- 网络攻击意图的识别

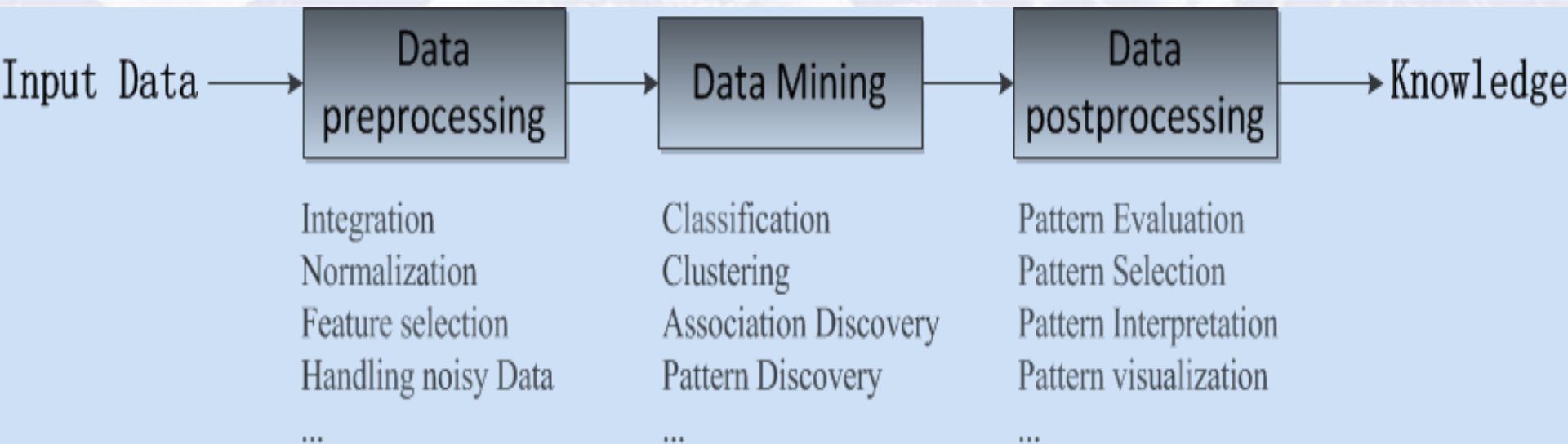


- 在数据分析与知识获取中，数据挖掘过程主要包括以下4个方面：
 - 数据获取
 - 数据规格化
 - 数据度量
 - 数据挖掘算法
- 在实现层面，主要考虑技术在单服务器上的实现效率



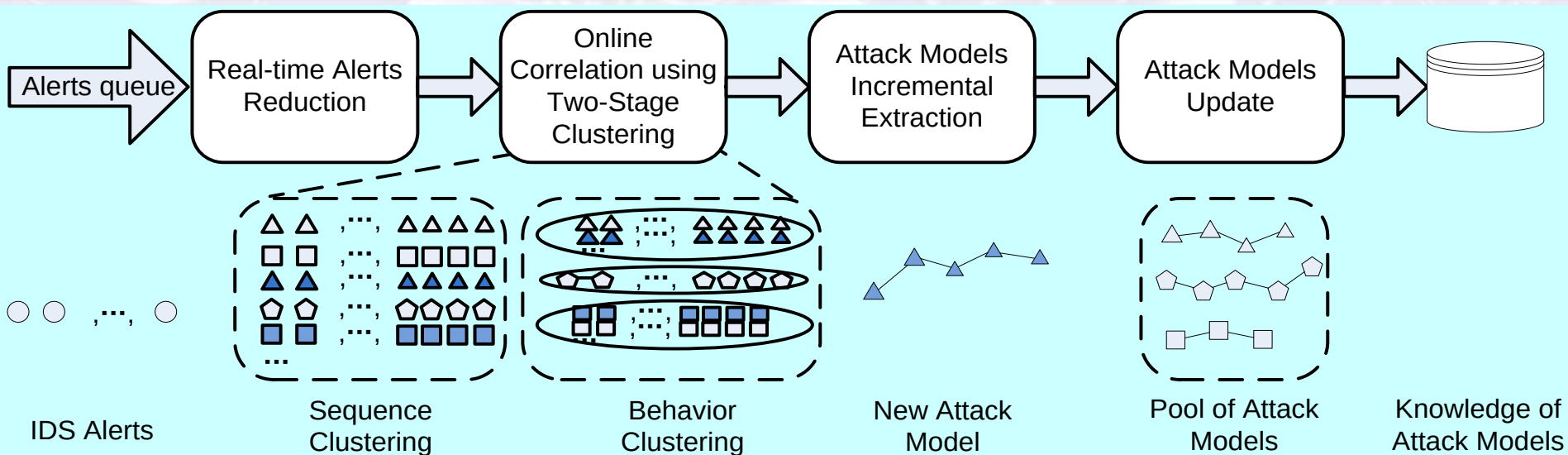
■ 一个典型的数据挖掘过程为：

- 数据预处理
- 应用数据挖掘算法
- 数据后处理



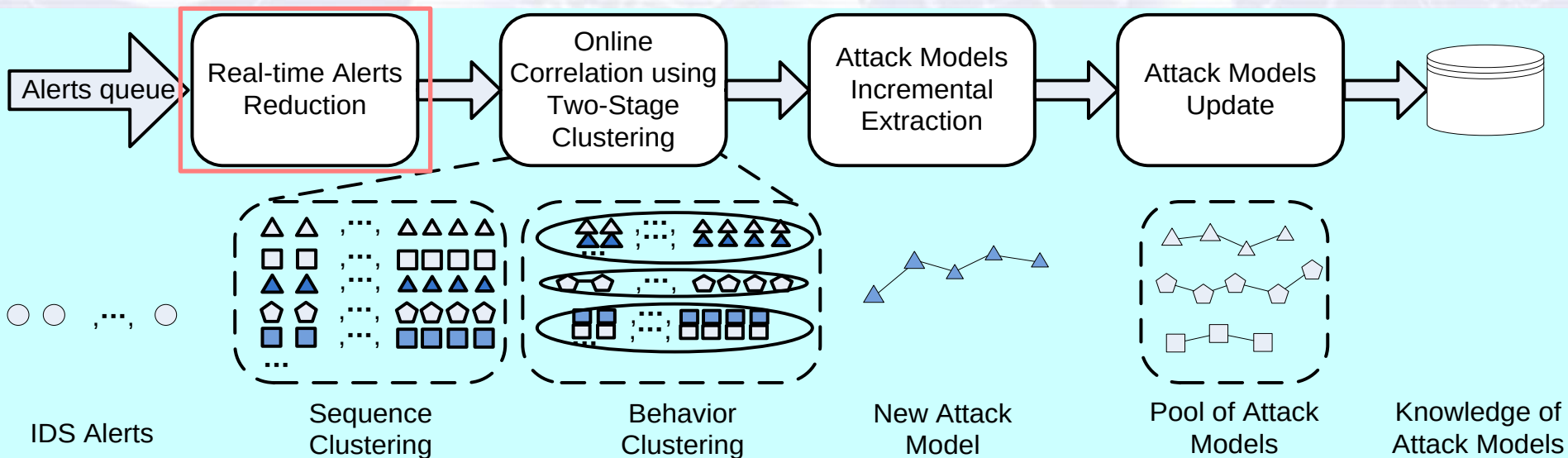


- 基于数据挖掘的高速网络安全事件分析——我们的框架
 - 实时的网络安全事件约减方法
 - 在线的事件聚类
 - 增量的攻击模型提取





- 基于统计冗余特征的安全事件约减方法
 - 减少传统入侵检测系统在高速网络环境下产生安全事件的数据量
 - 过滤冗余安全事件
 - 提高结果的生成质量和后续步骤的处理效率





■ 网络攻击事件约简指标与方法

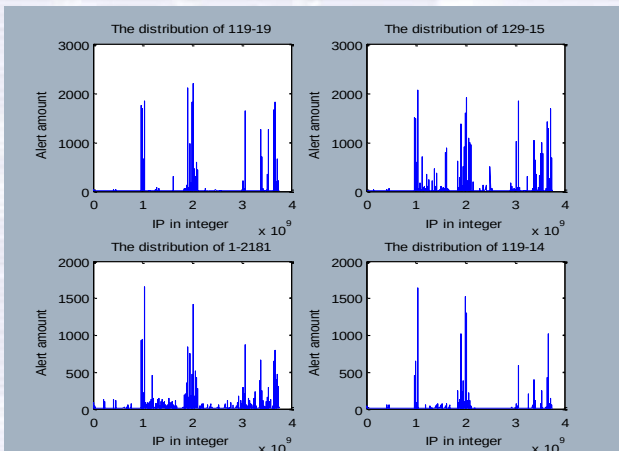
- 2%的事件占到91%的分布冗余，事件分布在不同IP空间相似
- 基于事件类型统计特征的冗余度指标
- 基于阶段性采样的冗余事件约简

数量冗余度 $\rightarrow$

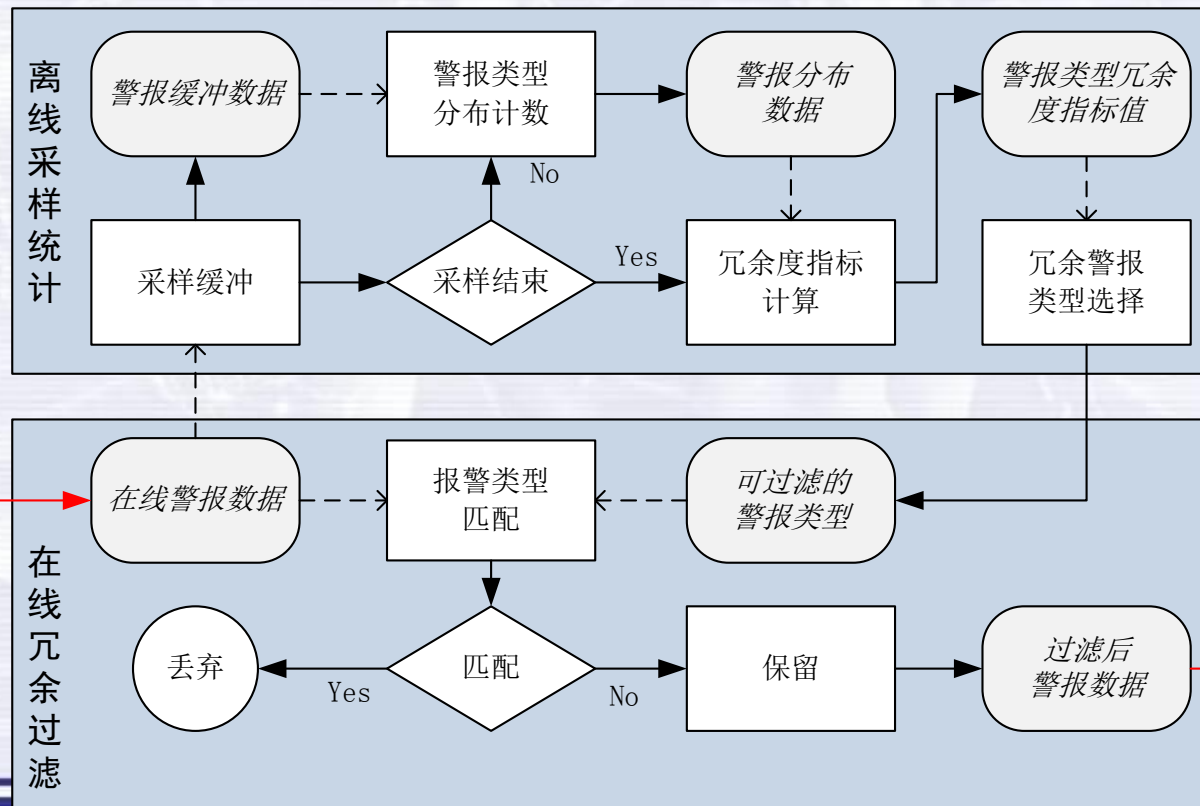
$$\tau(t) = \left(\frac{N(t)}{N} \right)^\alpha \cdot \frac{\sum_{i=1}^N M_i(t)}{\sum_{i=1}^N M_i}$$

会话冗余度

网络事件冗余度指标

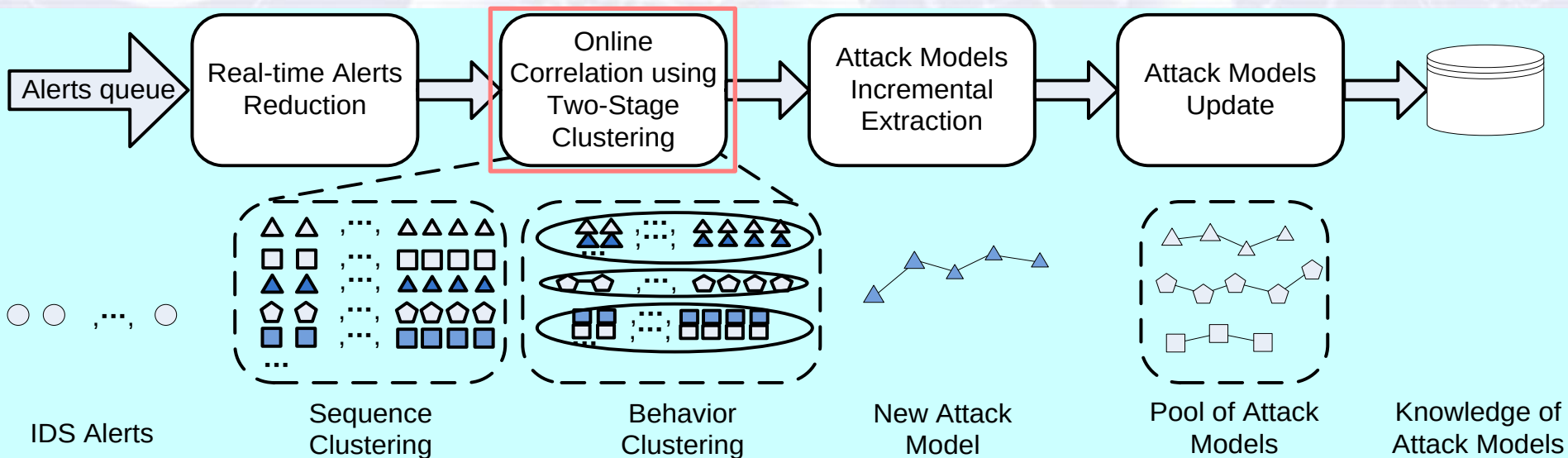


冗余网络事件分布相似性(采样依据)





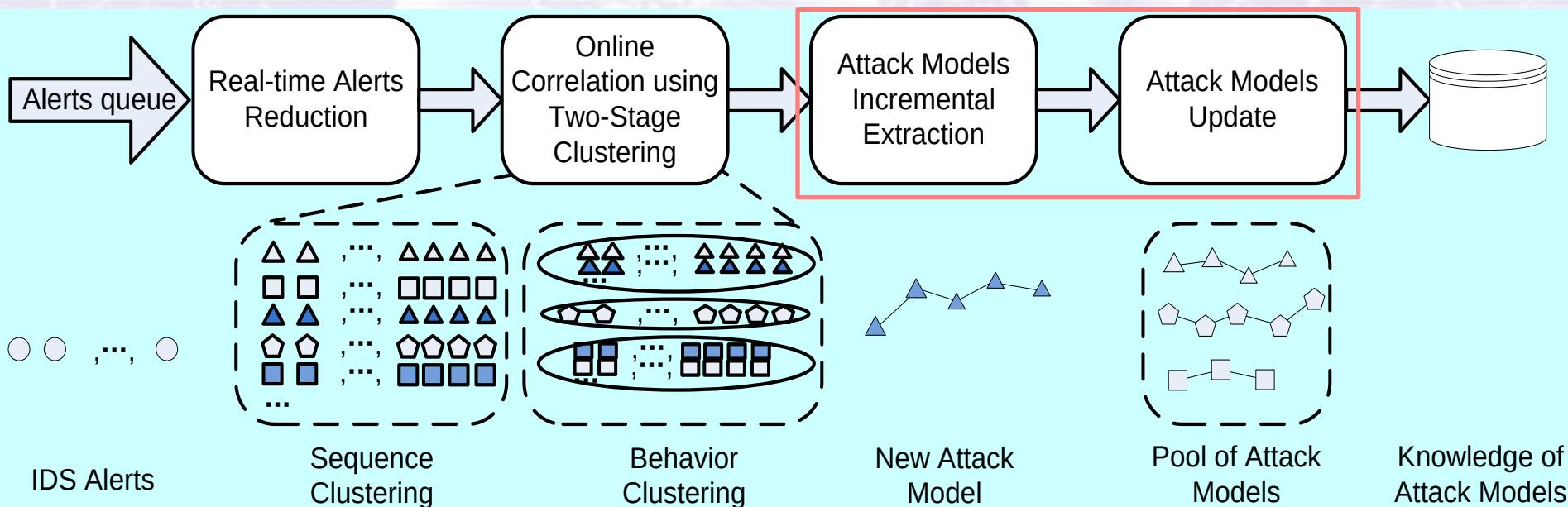
- 基于两阶段聚类的安全事件关联分析方法
 - 提升现有入侵检测系统对复杂网络威胁行为产生的安全事件数据层次
 - 将零碎的安全重新组织，提供高层次威胁行为视图





■ 增量的攻击模式生成方法

- 缩短网络威胁模式更新速度滞后时间差
- 减轻安全专家人工分析劳动强度
- 提高模式提取完整度和效率
- 生成可用于分析和检测完整攻击意图的知识





安全事件的相似性度量



■ 相似性度量的含义

- 按照一定标准，将难以相互比较的指标原始资料转换为可相互比较的数值
- 元素间距离的计算对于聚类算法的过程十分重要，直接影响着算法的有效性
- 数据挖掘中聚类算法的一个关键是对数据的度量，计算各元素间距离，判断各个元素之间的相似性
- 在多阶段次化聚类分析中，需要对每阶段聚类元素分别加以度量



- 在两阶段网络安全事件聚类分析中，需要分别度量的元素为：
 - 网络事件
 - 网络威胁行为
- 网络事件时空相似性度量
 - 根据网络事件时空差异进行度量
 - 网络事件特征向量
 - 事件源目的地址
 - 事件发生时间
 - 事件类型



■ 网络事件时空相似性度量

□ 事件源目的地址

- 相邻的主机遭受同一攻击行为的可能性较高
- 对地址空间上较为临近的事件赋予较高的相似度
- 根据IP地址前缀匹配长度计算
- 地址相似性度量函数: $SAD(a_1, a_2)$

$$h_1(a_1, a_2) = h(a_1.IPsrc, a_2.IPsrc) + h(a_1.IPdst, a_2.IPdst)$$

$$h_2(a_1, a_2) = h(a_1.IPsrc, a_2.IPdst) + h(a_1.IPdst, a_2.IPsrc)$$

$$SAD(a_1, a_2) = \frac{1}{L} \max\{h_1(a_1, a_2), h_2(a_1, a_2)\}$$



■ 网络事件时空相似性度量

□ 事件发生时间

- 同一威胁行为的各个攻击步骤时间分布上相近
- 对时间分布较近的事件赋予较高的相似度
- 根据事件发生的时间间隔来计算
- 时间相似性度量函数： $STM(a_1, a_2)$

$$STM(a_1, a_2) = \begin{cases} 1, & |a_1.time - a_2.time| < \Delta t, \\ \frac{\Delta t}{|a_1.time - a_2.time|}, & \text{else.} \end{cases}$$



■ 网络事件时空相似性度量

□ 事件类型

- 同一威胁行为所引发的事件类型多属于同一系列
- 事件类型属于同一系列的应当赋予较高的相似度
- 重复的事件类型应当滤除
- 事件类型相似度函数： $STM(a_1, a_2)$

$$STP(a_1, a_2) = \begin{cases} 0, & a_1.type \neq a_2.type, \\ 1, & a_1.category = a_2.category \\ 0, & \text{else.} \end{cases}$$



■ 网络事件时空相似性度量

- 事件之间的相似度函数定义为各特征相似度函数的加权和：

$$SIM(a_1, a_2) = \sum_{i \in I} w_i \cdot sim_i(a_1, a_2)$$

$$SAD(a_1, a_2) = \frac{1}{L} \max\{h_1(a_1, a_2), h_2(a_1, a_2)\}$$

$$STM(a_1, a_2) = \begin{cases} 1, & |a_1.time - a_2.time| < \Delta t, \\ \frac{\Delta t}{|a_1.time - a_2.time|}, & \text{else.} \end{cases}$$

$$STP(a_1, a_2) = \begin{cases} 0, & a_1.type \neq a_2.type, \\ 1, & a_1.category = a_2.category \\ 0, & \text{else.} \end{cases}$$



■ 网络威胁行为相似性度量

□ 序列距离的度量

□ 编辑距离 (Minimum Edit Distance, MED) 定义为： $MED(s_1, s_2) = \min\{len(OP \mid OP(s_1) = s_2)\}$

是指两个序列串之间，由一个转成另一个所需要的最少编辑操作次数。允许的编辑操作包括：将元素替换（substitution, s），元素插入（insert, i）或者元素删除（delete, d）

I	N	T	E	*	N	T	I	O	N
*	E	X	E	C	U	T	I	O	N
d	s	s		i	s				



■ 网络威胁行为相似性度量

- *MED*在序列长度差异明显时不能一致度量差异

s_1 : abcd

s_2 : abc**b**

s_3 : abcdefghijklmnopqrstuvwxyz**z**

s_4 : abcdefghijklmnopqrstuvwxyz**x**

$MED(s_1, s_2) = 1$  $MED(s_3, s_4) = 1$

s_3, s_4 比 s_1, s_2 之间的相似度应该更高?



■ 网络威胁行为相似性度量

□ 比例化编辑距离

- 比例化编辑距离可克服度量的不一致性

$$RMED(s_1, s_2) = \frac{MED(s_1, s_2)}{len(s_1) + len(s_2)}$$

是两个字符串编辑距离与同长度序列间最长编辑距离之比，也即是其编辑距离与两序列长度和之比。

- 比例化编辑距离消除了序列串长度异化带来的影响

$$MED(s_1, s_2) = 1$$

$$MED(s_3, s_4) = 1$$

$$RMED(s_1, s_2) = 0.125$$

$$RMED(s_3, s_4) = 0.0192$$



安全事件的聚类



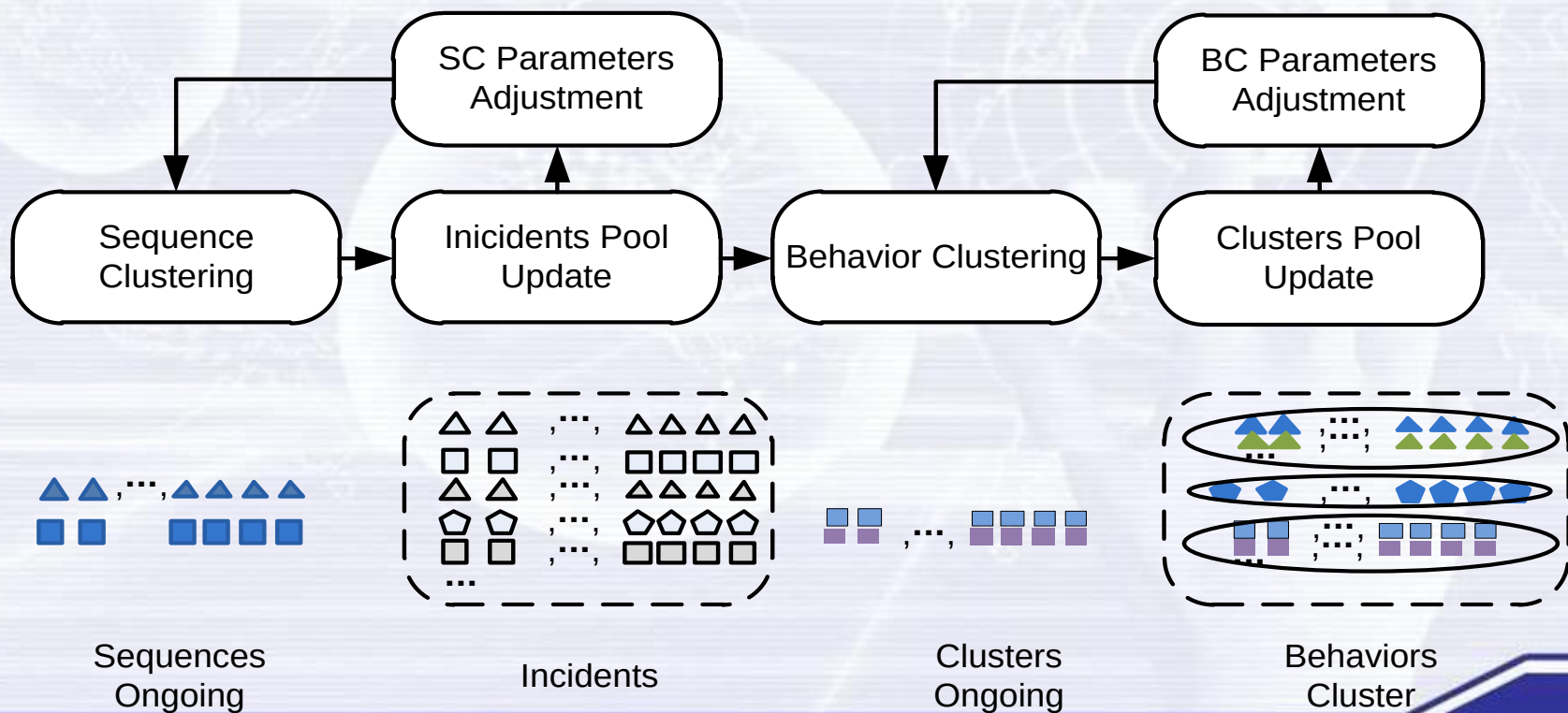
■ 聚类

- 聚类是一种非监督式学习，把相似的对象通过静态分类的方法分成不同的组别或者更多的子集，让在同一个组别或者子集中的成员对象拥有相似的属性
- 基于簇原型的聚类方法，创建数据对象的单层划分，簇是对象的集合，其中每个对象到定义该簇的原型比到其他簇的原型更加相似
- 基于簇原型的聚类具有较小的计算量，在数据量较大的情况下使用可以减少计算开销



■ 带参数调整的两个聚类框架

- 序列聚类：从事件还原可能的恶意序列
- 行为聚类：从序列到行为相似的行为簇





■ 网络事件聚类

□ 基本方法

- 通过网络事件时空相似性度量，计算网络事件时空相似度
- 相似度较高的分到同一类之中，相似度较低的分到不同类之中

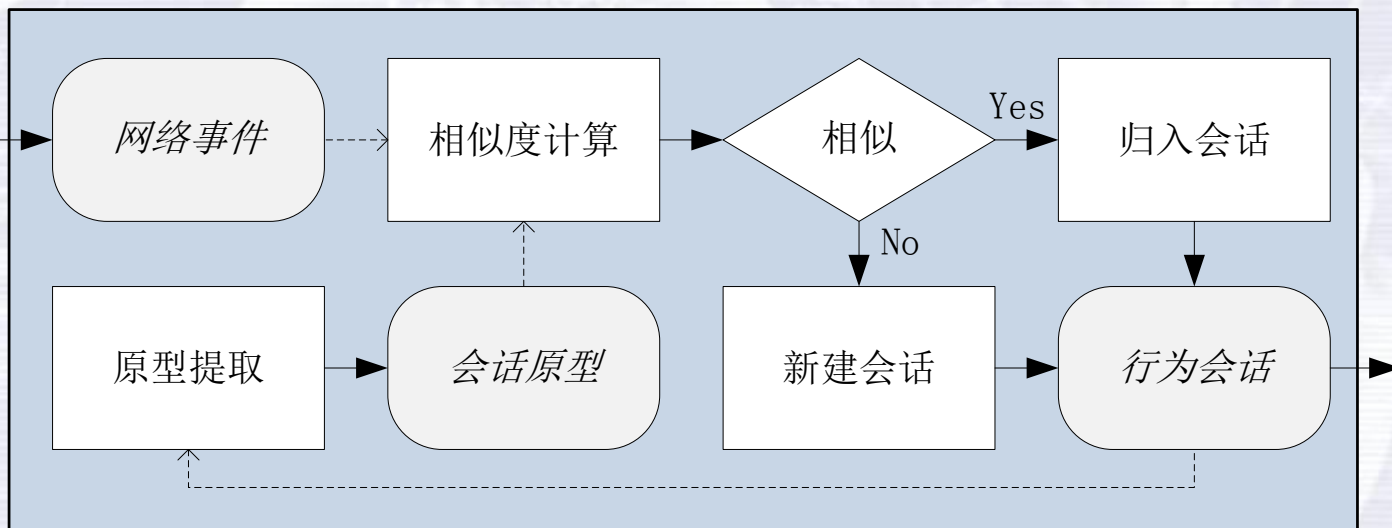
□ 改进的基于簇原型的聚类

- 构建簇原型并随着事件的到来而持续更新，这里需要持续性更新的是时间信息



■ 网络事件聚类

- 在网络事件聚类过程中，需要保持网络安全事件在时间特征上的先后顺序
- 对网络事件数据聚类的基本流程包括相似度计算、会话更新操作以及会话原型提取

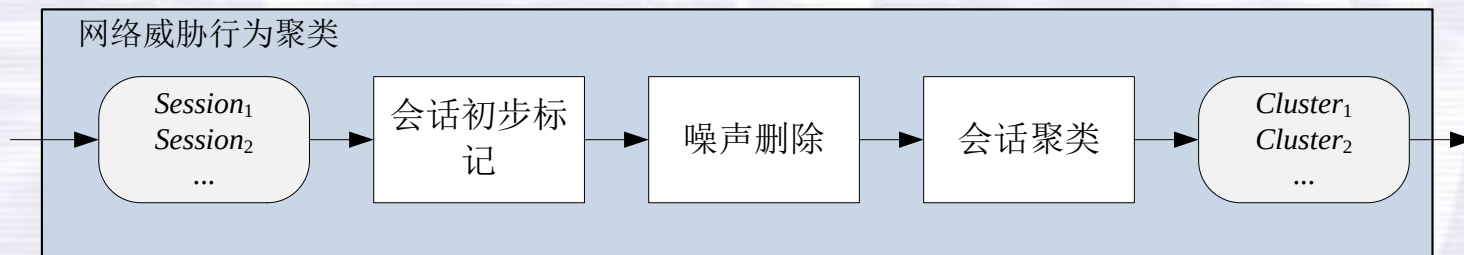


网络事件聚类流程图



■ 网络威胁行为聚类

- 对网络威胁行为行为的相似性分析选择比例化编辑距离 $RMED$ 度量，度量两个行为序列之间的步骤差异，全局归一化
- 在聚类过程称采用**DBSCAN**聚类算法，是一种基于密度的聚类算法，主要流程为会话警报分类、噪声去除和会话簇生成



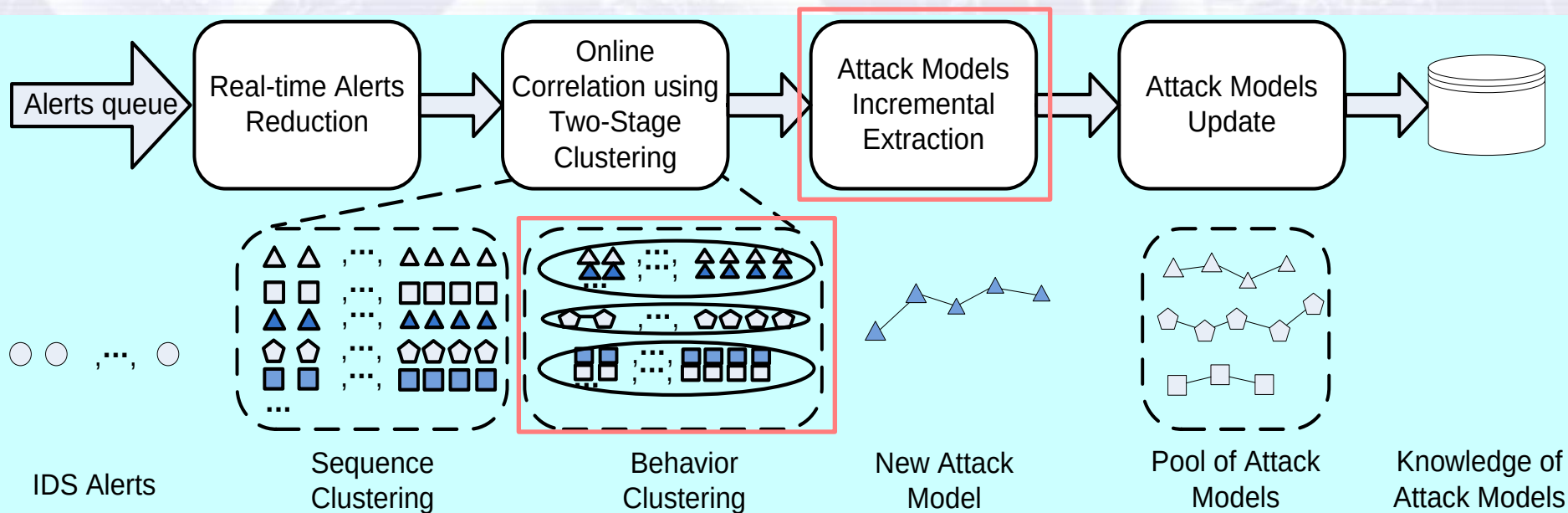
网络威胁行为聚类流程图



攻击模式提取



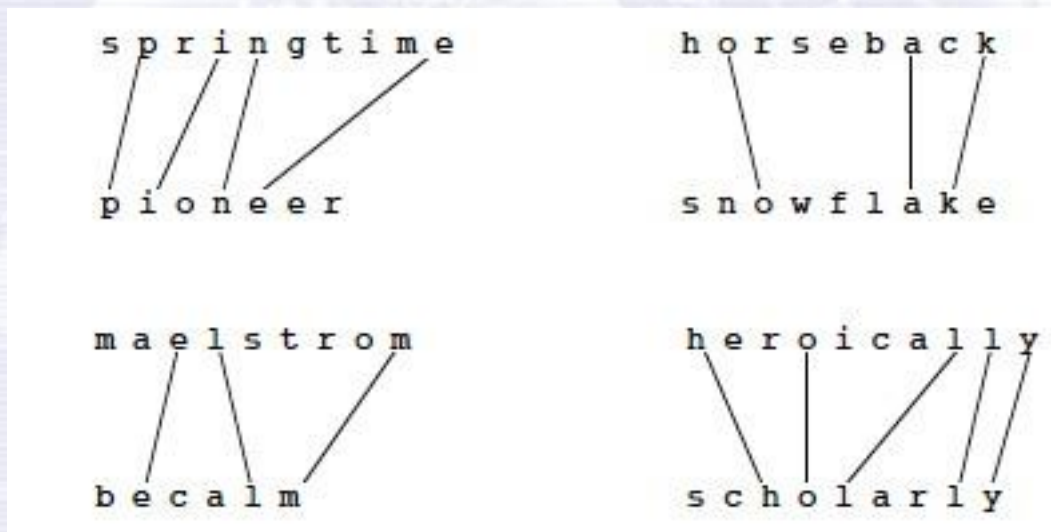
- 在分析框架中所处的位置
 - 是对类似序列的关联性和概括性分析
 - 分析攻击攻击工具惯用的手法
 - 攻击模式中通常蕴含了攻击意图信息





■ 最长公共子串(LCS)

- 最长公共子序列(Longest Common Subsequence, LCS)是计算机中的一个经典问题，寻找两个给定字符串中的长度最长的相同字符串子序列，LCS提取各序列的公共特征





■ 松散的最长公共子串(LLCS)

- LCS过强，但易造成步骤缺失
- 加入松散因子，攻击模式不需要完全覆盖每个序列，提高攻击模式挖掘能力

Input: Intrusion behaviors clusters C and threshold ε

Output: The models of the Intrusion behaviors S

- Initilize I
- Find Item In C with ratio over ε ;
- **Do**
- Add new subsequence into item-set I ;
- Generate new subsequences;
- Calculate the α of the new subsequences;
- **While** (exist ($\alpha > \varepsilon$))
- $S \leftarrow$ longest subsequence in I .

基于LLCS模式提取示意代码



□ 数据

- 真实骨干网13天数据
- 10E9+网络安全事件
- 10E7+主机地址

□ 实验结果

- 处理时间为4小时2分43秒
- 提取攻击模式14个

模式示例:

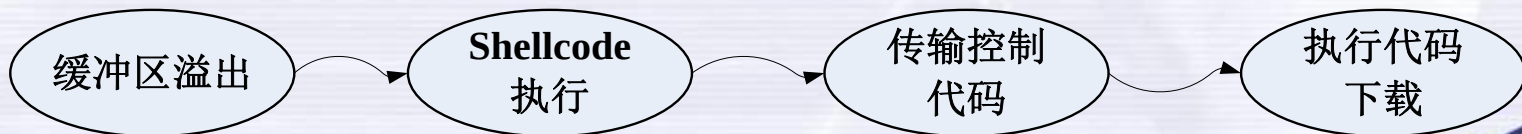
1-1394-12:SHELLCODE x86 inc ecx NOOP

129-16-1:FIN number greater than prior

1-648-10:SHELLCODE x86 NOOP

1-15306-4:WEB-CLIENT Portable Executable binary file transfer

1-16313-6:POLICY download of executable content - x-header





其他威胁模式

Attack Model 1:

1-687-8:SQL xp_cmdshell - program execution
1-686-13:SQL xp_reg* - registry access
1-683-7:SQL sp_password - password change
1-1387-11:SQL raiserror possible buffer overflow

Attack Model 2:

1-1419-12:SNMP trap udp
1-1413-13:SNMP private access udp
1-566-6:POLICY PCAnywhere server response
105-2-1:(spo bo) Back Orifice Client Traffic detected
1-634-5:SCAN_Amanda client-version request
1-1504-8:MISC AFS acces

Attack Model 3:

128-4-1:(spp_ssh) Protocol mismatch
119-21-1:(http_inspect) MULTIPLE CONTENT LENGTH
1-1939-5:MISC bootp hardware address length overflo
1-11976-4:VOIP-SIP overflow in URI type - SIP
1-1729-8:CHAT IRC channel join

Attack Model 4:

1-1394-12:SHELLCODE x86 inc ecx NOOP
129-12-1:Consecutive TCP small segments exceeding threshold
1-1990-2:CHAT MSN user search
1-540-12:CHAT MSN message
1-1986-7:CHAT MSN outbound file transfer request
1-1988-6:CHAT MSN outbound file transfer accept

Attack Model 5:

120-3-1:(http_inspect) NO CONTENT-LENGTH OR
TRANSFER-ENCODING IN HTTP RESPONSE
129-16-1:FIN number is greater than prior FIN
129-5-1:Bad segment, adjusted size <= 0

Attack Model 6:

1-11974-2:VOIP-SIP response too small
1-11969-2:VOIP-SIP inbound 401 unauthorized message
1-12007-2:VOIP-SIP outbound 401 Unauthorized message

Attack Model 7:

120-3-1:(http_inspect) NO CONTENT-LENGTH OR
TRANSFER-ENCODING IN HTTP RESPONSE
129-12-1:Consecutive TCP small segments exceeding threshold
129-5-1:Bad segment, adjusted size <= 0
129-7-1:Limit on number of overlapping TCP packets reached

Attack Model 8:

125-5-1:(ftp_telnet) FTP command parameters contained potential string format
1-491-10:FTP Bad login
1-2179-9:FTP PASS format string attempt
1-2417-5:FTP format string attempt

Attack Model 9:

120-3-1:(http_inspect) NO CONTENT-LENGTH OR
TRANSFER-ENCODING IN HTTP RESPONSE
125-8-1:(ftp_telnet) FTP bounce attempt
1-3441-4:FTP PORT bounce attempt

Attack Model 10:

120-3-1:(http_inspect) NO CONTENT-LENGTH OR
TRANSFER-ENCODING IN HTTP RESPONSE
129-5-1:Bad segment, adjusted size <= 0
1-17276-2:MISC Multiple vendor Antivirus magic byte detection evasion attempt

Attack Model 11:

1-1394-12:SHELLCODE x86 inc ecx NOOP
1-1390-8:SHELLCODE x86 inc ebx NOOP
133-34-1:(dcerpc2) Connection-oriented DCE/RPC - Request: Fragment length on non-last fragment (36) less than maximum negotiated fragment transmit size for client (4280)
1-2923-8:NETBIOS SMB repeated logon failure

Attack Model 12:

125-2-1:(ftp_telnet) Invalid FTP Command
125-1-1:(ftp_telnet) TELNET CMD on FTP Command Channel
125-9-1:(ftp_telnet) Evasive (incomplete) TELNET CMD on FTP Command Channel

Attack Model 13:

1-1394-12:SHELLCODE x86 inc ecx NOOP
129-16-1:FIN number is greater than prior FIN
1-648-10:SHELLCODE x86 NOOP
1-15306-4:WEB-CLIENT Portable Executable binary file transfer
1-16313-6:POLICY download of executable content - x-header

Attack Model 14:

129-12-1:Consecutive TCP small segments exceeding threshold
129-2-1:Data on SYN packet
124-3-1:(smtp) Attempted response buffer overflow: 596 chars
1-567-11:POLICY SMTP relaying denied



在线优化



■ 基于多核处理器的软件流水线

□ 采用软件流水线框架，提高并行性

- 每个处理器线程专职一个步骤（Task）
- 每个Task都有一个输入队列和输出队列
- 处理完当前task处理队列中的任务放入输入队列，然后进行下一个task

Input (1)	Input (2)	Input (3)	Input (4)	Input (5)		
	RMSR (1)	RMSR (2)	RMSR (3)	RMSR (4)	RMSR (5)	
		SC (1)	SC (2)	SC (3)	SC (4)	SC (5)
			BC (1)	BC (2)	BC (3)	BC (4) BC (5)
				LLCS (1)	LLCS (2)	LLCS (3) LLCS (4) LLCS (5)

软件流水线框架



- **懒惰的数据一致性策略，减少时空开销**
 - 事件时空聚类中，缓冲池包含大量状态数据，实时更新将带来给系统运行带来较大的时空开销，采用懒惰的数据一致性策略，减少系统运行的时空开销
- **并行化行为聚类算法，开发多核潜能**
 - 威胁行为聚类的开销远超过其余部分，存在较大的提升空间
 - 并行化威胁行为聚类算法充分利用多核平台性能，平衡软件流水线个阶段时空开销



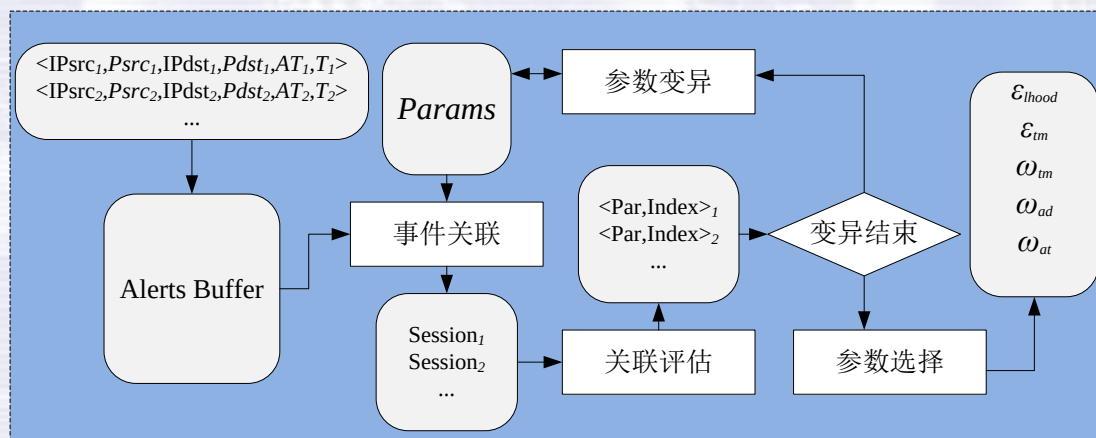
■ 并行化行为聚类算法

- 并行化威胁行为聚类将行为聚类过程中的相似度计算作任务划分之后，分配给各个处理器线程线程
- 从各个线程的计算结果中构造出威胁行为间的全局相似性度量
- 并行化聚类需要平衡开销，当并行化带来的优势不足以弥补并行化的额外开销时只进行单线程处理



■ 在线参数调整

- 威胁模式结果挖掘的质量和挖掘过程的时空开销与挖掘过程中使用的参数紧密相关
- 挖掘参数应与网络环境相对应
 - 对不同的网络环境制定不同的参数
 - 对变换的网络环境保持参数和环境的一致性



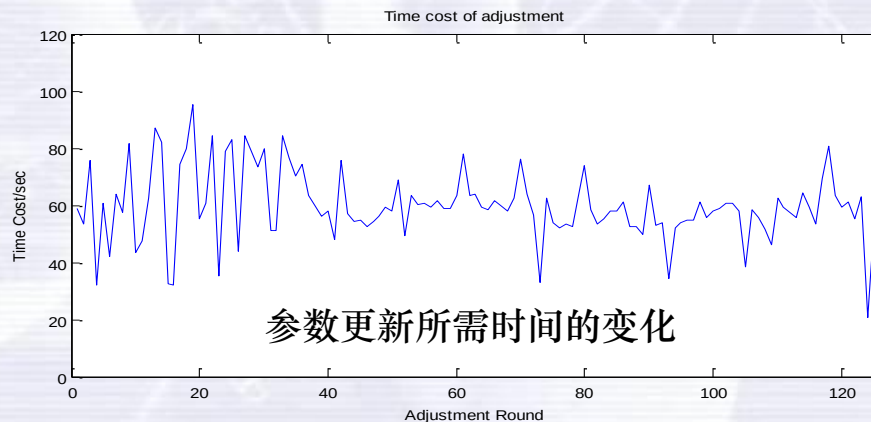
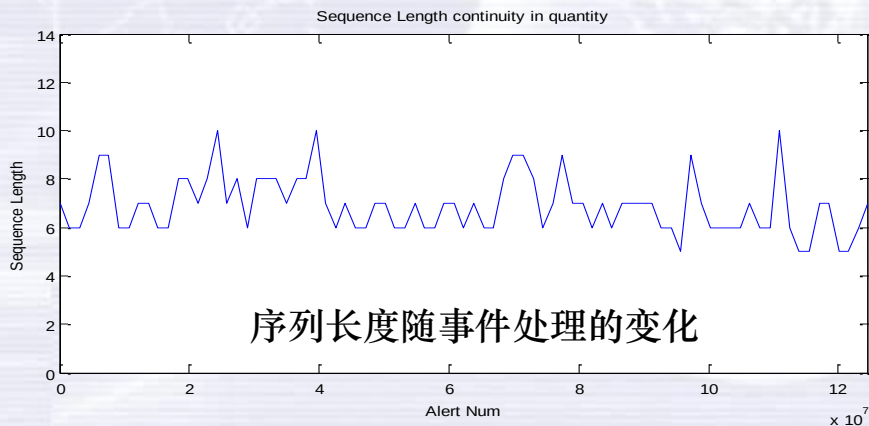
网络事件聚类参数在线调整流程图



参数优化结果

对真实骨干网上网络安全事件进行威胁模式提取

- 序列长度随时间的变化较为平稳
- 每次参数更新的时间开销基本稳定保持在40~90s





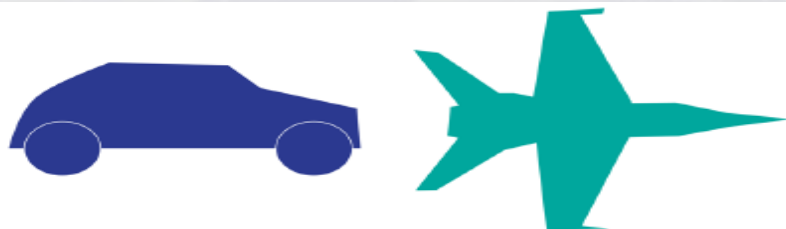
大数据带来的变革



■ 威胁模式挖掘面临着新的挑战



Volume: 数据量



Velocity: 数据产生速度



Variety: 结构和非结构化数据



- **传统方法在入侵检测和取证中分析日志、网络报文和系统事件时面临严峻挑战**
 - 大量数据难以存储保留，出于经济考虑一般60天后删除
 - 在大规模结构化数据中进行分析、复杂查询难以做到高效
 - 传统技术不适合于管理、分析非结构化数据
 - 计算可靠性难以得到保证



■ 数据驱动的信息安全

- 海量数据的分析处理是复杂网络威胁检测的一大挑战，而大数据相关技术能为此提供有力支撑
- 在2013年RSA大会上，ZB公司宣称相比于传统的SIEM工具，使用Hadoop集群和BI工具可以更快的解析更多的数据
- 在他们的研究中，使用传统系统对1个月的数据进行搜索需要20~60分钟，而使用大数据平台Hadoop仅需1分钟左右，性能对比显著



- Hadoop生态系统和流处理相关的数据库等大数据新技术能够以空前的速度和规模存储和分析大规模异构数据
- 这些技术将为安全分析带来革新：
 - 在超大尺度上收集各方数据
 - 对数据进行更加深入的分析
 - 提供全局统一的安全视图
 - 对流数据接近实时化的分析
- 但是最重要的问题
 - 保护的对象和成本



■ 谢谢！