管中窥豹，从DNS浅谈数据安全分析

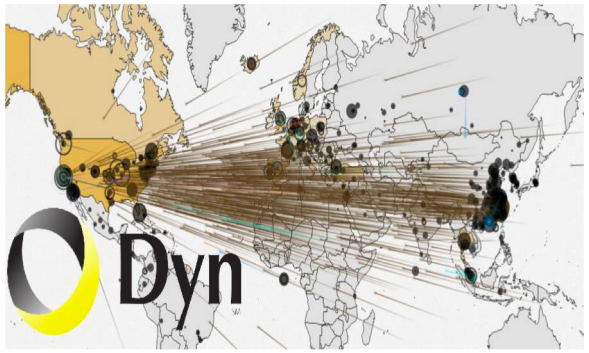
刘保君

清华大学计算机系 博士生

2020年07月19日

DNS + 安全

- 当大家谈起“DNS+安全”，大家一般指的是什么?
  - 第一类：攻击者破坏DNS的通信过程



拒绝服务

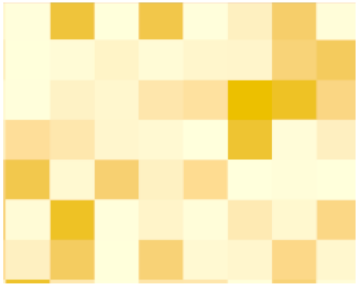
域名劫持

缓存污染

DNS + 安全

- 当大家谈起“DNS+安全”，大家一般指的是什么?
  - 第一类：攻击者破坏DNS的通信过程
  - 第二类：攻击者滥用DNS的灵活特性

恶意软件



色情赌博

隐蔽通信

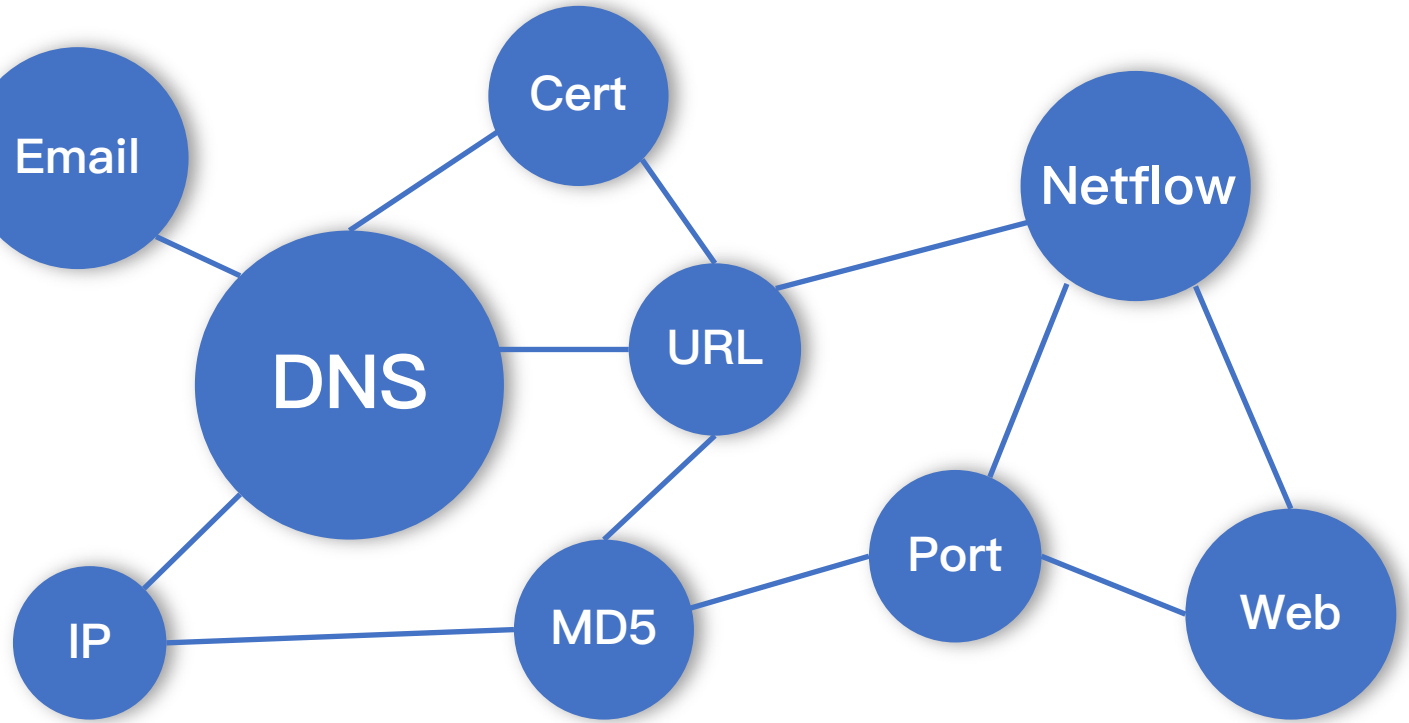
DNS + 数据安全分析

- 当大家谈起“DNS+安全分析”，大家一般指的是什么?

Source: Hongliang Liu (Phunter)

DNS数据安全分析的整体目标

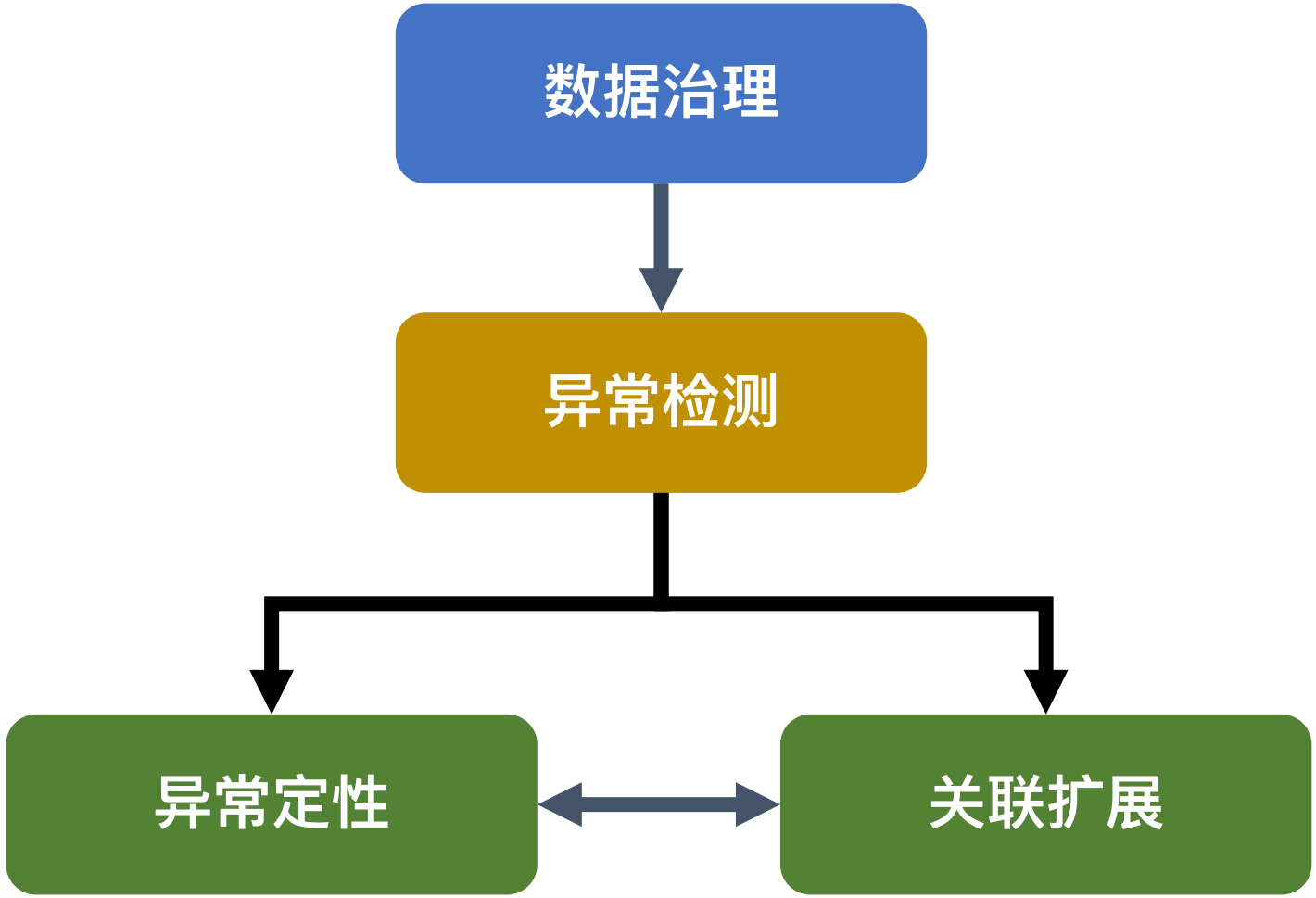
- 看得全
- 发现早
- 挖得深
- 串得广



Visibility: Security insights from big data

DNS数据安全分析的整体目标

- 看得全
- 发现早
- 挖得深
- 串得广



(1) 数据采集与治理

- **数据采集**
  - 数据采集的位置与方式
- **数据治理**
  - 格式化、归一化、去噪 ……
- **常见挑战**
  - 工程难度 & 数据多轮次加工及融合

1.1 数据治理，奇葩数据面面观

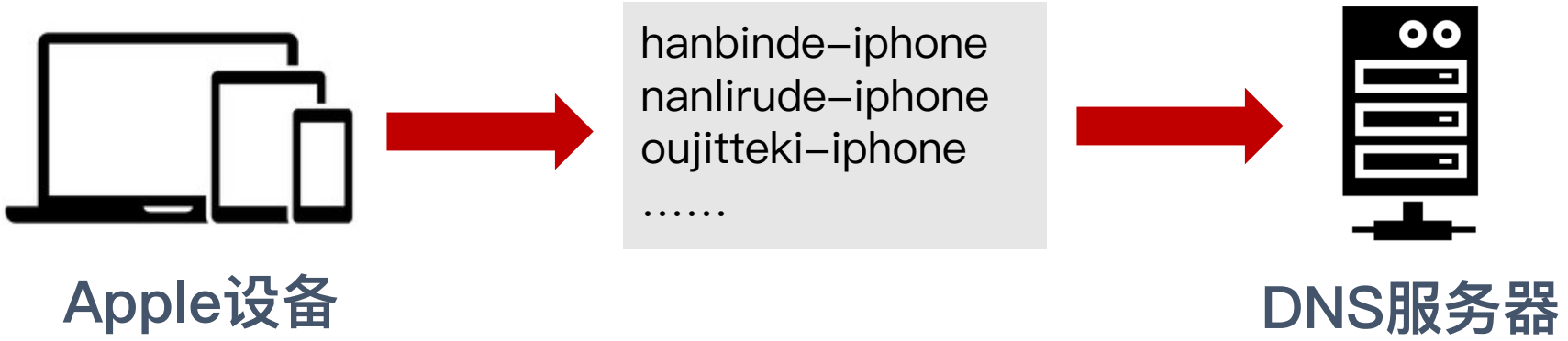
- **它们究竟属于数据噪音，还是属于安全问题?**
  - 第一类：奇葩IP（特殊用途的IP地址）
    - 1.1.1.1/8.8.8.8/114.114.114.114 …；保留地址，192.168.0.0/16 …

rdata	rrname	type	time_end
015ts8cep1beser6a9eci.32data.win	8.8.8.8	A	2016-07-17 00:30:16
00en7tsb.5r864.cn	8.8.8.8	A	2019-07-24 12:21:36
01a737ad191abffe8da2c705fa18dd23d.32data.win	8.8.8.8	A	2016-07-17 15:02:34
01dp6usca37tbiw7ry.32data.win	8.8.8.8	A	2016-07-17 09:15:36
0205.r6830l.com	8.8.8.8	A	2016-02-19 17:03:27

Source: www.passivedns.cn

1.1 数据治理，奇葩数据面面观

- **它们究竟属于数据噪音，还是属于安全问题?**
  - 第二类：奇葩域名 （无法解析的域名）
    - 安全危害：基于DNS流量，实现跟踪定位用户终端设备



(2) 异常检测

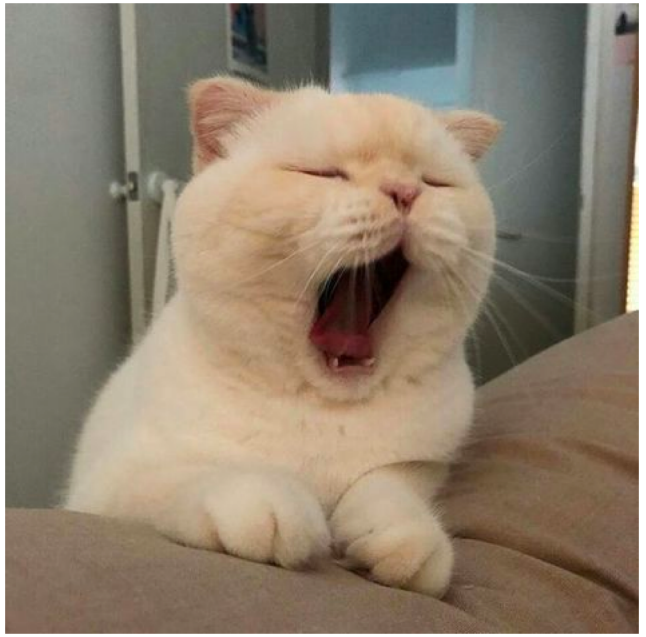
- **基于规则的路线**
  - 统计分析；选择白基线
- **基于学习的路线**
  - 威胁模型、特征工程、黑白样例、设计算法、评估验证

基于规则的检测更适合产出粗粒度告警

Source: https://scan.netlab.360.com

基于学习的检测真的很无聊吗?

- 观点：过程远比结果更重要
  - 威胁模型
  - 特征工程
  - 黑白样例
  - 设计算法
  - 评估验证



www.pinterest.com/pin/232779874474950066

2.1 明确安全威胁模型

- **真实案例**
  - 目标：检测“恶意域名”
  - 方法：收集黑白域名样例，采集网络流量，训练模型
- **误区**
  - 一：恶意域名不明确（DGA, SEO, Phishing）
  - 二：检测模型不可解释，难以落地部署

www.pinterest.com/pin/1221602086208588338

2.2 根据安全场景，推演关键特征

- **特征工程：以DGA域名检测场景为例**
  - 传统检测方法通常基于域名词法，包括长度、可读性、熵等

家族一（Bamital）
cd8f66549913a78c5a8004c82bcf6b01.info
aa24603b0defd57ebfef34befde16370.cz.cc

家族二（Banjori）
earnestnessbiophysicalohax.com
kwtoestnessbiophysicalohax.com

适用类型

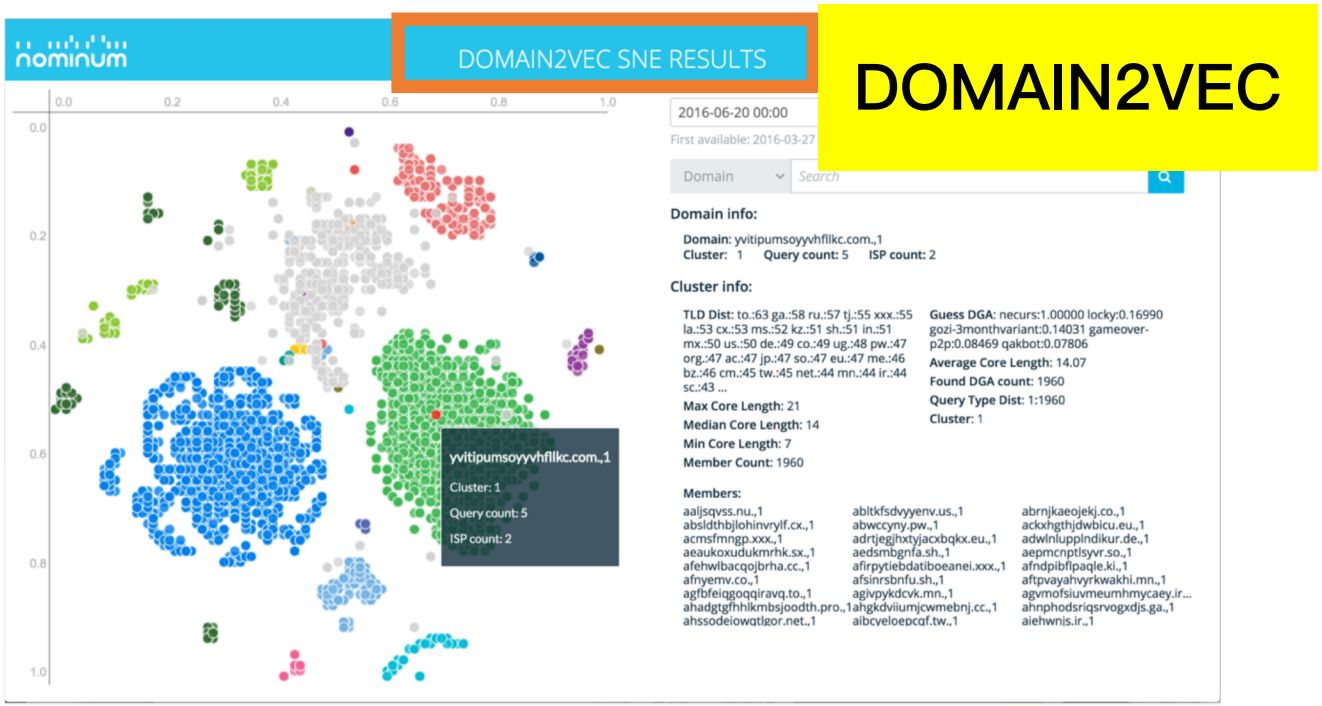
家族三（Matsnu）
world-bite-care.com
activitypossess.com

家族四（Suppobox）
sharmainewestbrook.net
tablethirteen.net

不适用类型

2.2 根据安全场景，推演关键特征

- 特征工程：以DGA域名检测场景为例
  - 技术路线：探索DGA域名在查询请求中的特征



安全检测场景的对抗特性

Source: Hongliang Liu (Phunter)

2.3 收集黑白样例时，确认数据集一致性

- 什么是恶意域名：以域名检测场景为例
  - 僵尸网络的C&C域名

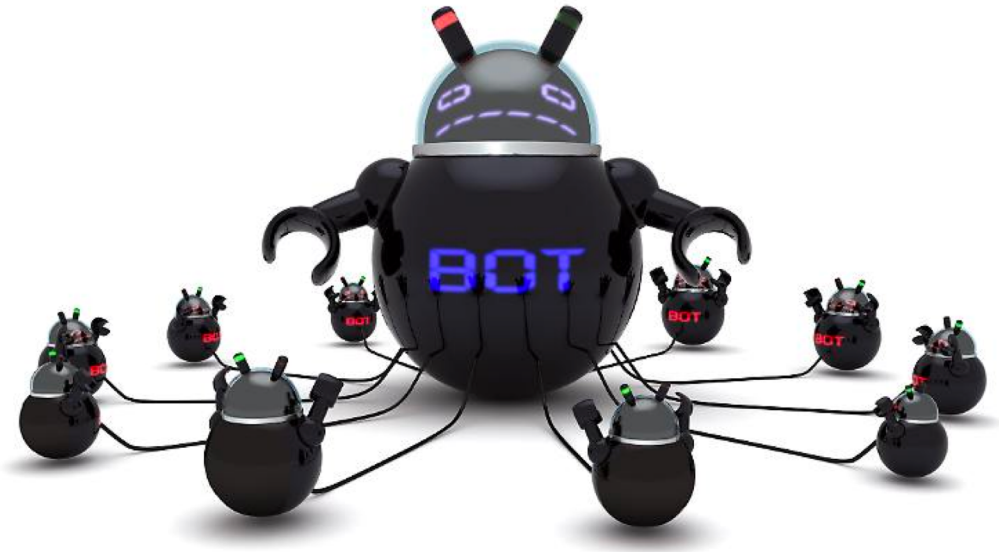
eef795a4eddaf1e7bd79212acc9dde16.net

fff1834cbcd5ba96ca75fdae9c44cf5d.net

35262768764bd6c908c386b532a3dc2f.net

7817b2bcf25367beb24b3270232e67e5.net

247b05f526ca169f6eff42dff26155d3.net



2.3 收集黑白样例时，确认数据集一致性

- 什么是恶意域名：以域名检测场景为例
  - 僵尸网络的C&C域名

eef795a4eddaf1e7bd79212acc9dde16.net

fff1834cbcd5ba96ca75fdae9c44cf5d.net

35262768764bd6c908c386b532a3dc2f.net

7817b2bcf25367beb24b3270232e67e5.net

247b05f526ca169f6eff42dff26155d3.net

2.3 收集黑白样例时，确认数据集一致性

- 什么是恶意域名：以域名检测场景为例
  - 僵尸网络的C&C域名

eef795a4eddaf1e7bd79212acc9dde16.net

fff1834cbcd5ba96ca75fdae9c44cf5d.net

35262768764bd6c908c386b532a3dc2f.net

7817b2bcf25367beb24b3270232e67e5.net

247b05f526ca169f6eff42dff26155d3.net

域名注册局

VERISIGN

域名注册商

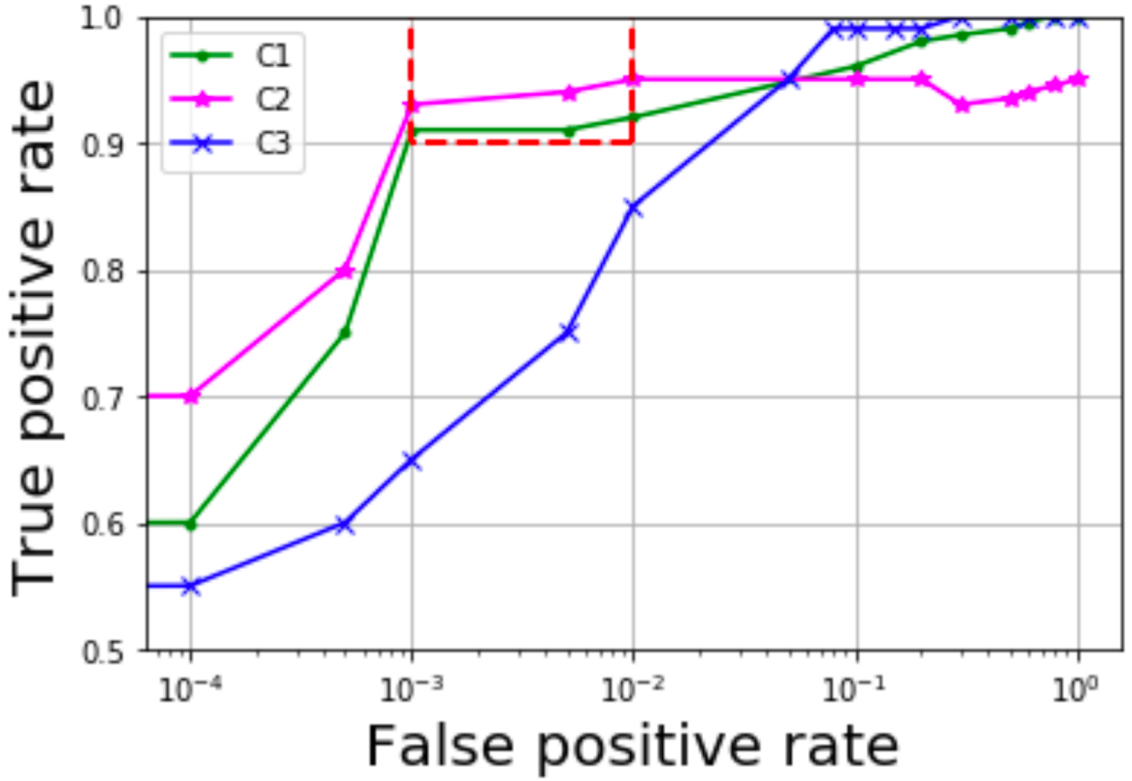
GoDaddy

域名注册人

Depends on WHO you ask and WHEN you ask.

2.4 模型评估

- 如何评价不同模型效果的优劣?



Source: Sadia Afroz (ICSI), Enigma 2020.

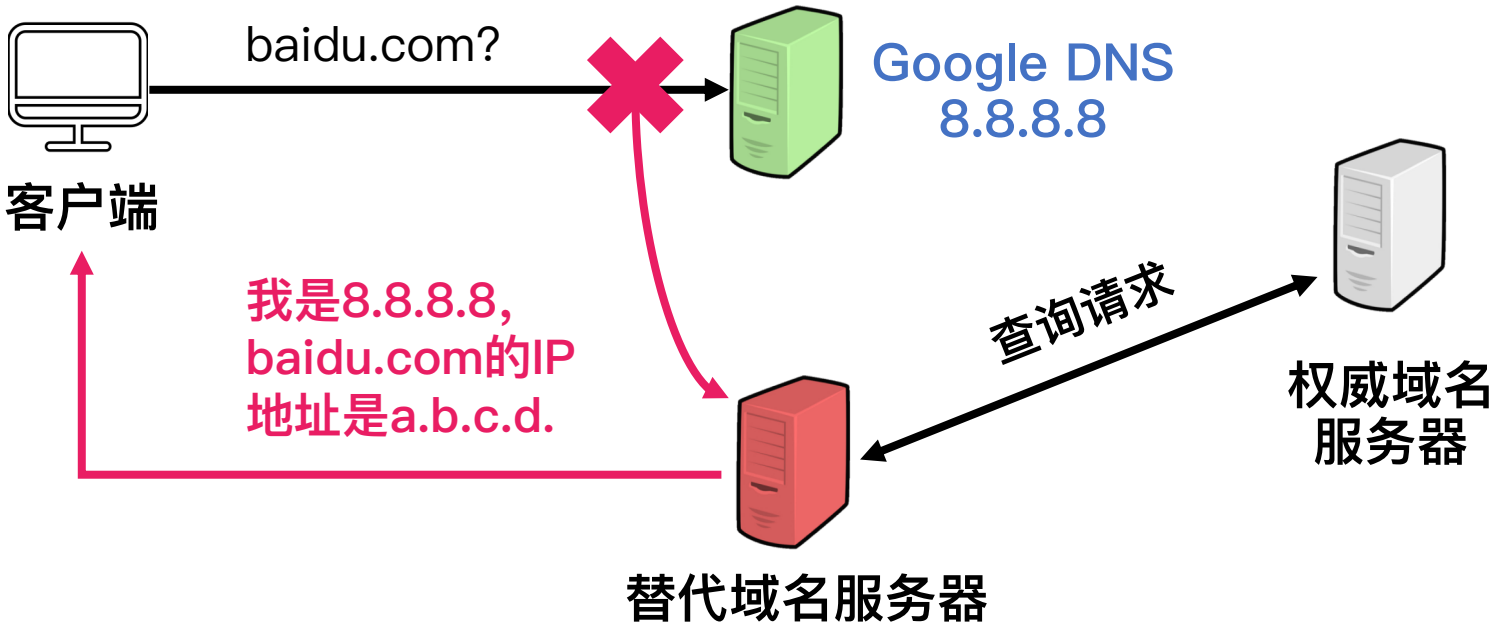
(3) 异常定性

- 专家知识
- 跟踪系统
  - 网络爬虫、高交互蜜罐、沙箱日志 ……
- 用户端数据

我们要接受一个残酷的现实

3.1 利用网络测量平台，补充用户端数据

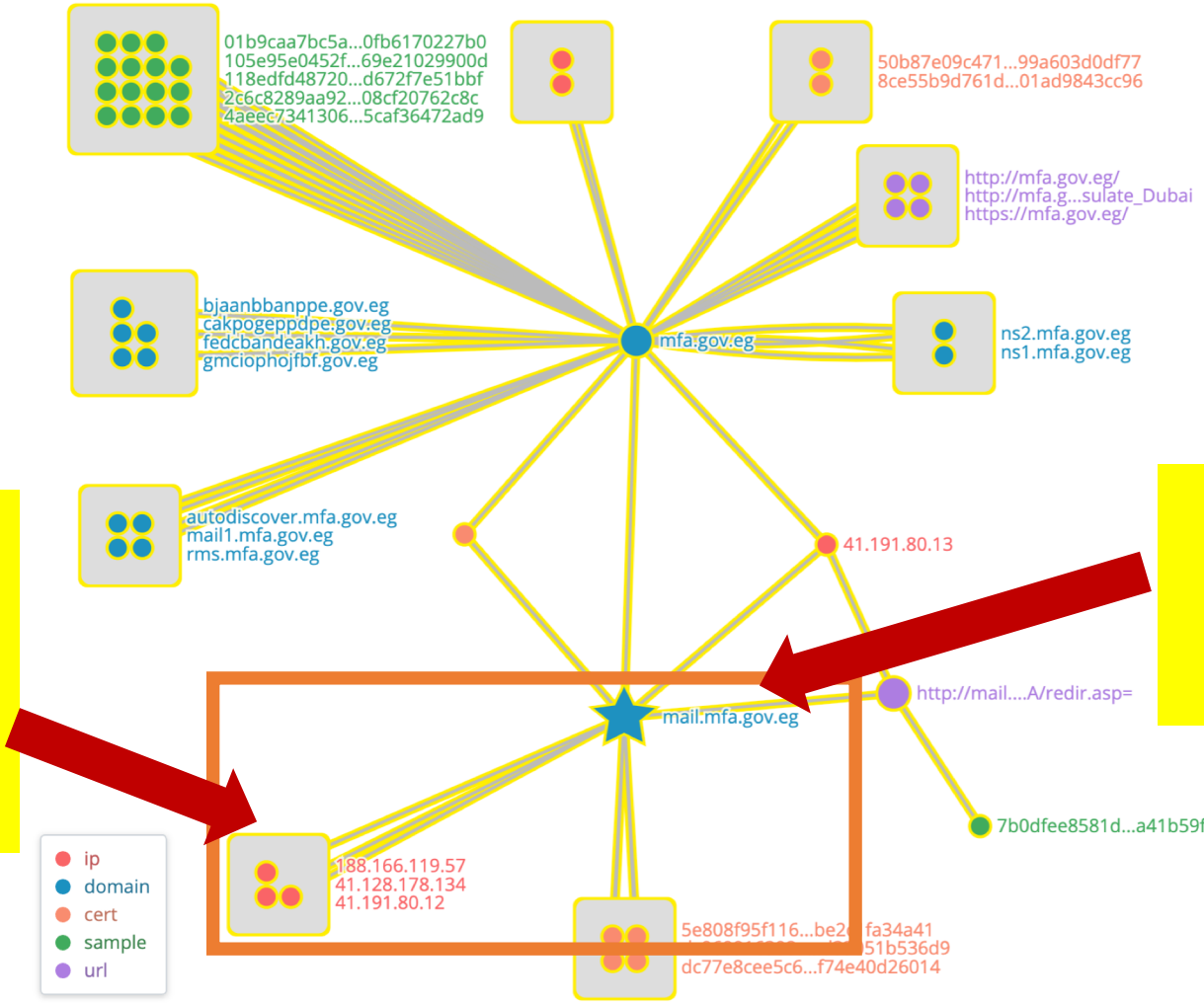
- **域名解析链路劫持**
  - 劫持者伪造数据包源地址，劫持用户查询请求



(4) 关联扩展

- **图数据系统的威力**
  - 深入的数据融合
  - DNS, IP, Port, Certificate, MD5, NetFlow, URL ……
- **真实案例：DNSpionage事件**
  - 典型的高级可持续攻击（APT）
  - 影响50多个中东地区的政府机构，雇员邮件通信记录被窃取

4.1 图系统在关联扩展中的作用



188.166.119.57

41.128.178.134

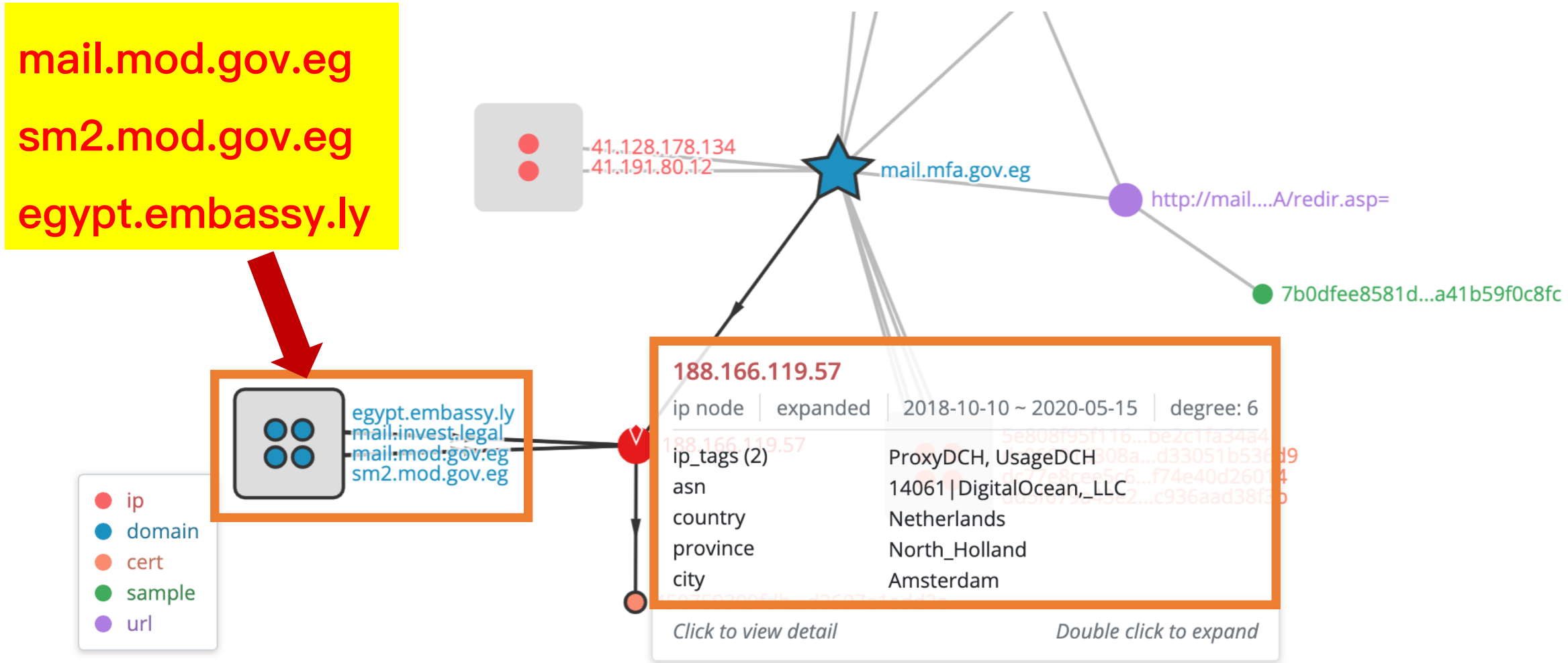
41.191.80.12

mail.mfa.gov.eg

域名解析异常

Source: 360 Netlab Graph Query System

4.1 图系统在关联扩展中的作用



Source: 360 Netlab Graph Query System

总结

- DNS安全分析只是大数据驱动安全能力在DNS领域的应用

- 祝各位参赛同学在DataCon 2020中取得好成绩！