东南大學 SOUTHEAST UNIVERSITY
网络空间安全学院
School of Cyber Science and Engineering

InForsec“走进蚂蚁集团”

网络空间安全2020大学生夏令营

加密网络流量测量和分析

东南大学 程 光

2020 年 7 月18日

◆提 纲

- 学院情况介绍
- 相关研究情况介绍
- 实验室师资和环境介绍
- 加密网络流量测量和分析
- 小结

学院情况介绍

School Introduction

◆一流网络安全学院示范项目

- 《关于加强网络安全学科建设和人才培养的意见》（中网办发文[2016]4号）、《一流网络安全学院建设示范项目管理办法》（中网办密字[2017]573号）的要求，创新网络安全人才培养机制，争创一流网络安全学院。
- 2017年9月16日，中央网信办、教育部公布了7所“一流网络安全学院建设示范项目高校”名单。2019年8月公布了第二批四所学校名单。

序号	学校
1	西安电子科技大学
2	**东南大学**
3	武汉大学
4	北京航空航天大学
5	四川大学
6	中国科学技术大学
7	战略支援部队信息工程大学
8	华中科技大学
9	上海交通大学
10	北京邮电大学
11	山东大学

◆学院历程

1978 顾冠群院士创建网络学科

2016.1 获批网络空间安全博士点

2016.12 成立网络空间安全学院

2017.3 任命于全院士担任学院院长

2017.9.16 获批一流网络安全学院建设示范项目高校

2018.3.5 获批网络空间安全本科专业

2018.7.5 获批网络空间国际治理研究基地

2019.12 获批网络空间安全博士后流动站

2019.12 首批教育部一流本科专业建设“双万计划”

实验室研究

laboratory Research

◆实验室研究

- ✓ 网络测量关键技术
- ✓ 下一代网络恶意流量检测
- ✓ 面向5G的流量测量技术
- ✓ 网络攻击主动防御技术
- ✓ 加密流量系统化智能感知与分析关键技术
- ✓ 暗网关键技术
- ✓ 物联网威胁感知与协同治理研究
- ✓ 高速网络基础设施的流量大数据分析

01

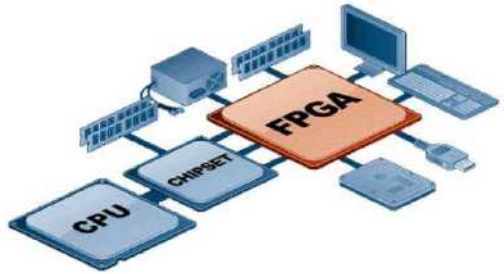
网络测量关键技术

◆ 面向自适应网络测量架构及威胁态势感知的基础设施关键技术

创新点：

（1）硬件创新点

基于Intel Atom CPU和Altera arria V FPGA实现可编程交换设备OpenBox，是一个基于多核CPU 和FPGA 的开源平台，支持下一代网络体系结构、新型网络协议和分组处理机制研究的开放可重构交换平台。



（2）FAST框架创新点

FAST

FAST是一个可扩展的框架，实现了基本的网络交换功能，例如L2 交换，L3 路由和openflow 交换等。用户可以在FAST架构基础上实现自己的创新工作，而不用重新设计一个完整的数据平面。

（3）基于OpenBox硬件和FAST框架开发了DDoS攻击检测平台

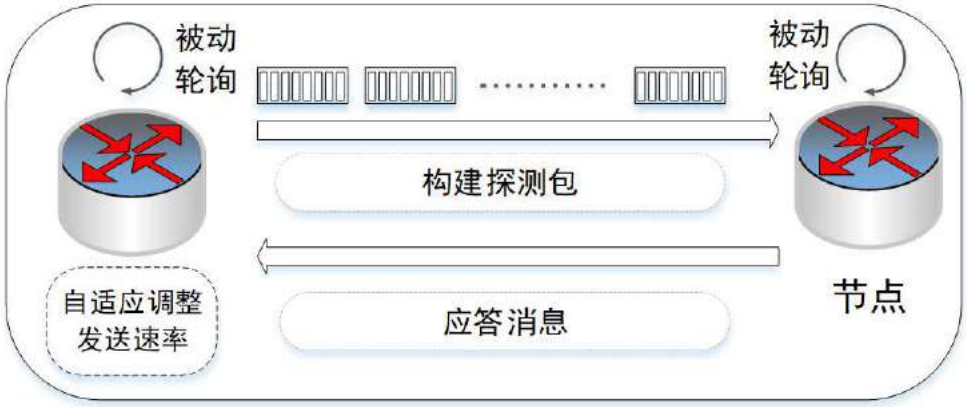
DDoS攻击检测应用支持IPv6网络，使用OpenBox硬件计数器提前感知网络状态，使用机器学习技术识别DDoS攻击，使用数据库存储异常流量特征向量，并且基于Web技术实现了检测平台的控制和展示系统。

◆ 网络自适应弹性测量方法

项目工程性创新点：

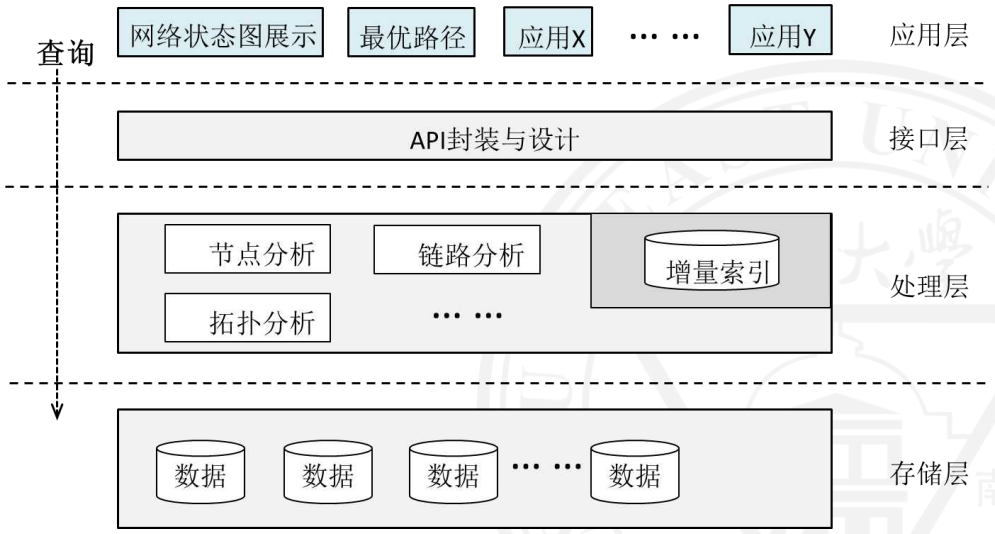
（1）自适应测量

采用自适应测量调度控制器，根据测量任务和当前全网状况调整测量粒度。



（2）网络动态图谱

动态网络图谱主要为网络中多维资源视图而设计分布式图数据库，除了在内存数据库(比如Redis)上封装一层逻辑的图结构，还增加网络拓扑状态增量索引，保证版本控制。

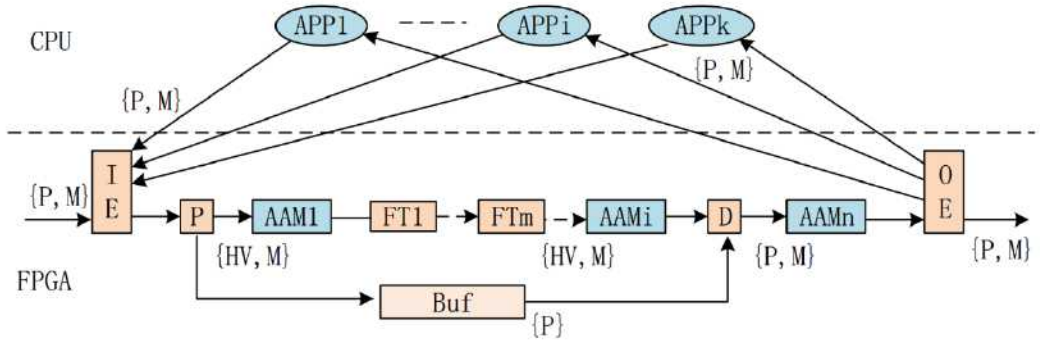


◆ 下一代网络处理器体系结构及关键技术研究

创新点：

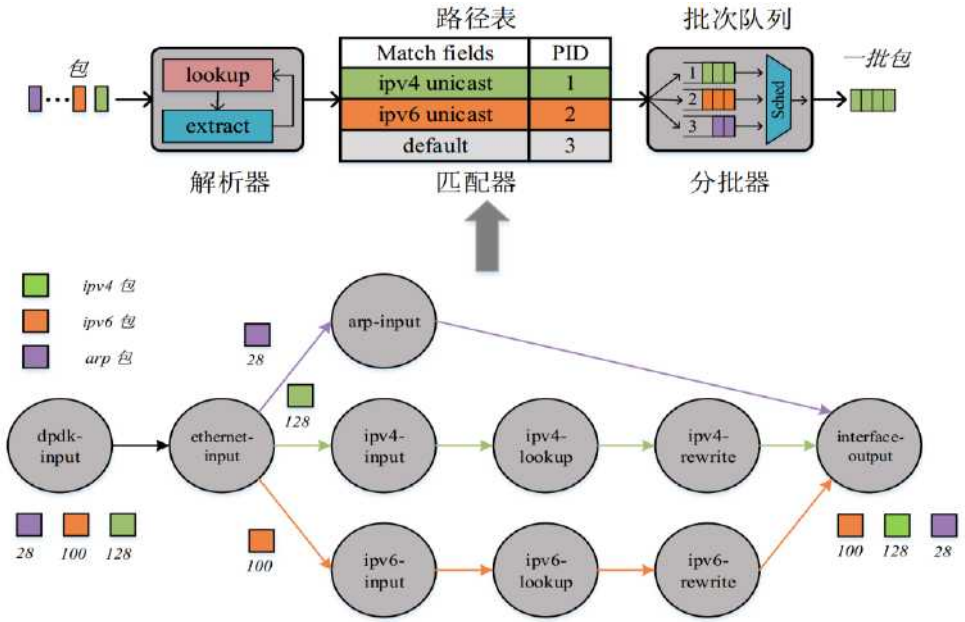
（1）提出软硬件协同可扩展的下一代网络处理器架构

项目提出基于通用多核 CPU 加 FPGA 构建下一代网络处理器，设计软硬件协同的可扩展分组处理架构



（2）提出通用多核处理器感知的分组优化调度技术

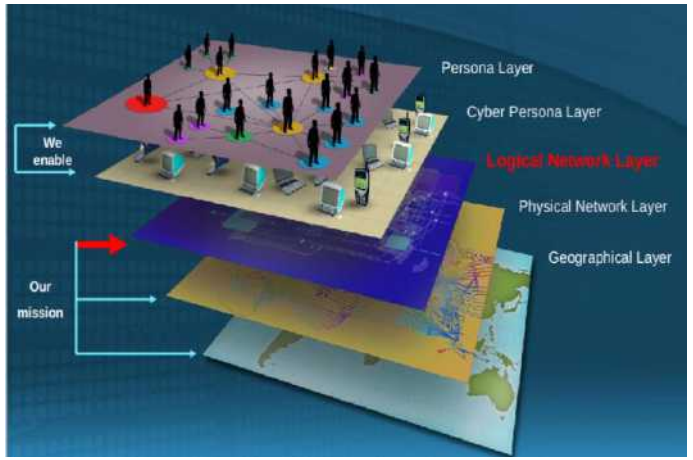
提出一种通用多核处理器感知的分组调度技术，通过面向访问局部性优化的分组优化调度机制，有效降低通用多核处理器 TLB 和 Cache 的失效率。



◆ 网络空间作战力量全球布势及感知管道体系化建设研究

研究问题：

本研究关注的是网络空间作战力量全球布势及感知管道体系化建设，将系统化分析美及其盟友在全球范围部署的“藏宝图”、“湍流”、“精灵”等体系化侦察感知和网络攻击项目。



藏宝图

湍流

(U) COMPUTER NETWORK OPERATIONS
(U) GENIE

This Exhibit is SECRET//NOFORN

	FY 2011[1] Actual	FY 2012 Enacted			FY 2013 Request			FY 2012 – FY 2013	
		Base	OCO	Total	Base	OCO	Total	Change	% Change
Funding ($M)	615.2	562.0	74.2	636.2	589.2	62.5	651.7	15.6	2
Civilian FTE	1,110	1,238	–	1,238	1,231	–	1,231	-7	-1
Civilian Positions	1,110	1,238	–	1,238	1,231	–	1,231	-7	-1
Military Positions	456	572	–	572	639	–	639	67	12

[1]Includes enacted OCO funding. Totals may not add due to rounding.

精灵

02

下一代网络恶意流量检测

实验室研究

◆ 下一代网络拒绝服务攻击检测和防护关键技术研究

研究内容：

① 研究拒绝服务攻击流量实验环境；

② 研究拒绝服务攻击的智能检测框架和核心算法；

③ 研究基于SDN的全网网络设备和清洗设备联动防护处置方案；

④ 基于网络切片和威胁情报共享的下一代网络的DDoS检测与防御框架。

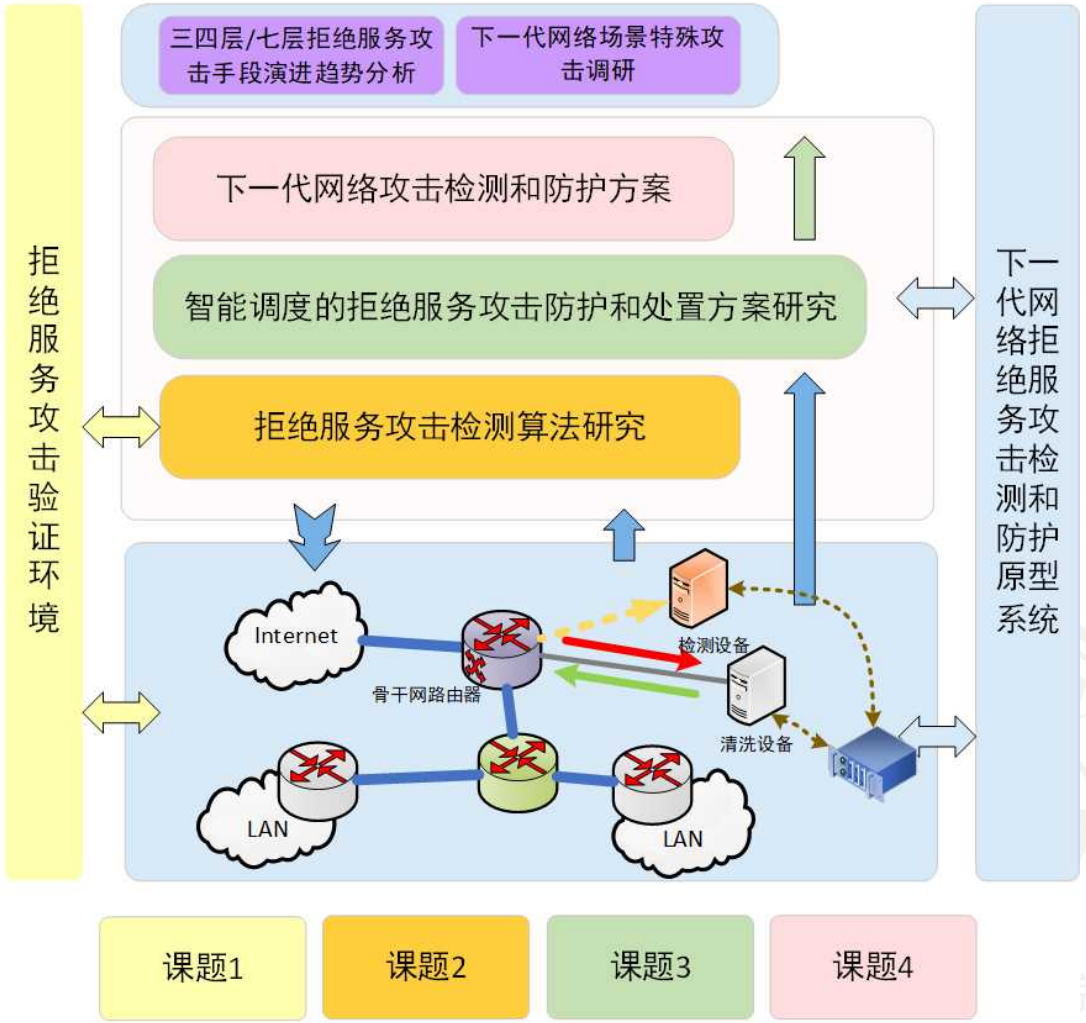
创新点：

① 拒绝服务攻击流量生成框架及其实现技术；

② 基于多源异构大数据的DDoS智能检测方法；

③ 基于SDN的云化流量清洗和多点协同DDoS防御框架；

④ 基于多层动态栈式处理的下一代网络智能DDoS检测与防御框架。

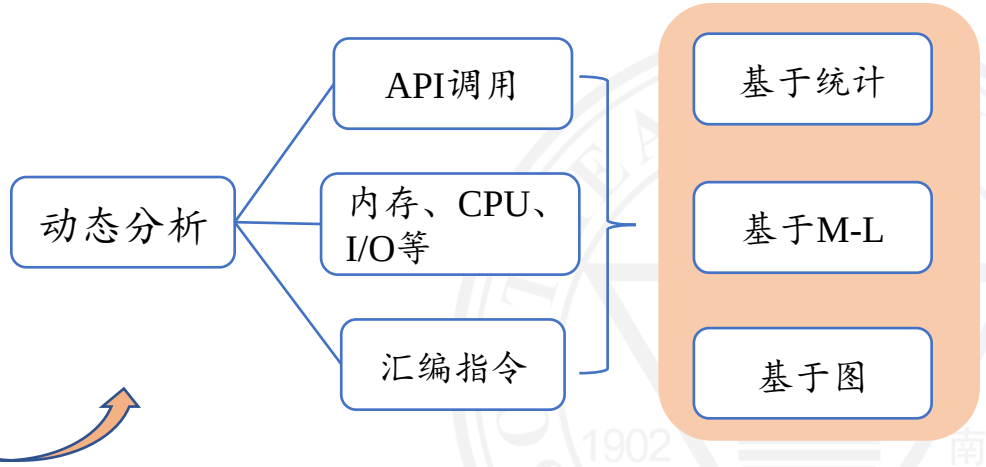


项目系统架构图

基于动态分析的勒索软件检测方法

- CNCERT发布的《2019年我国互联网网络安全态势综述》
  - 勒索软件攻击持续活跃；
  - 在一年之内样本就捕获了73.1万余个，较2018年增长超过4倍；
  - 勒索软件攻击活动越发具有目标性，并形成互联网地下黑灰产；
  - 如何准确的检测勒索软件和有效的恢复被加密文件是目前研究的热点问题。

- 勒索软件检测方法目前主要分为两类：
  - 静态分析技术：不执行代码；无法适用于混淆、变形等技术。
  - 动态分析技术：执行软件并分析其行为；软件行为模型；可以弥补静态分析技术中的缺点。



03

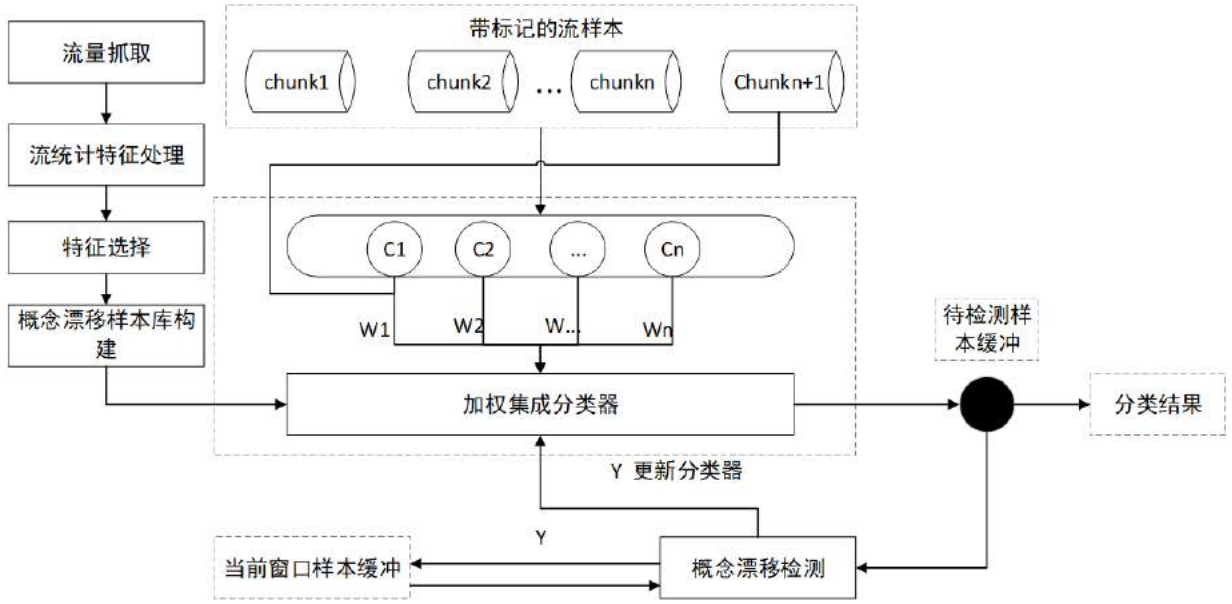
面向5G的流量测量技术

实验室研究

网络流概念漂移检测与分类方法研究

创新点：

提出一种基于散度的网络流概念漂移分类方法，采用双层窗口机制，从信息熵的角度出发，根据流量特征分布的Jensen-Shannon散度来度量滑动窗口内数据分布的差异，从而检测概念漂移。采用增量集成学习的策略在概念漂移点引入漂移流量建立分类器，通过分类器权值比较替换掉性能下降的分类器，加权集成分类结果对样本进行分类。



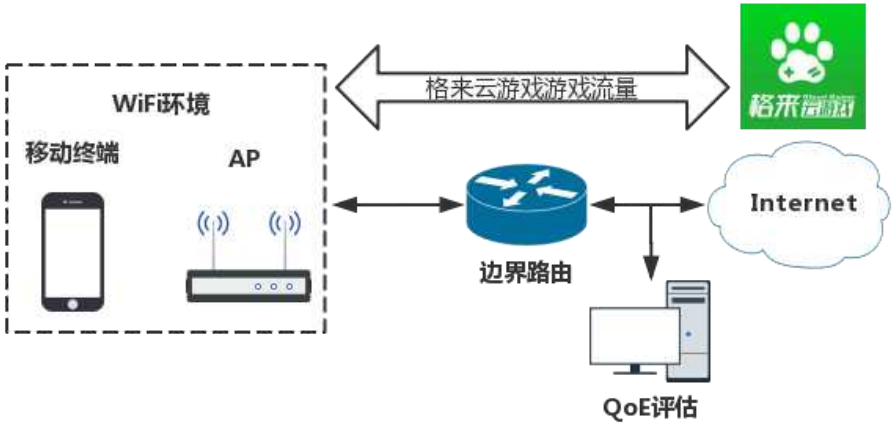
基于散度的网络流概念漂移自适应分类方法

AR/VR识别算法研究

研究内容：

① 移动云游戏QoE评估

② 爱奇艺4K视频识别与QoE评估



研究目的：

5G下，移动云游戏、VR/AR、4K高清视频等数据流量将占据互联网主要的流量，对其识别与QoE评估，可以给服务提供商动态调整网络服务参数、优化网络资源、提高网络利用率，提供参考。

创新点：

① 提出了一种移动云游戏QoE评估模型，使用FAHP的方法对移动云游戏QoE影响因素进行建模。

② 构建移动云游戏数据库，通过模拟不同的网络环境，在不同的设置条件下，针对移动云游戏平台“格来云游戏”进行测量，采集数据流量。

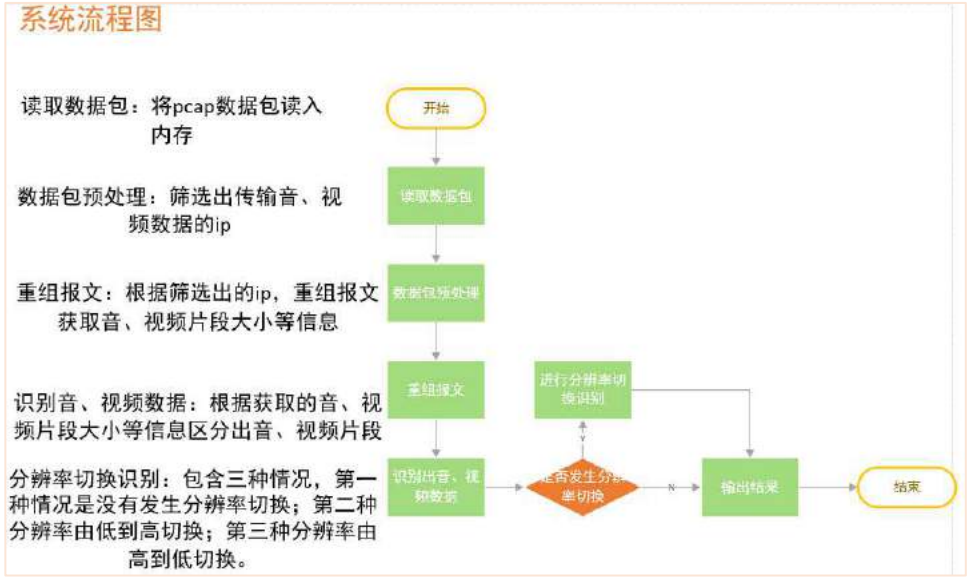
加密流媒体分辨率切换识别

研究内容：

① 调研加密流媒体的传输机制和特征；

② 完成加密流媒体视频数据的重组；

③ 完成视频分段分辨率的识别和视频分辨率切换识别。



研究目的：

由于采用了加密协议传输流媒体数据，数据内容不可见，网络服务提供商需要在不侵犯用户隐私的情况下了解用户的习惯和行为以提高网络服务，因此需要找到一种识别加密流媒体分辨率切换的方法。

创新点：

① 基于ACK number的方法，从加密网络数据中重组出音频和视频分段，再通过特征提取和视频分段构建样本库，采用机器学习的方法识别视频分段的分辨率，最后通过滑动窗口判断分辨率切换；

② 项目三期重新分析了数据的传输机制和特征，结合请求间隔和其他特征重组出了音视频数据。

◆ 4K高清视频相关（VR类）

研究内容：

研究VR高清视频的识别和体验评估。包括业务识别算法设计及源码、QoE建模分析报告、VR主辅视野的识别。调研后发现Visbit 8K应用中符合研究要求。

研究目的：

识别Visbit 8K应用、识别场景中的主辅视野、源码设计、QoE建模设计与分析；研究VR应用的流量特征，构建QoE用户体验评估模型。

创新点：

VR业务中的主辅视野机制（FOV）对于运营商和用户来说是一个新的研究领域。经过测算，普通4K视频码率在0.45mbps左右，而无主辅视野机制的4K VR视频的码率在1.2~1.5mbps左右。主辅视野机制允许在VR的主视野区域传输高清视频，而在非主视野区域（辅视野及背后的视野）传输普清视频，能在几乎不影响用户体验的前提下，大大节约网络带宽资源。

04

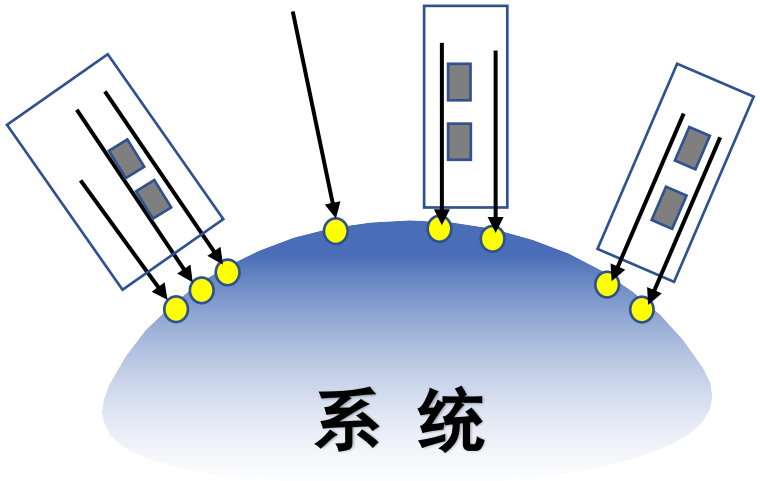
网络攻击主动防御技术

实验室研究

动态攻击防御技术

研究意义：

网络安全防护能力的对于保障网络信息体系效能具有至关重要的作用，动态目标防御通过多样的、不断变化的构建、评价和部署策略来增加攻击者的攻击难度及代价，有效限制脆弱性暴露并增强系统弹性。



静态防御技术

确定性、静态性、相似性

攻防双方时间、信息的不对称性

网络安全“易攻难守”的格局

网络空间欺骗防御技术

研究意义：

形成全方位、多层次的欺骗防御效能监控与评估体系，弥补现有欺骗防御体系中隐蔽性量化的缺失，能够根据采取的欺骗防御方案，定量评估当前欺骗防御场景下的多层次系统资产的隐蔽性与欺骗防御效能。

研究目标：

研究提出欺骗防御效能的监控评估方法，在网络数据、系统数据、诱饵使用事件高效采集的基础上，突破基于贝叶斯网络的欺骗防御隐蔽性量化、基于攻击图变换的欺骗防御有效性评估等关键技术，实现对欺骗防御的有效性和隐蔽性进行评估。

系统结构：

网络空间欺骗防御系统主要由控制子系统、隐藏混淆子系统、诱捕子系统、监控评估子系统组成。建立监控评估子系统，在信息收集的基础上对欺骗防御的效果进行评估。

05

加密流量系统化智能感知与分析关键技术

实验室研究

◆ 加密视频流量识别

研究意义：

近年来，视频流量在互联网流量中的占比的逐年攀升，与此同时，为了保护用户隐私，加密技术被广泛应用。如何从加密的视频数据中抽取出网络安全防护和网络管理需要的信息已经成为网络管理中亟待解决的问题。在做到保护用户的隐私的同时，及时发现因特网中传递的危害国家和社会安全的信息，是目前加密视频流量研究中的一大挑战。

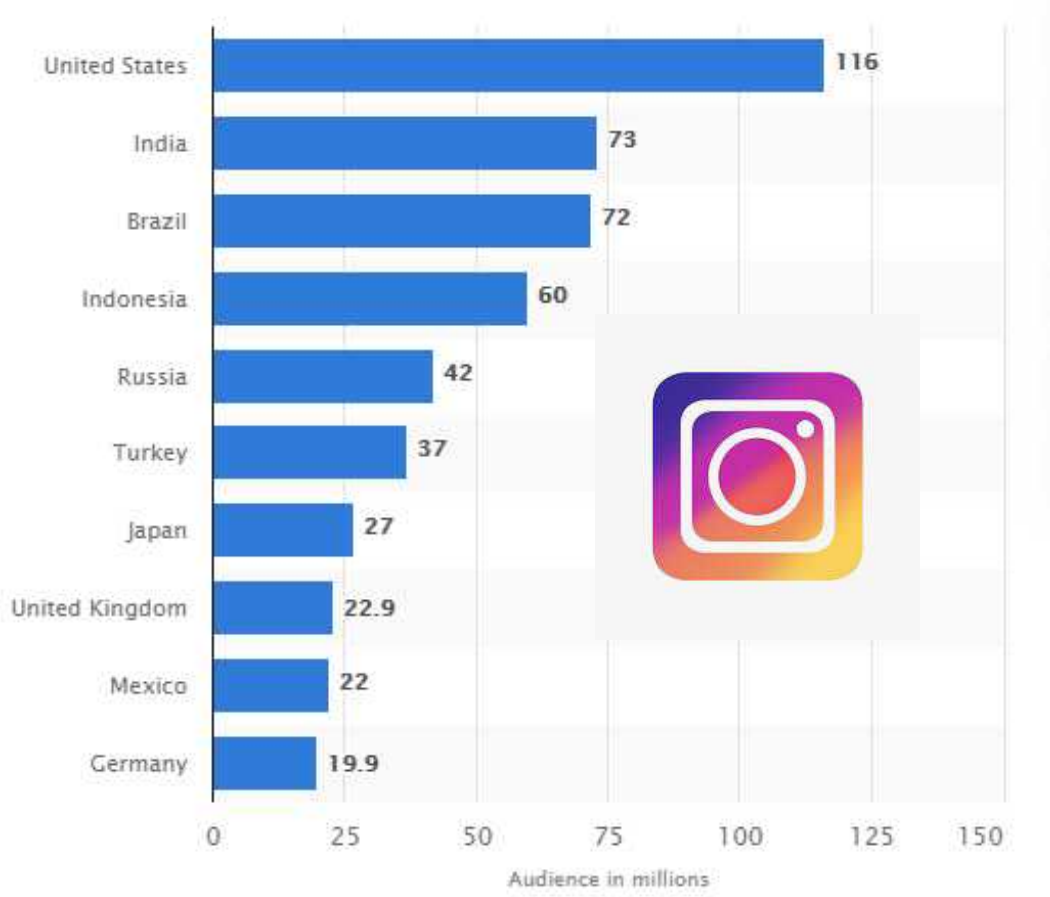
研究思路：

现有的针对加密视频流量的研究普遍采用流特征，但是方法的效果差强人意。为了提高加密视频流量识别的准确率与可扩展性，本方法首次提出将HTTP头部长度和TLS片段相关信息作为主要特征，并提出使用视频片段的长度作为构建视频指纹的关键信息。

◆ 加密流量用户行为识别（Instagram）

研究现状：

在加密流量用户行为识别的研究问题中，基于端口的检测和基于深度包检测（数据流特征分类技术） 不再适用，这些方法中分类算法多采用K近邻、决策树、朴素贝叶斯等，特征选取往往基于数据包大小分布、往返时间、时间差等。这些方法的缺点是数据丢包、重传影响大，适用性低。



截止2019年10月的Instagram用户数量排名领先的国家

06

暗网关键技术

实验室研究

暗网关键技术研究

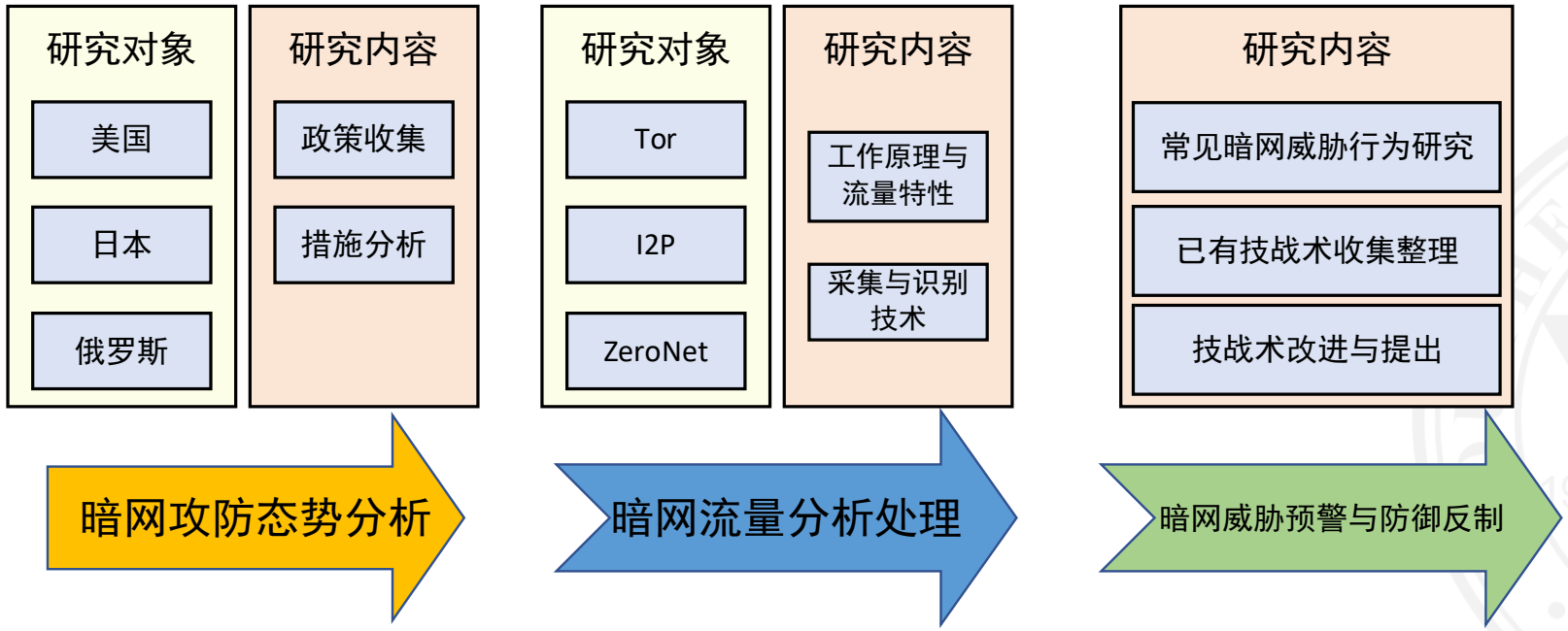
研究体系结构：

针对暗网环境的复杂性、隐蔽性，研究：

① 全球主要国家的暗网攻防态势情况，各国针对暗网提出的相关政策等；

② 暗网流量采集、处理、识别技术，实现暗网流量的精准识别；

③ 暗网威胁预警与防御反制技术，对于暗网威胁及时预警并处理。



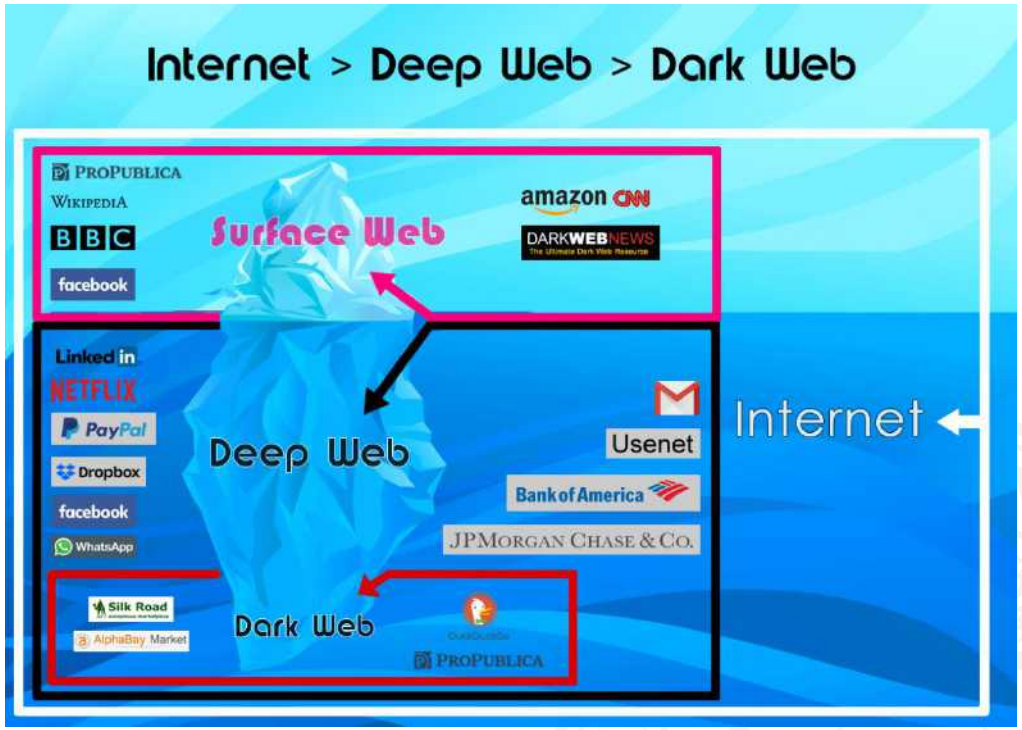
◆ 暗网关键技术研究

创新点：

① 对全球主要国家的暗网攻防态势相关政策与措施进行了收集与整理，并通过总结其发展历程，为暗网攻防的发展方向提出建议。

② 对常见暗网技术中的加密流量进行研究。在收集、整理、研究其工作原理的基础上，实现对暗网加密流量的精准识别。

③ 提出新的暗网威胁预警与防御反制技战术。在识别暗网加密流量的基础上，实现对暗网威胁行为的检测与预警，并进行及时防御反制。



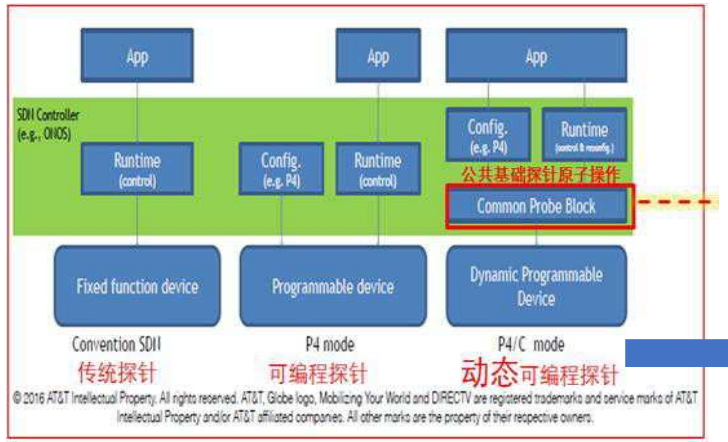
08

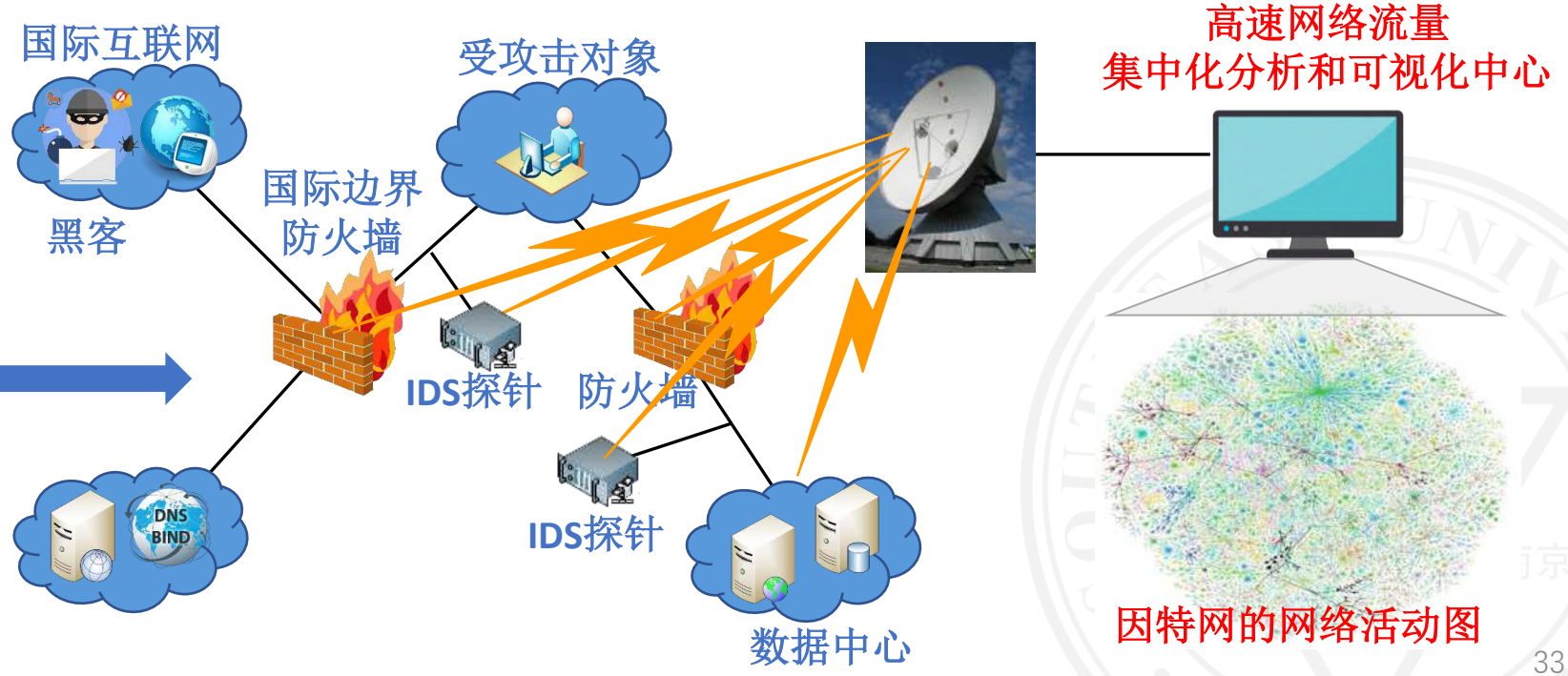
高速网络基础设施的流量大数据分析

高速网络基础设施的流量大数据分析

研究问题介绍：

在网络传输层，通信网络基础设施（如骨干网、数据中心网、5G核心网）承载了端设备间的数据通信。网络中存在大量未知、隐蔽、碎片化攻击流量穿越机构的网络边界，其来源不明、意图不明，需要在边界网关和内部主机部署泛在探针，分析通信网络中的巨量高速流量，检测未知网络攻击。





A-CORD（AT&T eComp测量子项目）：软件定义的动态可编程测量架构:

(1) 公共的基础探针原子操作(primitives vProbe)：如base DPI，主动拨测，QOS监测， SLA验证

(2) 基础探针原子可下发到在ebpf框架中，进行动态组合与编排，满足各种网络监测场景测量需求。

高速网络基础设施的流量大数据分析

课题：跨时空域流式大数据融合分析与复杂网络行为检测

理论研究内容

研究内容一	研究内容二	研究内容三
亚线性存储消耗的流式数据分析，用于实现泛在探针网	亚线性时间消耗的动态图数据分析，用于集中计算大图	基于深度学习，实现复杂网络行为的准确检测、关联、预测

应用研究内容

研究内容四

- 针对国家网络基础设施的智能运维需要，部署泛在探针，构建全局视图，支持意图驱动调度
- 针对国家基础设施安全的需要，智能检测已知/未知网络异常，实现攻击溯源，识别攻击意图

- 创新点1：完善跨时空域的流式数据分析理论，实现泛在探针的高效协同，准确查询抽取复杂全局状态
- 创新点2：针对带时空间属性的大图数据，形成网络对象画像，归纳事件时空、内容和主客体关联性，推理事件潜在关联和演化趋势，识别社区化长期复杂行为
- 创新点3：推测网络行为主体的意图，推测其主体的身份、组织、国家等属性，理论联系实际，开发原型系统

关键技术

硬件加速 (Offloading)

高精测量 DPM
- 抽样技术 (Sampling)
- 数据流技术 (Data Streaming Algorithm)

语义识别 SPR
- 分类技术 (Classification)
- 深度报文、流、语义检测技术 (DPI、DFI、DSI)
- 通信模式识别 (Anomaly Detection, Similarity Calculation)

网络智能感知实验室概况

Network intelligent awareness LaB

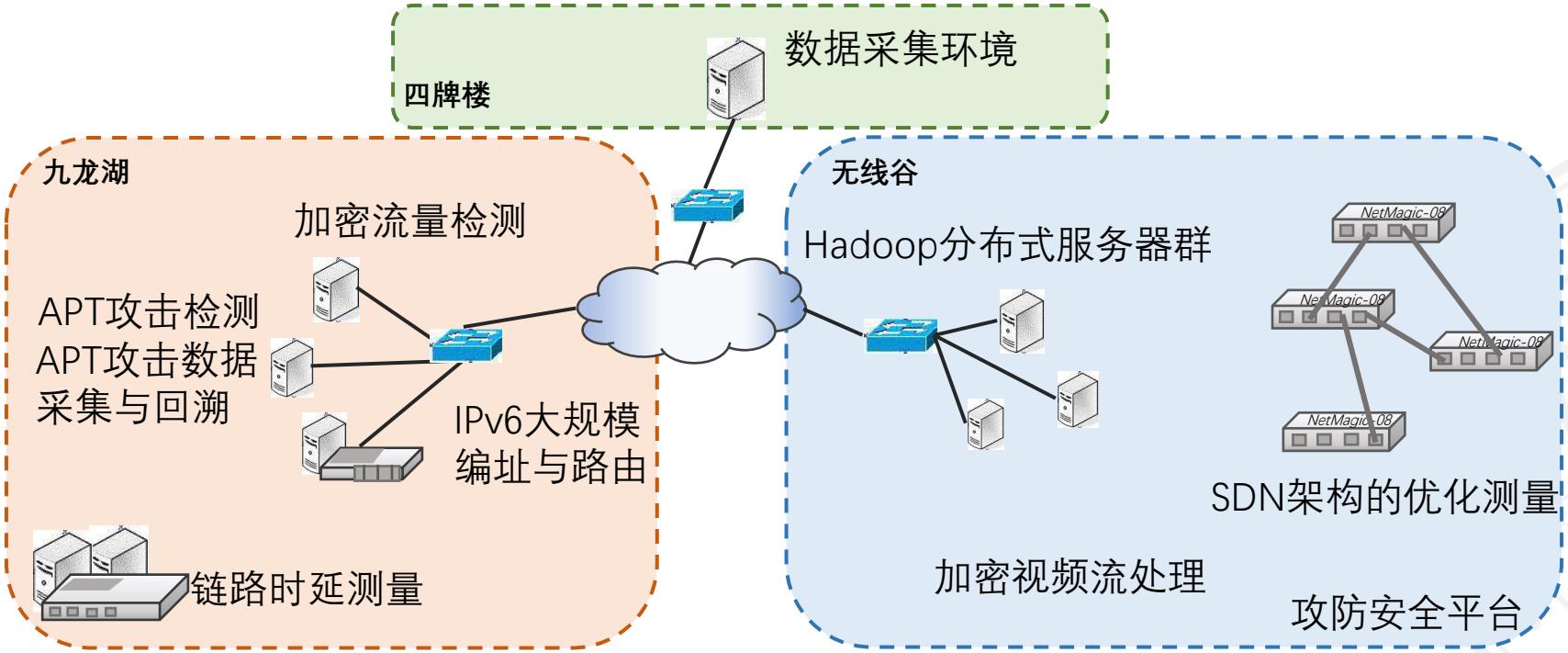
依托平台

- 计算机网络和信息集成教育部重点实验室（东南大学）
- 江苏省计算机网络技术重点实验室
- 中国教育和科研计算机网CERNET华东（北）地区网络中心
- 江苏省高速网络管理与服务品质监测工程技术研究中心

实验环境

包括iRouter，NetMagic-08，服务器，路由器，刀片服务器等。

- 基于NetMagic-08搭建的SDN架构的优化测量实验环境
- APT攻击数据采集及回溯实验环境
- 自适应流媒体视频传输测量实验环境
- 僵尸网络检测实验环境
- 基于iRouter的IPv6大规模编址与路由实验环境，及链路时延测量环境



构建的实验系统

测量架构：

- 基于可编程交换机的软件定义网络测量系统架构
- 面向层叠网的网络测量系统

视频测量：

- 加密流量视频码率&分辨率识别系统

精确测量：

- 基于精确测量的TCP拥塞控制机制
- 基于链路精确反馈的拥塞控制机制

安全检测：

- 基于海量大数据APT攻击检测实验方案

加密网络流量测量和分析

◆ 加密流量现状

- Netmarketshare，2019年10月，全球使用HTTPS加密的Web流量的比例超过九成。
- Google，2019年5月，94%web流量都被加密.
- 在全球范围内，美国的HTTPS比例为92%，俄罗斯为85%，日本为80%，印尼为74%。
- 互联网上排名前100位的非Google网站中默认使用HTTPS的有90位。
- 百度一直强制网站更换HTTPS，百度使用HTTPS的网站占比预计大于60%。
- Barac的创始人Omar Yaacoubi指出所有流量的80%将在2019年加密。

◆ 加密流量系统化智能感知与分析关键技术研究

研究意义：

形成加密流量已经成为当今网络流量的主体形式，大规模加密流量在给用户带来安全与隐私保护的同时，也给网络管理与安全监管带来了挑战。目前，加密流量感知与分析已经成为国家在网络管理与安全相关领域的一大瓶颈，在国际上也是尚未解决的一大难题。

对加密流量的检测、分类与识别，可以有效的打破当前网络流量管理与网络空间安全领域现存的壁垒，大幅度提高我国网络管理运行效率及网络安全管理能力，进而提高网络经济运行效率与网络安全防护能力，保障国家网络环境的良好发展与长治久安。

加密流量感知与分析三部曲

国内外加密网络流量的研究

- **国外研究机构**：密歇根州立大学、加州大学伯克利分校、Cisco 公司、英国牛津大学等
- **国内研究机构**：东南大学、中科院信工所、清华、北邮、西安交大等
- **国外厂家**：Cisco、ENEA、Vectra、Awake Security、Bricata、 Corelight、Corvil、Darktrace、ExtraHop、FireEye、GREYCORTEX、 Lastline、allot等。
- **国内厂家**：华为、绿盟科技、安天、观成科技、上海观安、科来软件、东华软件、网鼎芯睿、上海纽盾、恒安嘉新、Panabit、亚信安全等

现有加密检测方法

- 使用与协议详细信息无关的数据元素（例如流中数据包的长度、到达时间），这些数据元素同时适用于加密和未加密的数据流。
- 思科的ETA通过被动监测、提取相关数据元素、行为建模和机器学习与基于云的全局视图相结合，识别加密流量中的恶意流量。
- ETA中的加密流量分析提取四个主要数据元素：初始数据包、数据包长度和时间序列、字节分布和TLS特定特性。

未知加密协议检测-研究问题和目标

研究背景：

- 较多网络应用采用公开标准的加密协议（如TLS）进行通信，但也有不少应用、恶意软件采用**私有加密协议**，而对这类未知的加密流量的检测和分析尚存研究空缺

研究目标：

- 识别加密流量中存在的非加密内容并分析其特征，构建不同加密应用协议的指纹特征，实现对加密应用协议流量的识别

研究问题和目标

通过对不同类型加密流量的负载的观察分析，可以将加密流量中的数据类型划分为三类：

① 非加密（非随机[1]）：NN

② 非加密（随机[2]）：RN

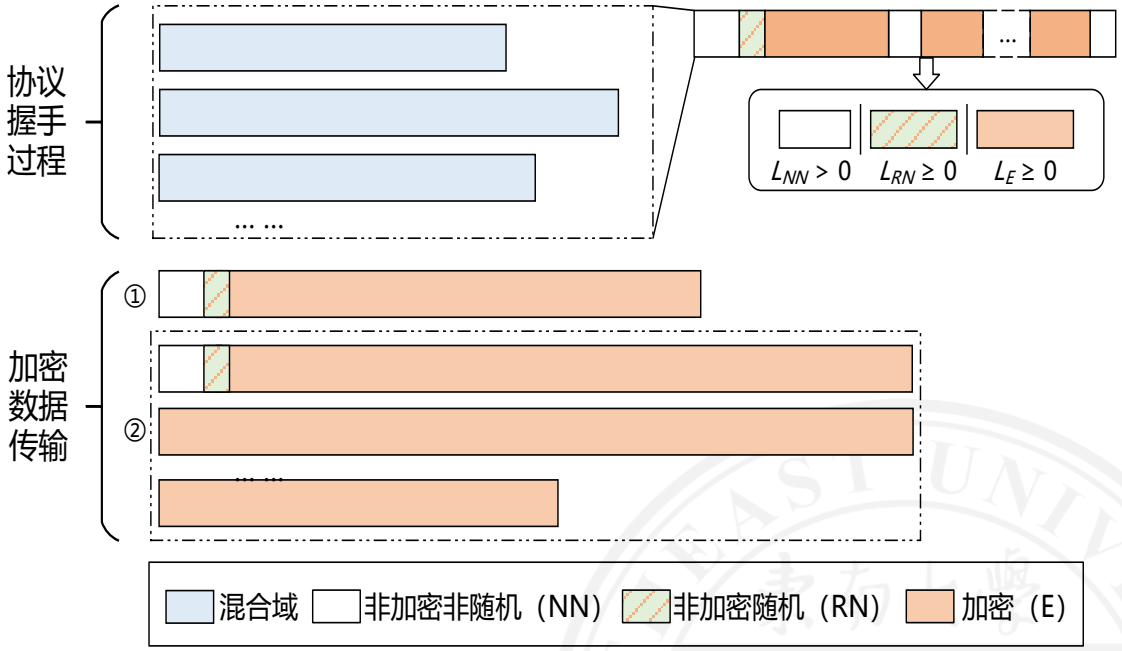
③ 加密[3]：E

TCP加密会话通常包括协议握手和数据传输两个阶段：

- 握手过程（*较多随机性较低的明文字段*）
- 数据传输（*大部分为随机数据*）

UDP加密会话由于无连接的不可靠性，每个报文中需要有额外的非加密字段保证数据正常有序接收，这些字段通常呈现为：

- *低随机性*
- 在多个报文中具有*相同或有规律的值*
- 通过其它非加密方式混淆成*高随机数据*



典型加密应用协议流量的负载随机性分布特征

报文头部随机性分析

- 基于对已知加密应用协议流量数据的分析，较多加密协议在报文头部存在 **非加密字段（随机/非随机）**，这类字段在报文头部格式中的位置、长度通常是固定不变的，因此首先对报文头部进行随机性检测和字段类型识别

在多个报文中具有相同或有规律的值，通过非加密方式混淆成的高随机非加密部分

分析过程

由于不同方向报文格式内容上的差异性，对每条单向流的数据包分别进行处理和分析：

- 对每个报文，选取头部N bit，计算前M个报文的同个b- bit块偏移的熵：

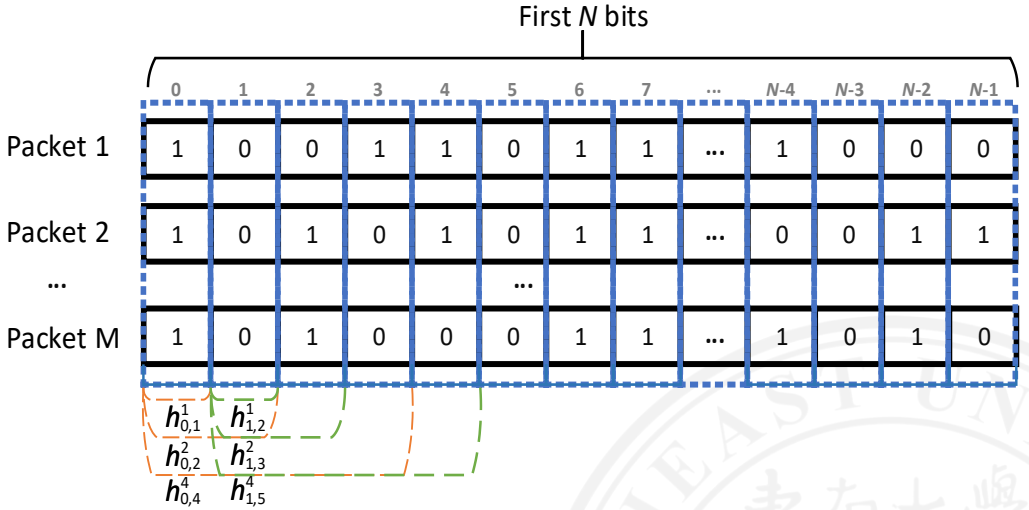
$$h^b_{(i,i+b)} = \sum_{k=0}^{2^b-1} \frac{m_i^k}{M} \log \frac{m_i^k}{M}$$

- 综合考虑在不同b（ $b = 1,2,4$）取值的熵值，每个单位bit（i^{th} bit）的随机性评估值计算如下公式，，

$$E_i = \sum_{b\in\{1,2,4\}} \lambda_b \cdot t^b_{(i,i+b)}, \qquad t^b_{[i/4]} = \frac{h^b_{(i,i+b)} - \mu^b}{\sigma^b}$$

λ_b为常数项，e.g. $\lambda_b = 1$

μ^b和σ^b分别为N-截断熵估计值在样本大小为M时的均值和方差



若$E_i < \tau$（e.g. $\lambda_b = 1$, $\tau = 9$，与**N-截断熵**的区间值相关），则该bit位置为在大多报文头部具有相对固定的值，即属于**非加密字段**；反之，则认为该bit位置为在大多数据包中的值随机性较高，标记为属于**加密字段**

实验结果

未知加密协议检测

- 在分析的结果中，根据加密数据占比以及报文头部格式，检测到一些未知的加密协议

- SRC_TO_DST方向：共有794条流满足加密比例 > 0.5，且报文头部被识别为有非加密结构
- DST_TOSRC方向：共有898条流满足加密比例 > 0.5，且报文头部被识别为有非加密结构

- 报文头部字段位置（字节偏移）：

[0, 4]

- 样本：

203.X.X.161_222.X.X.192_80_56423_TCP

```
48 a5 50 18 1f 68 f0 78  00 00 f4 00 03 f4 46 40
de 60 5e 28 f8 00 a8 ec  e5 5d 09 ae a5 07 75 b7
72 31 72 9a e8 d4 de c6  9e 69 94 09 96 60 e4 f7
de 57 26 10 d8 1e 2c ab  6a 5e 2e 52 ad 9a f5 38
```

137.X.X.90_121. X. X.198_54058_57163_UDP
157. X. X.160_121. X. X.25_9041_34947_UDP
137. X. X.90_121. X. X.198_54058_57161_UDP
183. X. X.228_121. X. X.18_5855_42634_UDP

VPN加密流量检测--研究问题和目标

研究问题

- VPN技术作为加密网络流量的主要使用技术之一，依靠加密隧道等手段，向用户提供便利的同时，也给网络监管带来了难处。因此，需针对VPN流量报文信息缺失，流量特征混淆的问题，研究VPN流量检测与识别方法。

研究目标

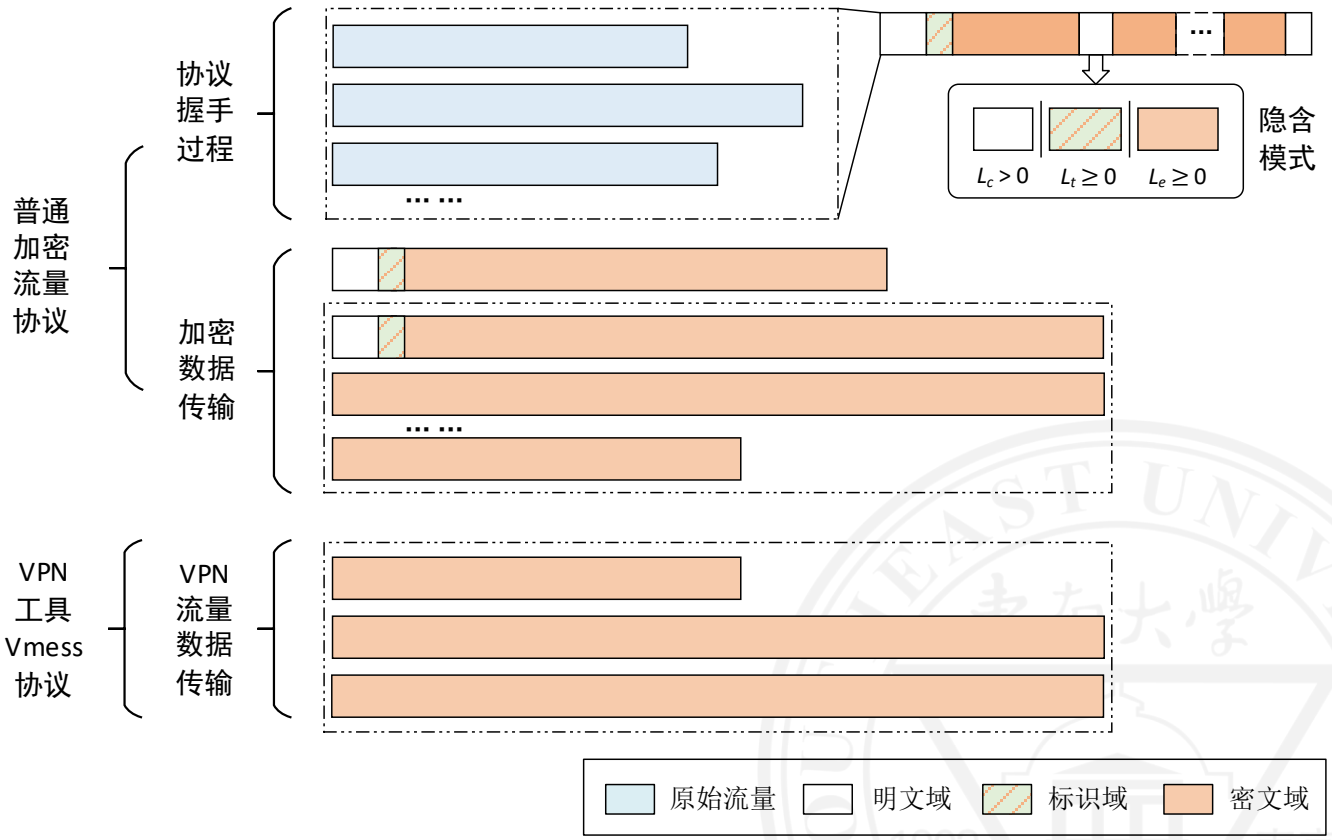
- 研究重点关注VPN加密流量的信息熵分布特性，实现对VPN加密流量的检测与识别

方　法

提出一种基于分段熵分布的VPN加密流量检测与识别方法，对VPN加密流量的信息熵分布特性进行研究，利用滑动窗口方法对VPN加密报文序列高熵、低熵区域进行划分，计算VPN加密流量的高熵低熵分布情况，并以此作为流量特征，接着使用机器学习方法进行流量识别分类，实现VPN加密流量的精准检测与识别，加强对VPN流量的有效监管，提高网络监管的效率与效果。

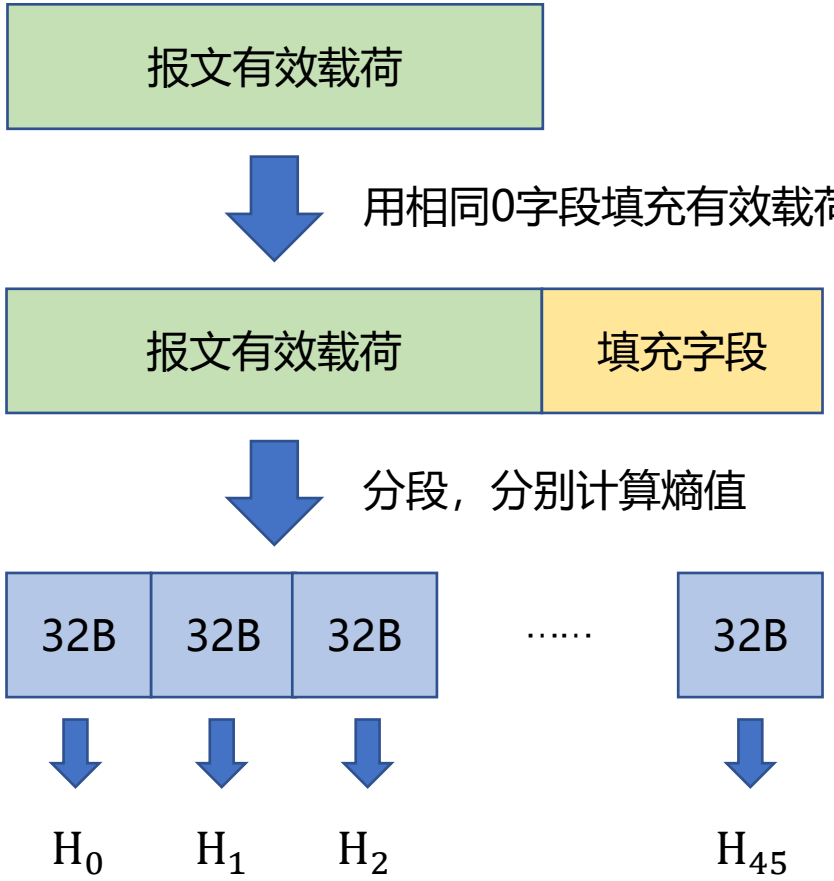
VPN加密流量熵分布

- VPN工具使用私有协议Vmess实现数据的随机化加密传输。
- Vmess协议是一种基于TCP协议的无状态协议，协议本身没有握手过程，并可以通过与其他协议的组合实现完全的流量混淆与匿名化。
- VMess协议下的VPN加密流量流量呈现高度均匀随机分布，且不含明文头部的特点



VPN实验部分-胶囊神经网络

数据包特征选取



实验使用的51组特征

编号	特征
1	报文方向
2	TCP ACK flag
3	TCP PUSH flag
4	报文长度与32的比值
5	报文有效载荷总熵值
6	第1段有效载荷熵值特征 ($\lvert H_N(U) - H_0 \rvert$)
	……
51	第46段有效载荷熵值特征 ($\lvert H_N(U) - H_{45} \rvert$)

实验数据集

ISCX VPN-nonVPN 公开数据集

流量类型	应用内容
邮件	Email、Gmail(SMPT.POP3,IMAP)
VPN-邮件	
聊天	ICQ、AIM、Skype、Facebook
VPN-聊天	
视频播放	Vimeo、Youtube、Netflix
VPN-视频播放	
文件传输	Skype、FTPS、SFTP
VPN-文件传输	
语音通话	Facebook、Skype、Hangouts
VPN-语音通话	
文件下载	uTorrent、Bittorrent
VPN-文件下载	

实际采集得到的VPN数据集

流量类型	应用内容
邮件	QQ邮箱、网易邮箱、
VPN-邮件	
聊天	QQ、微信、TIM
VPN-聊天	
视频播放	Youtube、腾讯视频、爱奇艺、
VPN-视频播放	
文件传输	微信、QQ
VPN-文件传输	
语音通话	微信、QQ
VPN-语音通话	
文件下载	迅雷、百度网盘
VPN-文件下载	

实验结果

胶囊神经网络具体分类结果

数据集	ISCX 公开数据集	实际采集得到的VPN数据集
准确率	99.87%	96.34%
精确率	99.74%	93.19%
召回率	100.00%	100.00%

不同机器学习对比分析

方法	SVM	决策树	随机森林	胶囊神经网络
分类准确率	90.21%	92.95%	95.89%	96.34%
召回率	86.24%	88.80%	94.44%	100.00%

◆ 加密视频流量识别

研究意义：

近年来，视频流量在互联网流量中的占比的逐年攀升，与此同时，为了保护用户隐私，加密技术被广泛应用。如何从加密的视频数据中抽取出网络安全防护和网络管理需要的信息已经成为网络管理中亟待解决的问题。在做到保护用户的隐私的同时，及时发现因特网中传递的危害国家和社会安全的信息，是目前加密视频流量研究中的一大挑战。

研究思路：

现有的针对加密视频流量的研究普遍采用流特征，但是方法的效果差强人意。为了提高加密视频流量识别的准确率与可扩展性，本方法首次提出将HTTP头部长度和TLS片段相关信息作为主要特征，并提出使用视频片段的长度作为构建视频指纹的关键信息。

加密视频流量识别

研究方法：

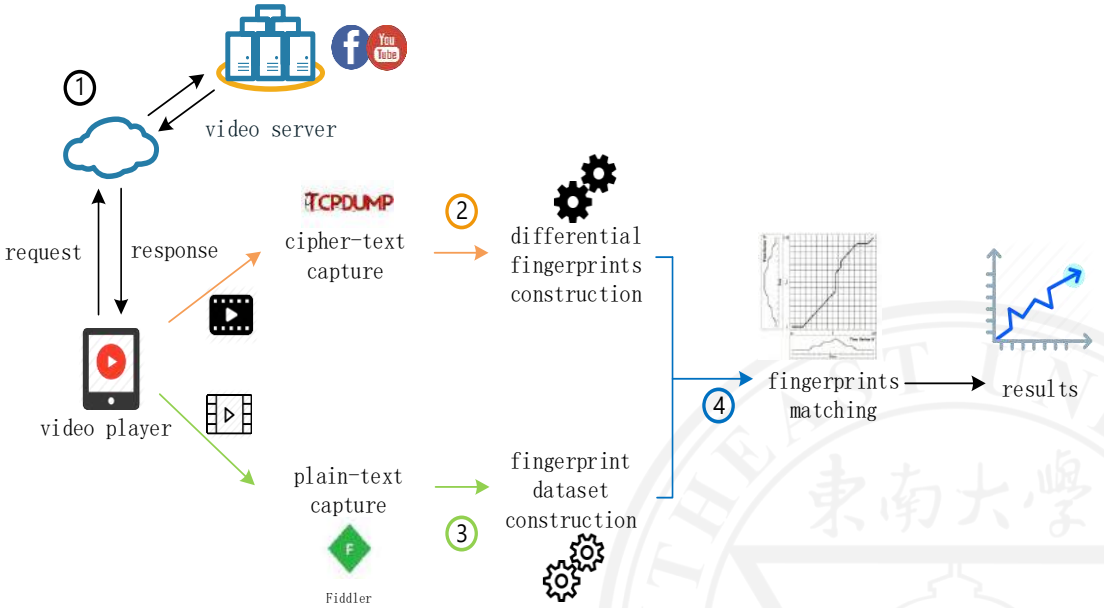
（1）方法流程

① 分别利用tcpdump和Fiddler抓取视频密文和明文数据；

② 利用DASH传输视频的特点，从密文数据最大程度还原加密视频片段大小，并根据视频片段的大小构建视频指纹；

③ 使用抓取的明文数据构建视频明文指纹库；

④ 利用序列匹配算法将待匹配的加密视频指纹与指纹库中的指纹进行匹配，最终确定出待匹配加密视频的身份。

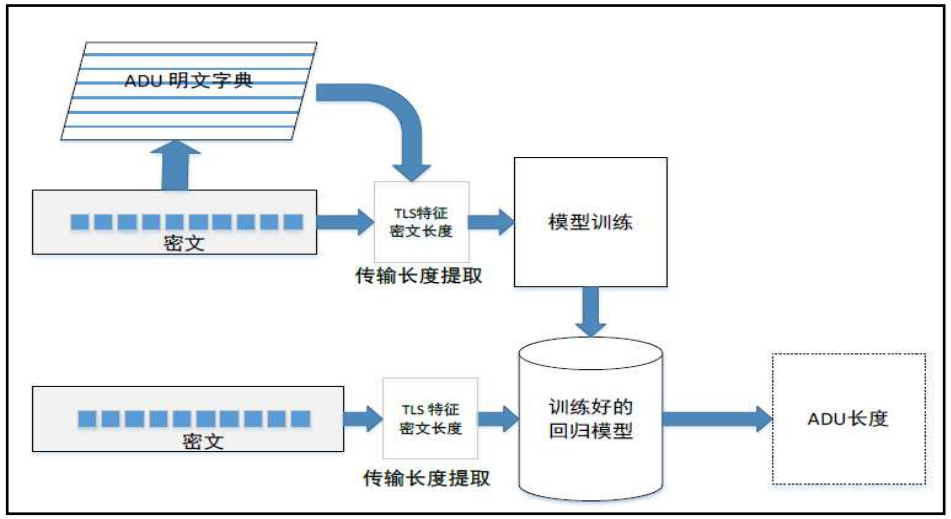


◆ 加密视频流量识别

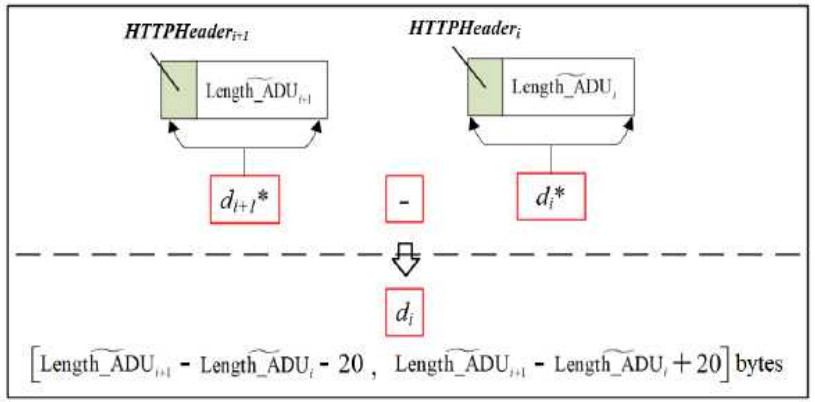
研究方法：

（2）加密视频指纹构建方式

针对加密视频数据封装格式的不同，本方法分别提出了两种构建指纹的方法，如下图所示。在方式1中，可以直接利用还原后的视频片段大小及其序列构成代表该视频的指纹与指纹库中进行匹配。在方式2中，利用还原后的相邻加密视频片段的差值构建指纹。



▲加密视频指纹构建方式1



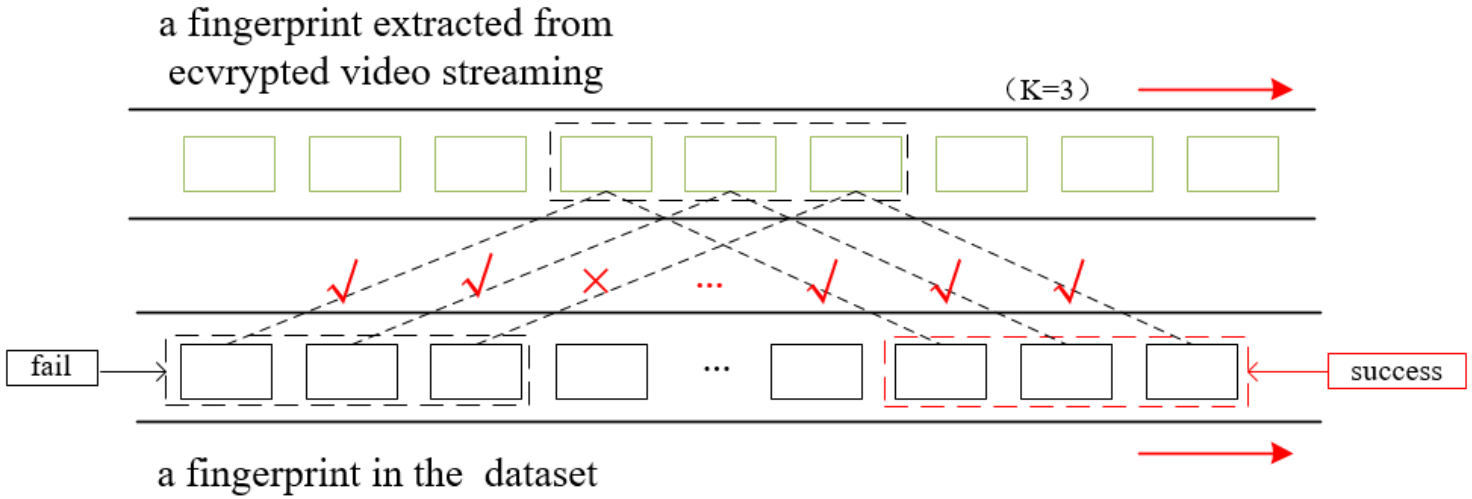
▲加密视频指纹构建方式2

加密视频流量识别

研究方法：

（3）序列匹配算法

在指纹匹配过程中，采用基于欧氏距离的K段匹配算法进行加密指纹与明文指纹之间的匹配。当加密指纹中连续K个元素与明文指纹的K个元素一致时，认定此时匹配为成功。



◆ 加密视频流量识别

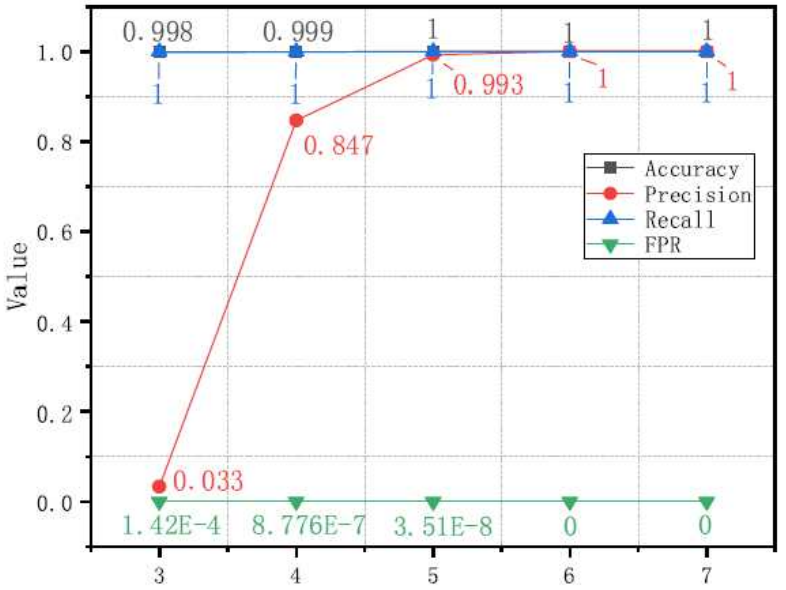
研究方法：

（4）实验结果

为了验证本方法的真实性能，基于真实视频构建了一个规模为20万级的指纹库，并在该指纹库中进行了实验。结果显示只需要使用由超过5个连续视频片段构成的指纹，本方法就可以达到准确率、查准率、查全率为100%，假阳率为0的效果。

长度指纹修正方法	k	准确率	查准率	查全率	假阳率
HHTF 方法	2	99.9997%	65.7%	100%	2.54×10^{-6}
	3	**100%**	**100%**	**100%**	**0%**
Reed 方法	2	49.58%	9.64×10^{-6}	100%	50.41%
	3	99.89%	0.47%	100%	0.10%

▲指纹构建方式1中实验结果

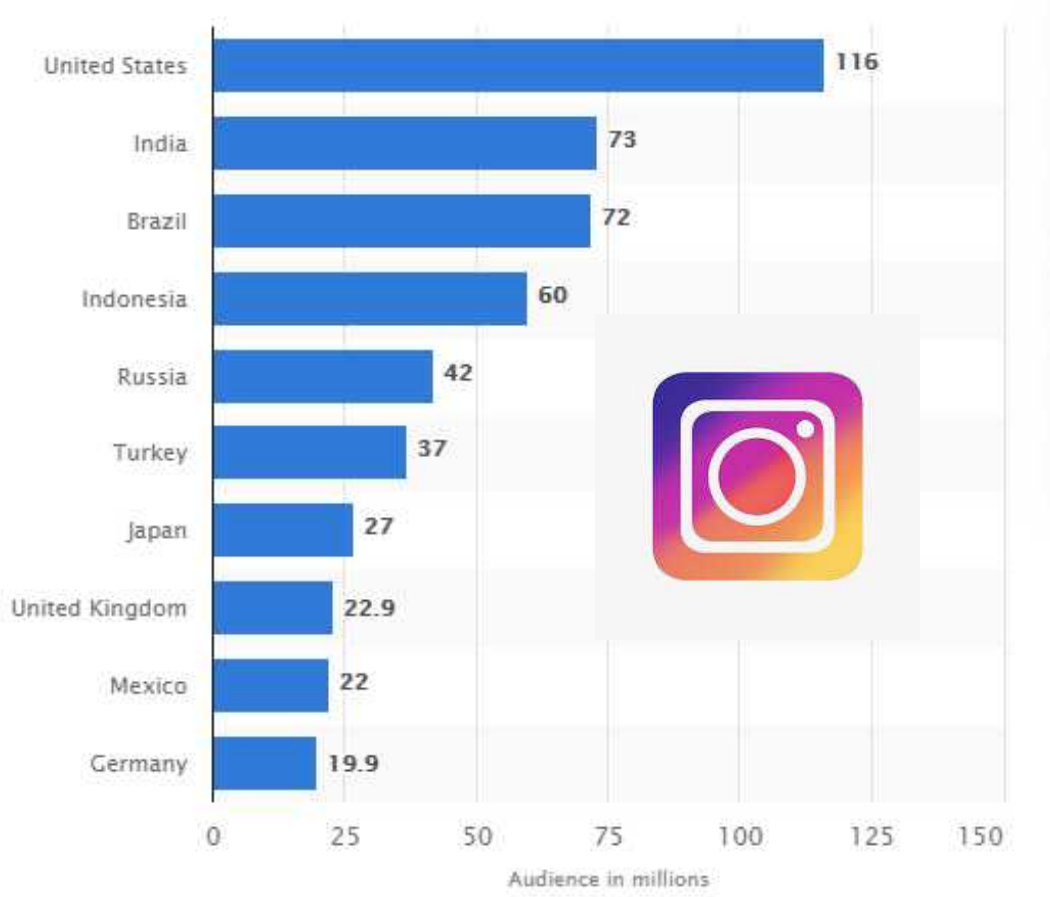


▲指纹构建方式2中的实验结果

◆ 加密流量用户行为识别（Instagram）

研究现状：

在加密流量用户行为识别的研究问题中，基于端口的检测和基于深度包检测（数据流特征分类技术） 不再适用，这些方法中分类算法多采用K近邻、决策树、朴素贝叶斯等，特征选取往往基于数据包大小分布、往返时间、时间差等。这些方法的缺点是数据丢包、重传影响大，适用性低。



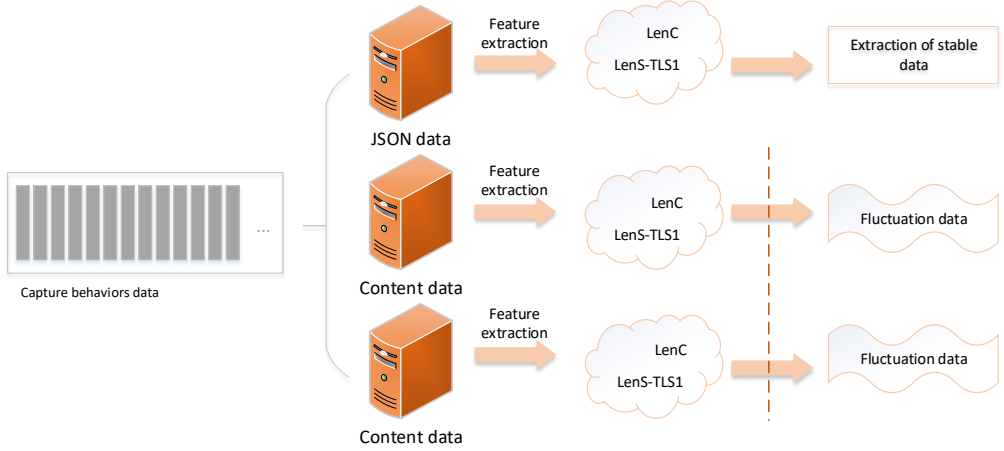
截止2019年10月的Instagram用户数量排名领先的国家

◆ 加密流量用户行为识别（Instagram）

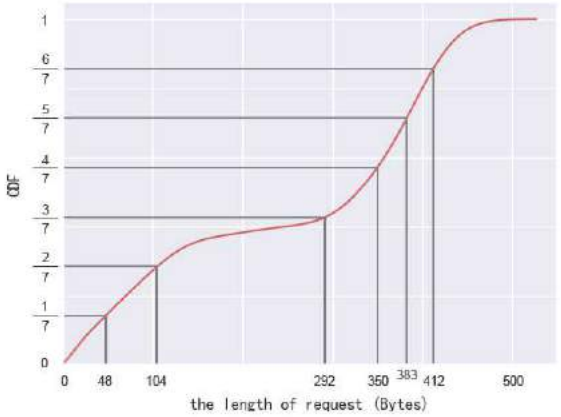
研究方法：

① Instagram社交应用采用RESTful架构，该架构在应用中主要包括JSON、JPEG以及MPEG类型的数据，我们将JSON数据归类于稳定数据，JPEG和MPEG归类于波动数据，利用相关特征将两种数据分类，提取出稳定的数据。

② 从稳定的数据中提取主要服务器的数据长度特征，根据不同服务器的数据量划分区间，组合获得16*16+12*12+7*7=449维的高维特征空间。



稳定数据的提取



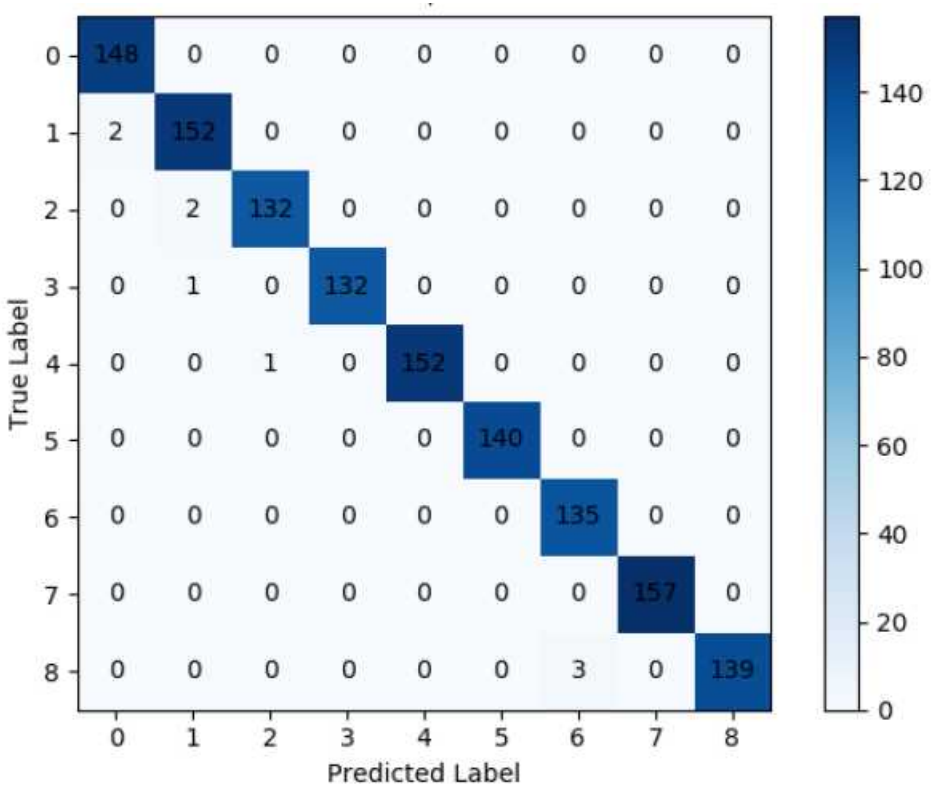
服务器长度范围计算

Server Name	Ratio of Data	Number of Ranges
Server a	56%	16
Server b	32%	12
Server c	12%	7

◆ 加密流量用户行为识别（Instagram）

研究方法：

③ 使用SVM算法对数据进行分类



用户行为分类的混淆矩阵

用户行为分类的各个指标结果

User Behavior	Accuracy	Precision	Recall	FPR
Enter	99.8%	98.7%	100%	0.17%
Login	99.6%	98.1%	98.7%	0.26%
Posting	99.8%	99.2%	98.5%	0.09%
Browse	99.9%	100%	99.2%	0%
Repost	99.9%	100%	99.3%	0%
Return	100%	100%	100%	0%
Favorite	99.8%	97.8%	100%	0.26%
Set up	100%	100%	100%	0%
Exit	99.8%	100%	97.9%	0%
Average	99.8%	99.3%	99.3%	0.09%

◆ 小结

研究方法：

① 加密流随机性分析：加密流中混合着可识别的、未加密的内容，提出了熵的方法将单条流中的加密和非加密比特流分离的方法，进而建立加密流指纹，采用机器学习等方法进行加密流量分类。

② 对VPN加密流量的信息熵分布特性进行研究，提出一种基于分段熵分布的VPN加密流量检测与识别方法，实现VPN加密流量的精准检测与识别，加强对VPN流量的有效监管。

③ 加密视频流特征分析应用：加密视频是目前主要的加密流量，通过深入分析自适应码流传输模式HLS和DASH的基础上，建立根据视频块的统计特征建立识别模型，视频块特征从网络层提取，实现对加密视频的QoE特征提取。

東南大學 SOUTHEAST UNIVERSITY

网络空间安全学院
School of Cyber Science and Engineering

InForsec“走进蚂蚁集团”

网络空间安全2020大学生夏令营

欢迎各位同学报考我院！

程 光

2020 年 7 月 18 日