DataCon 2019 赛题分享

刘保君

2019年6月5日

About

- 关于DataCon比赛
  - 全称：大数据 安全 分析比赛
  - 赛题方向设置
    - DNS安全方向
    - ~~加密流量方向~~（跑路了）
    - 恶意代码方向
    - 攻击溯源方向
- 关于我
  - 360 Netlab 菜鸡实习生
  - DataCon全程负责打嘴炮
  - 主页：https://www.liubaojun.org

DNS方向赛题 设想 & 结果

- 面向人群尽量广泛
  - 前期知识：零储备
  - 专业方向：机器学习 or 安全 均可
- 赛题设置具有区分度
  - 第一道题：DNS基础安全知识
  - 第二道题：DGA域名检测
  - 客观分数： 80分 （比例 6:4）
  - Writeup分数：20分
- 实际结果还能凑合
  - 96支队伍，共500余人

赛题设计（一）

- 题目：DNS基础安全知识
- 场景说明
  - 模拟网络管理员的攻击分析过程。
  - 给定的流量中，包含五种DNS攻击流量。选手需要准确判断出五种DNS攻击，并说明pcap文件中哪些数据包是攻击流量。
  - 提供流量为pcap格式数据包，大小为2.6GB（人造流量）
- 答题时间
  - 跨度：4月1日 -- 4月8日；每天只判一次分；
- 提交形式
  - 提交CSV文件：PacketID，AttackID

赛题设计（一）

- 判分规则
  - 大原则：惩罚误报
  - 该题得分由每一类得分累加而成，每一类攻击占比20%
  - 选手提交的5个类别，将依次与标准答案中的5类进行比较
  - 细节：全排列计算

$$Score_{i,j} = max\left(\frac{1}{N}\cdot w\cdot\sum_{k=1}^{m}([P_k \in FTrue_j]),0\right)$$

各项符号的含义：

$FTrue_j$：标准答案的第j类

N：$FTrue_j$的长度

m：选手提交的第i类的长度

P_k：选手提交的第i类中的第k个数据包

[]：其中为判断语句。语句成立时，值为 1；不成立时，值为-1

$max()$：取最大值函数

w：积分权重

赛题设计（二）

- **题目：DGA域名家族聚类**
- **场景说明**
  - 选手需从给定的数据包中，通过数据分析，发现其中存在的DGA域名，并自行设计算法完成DGA域名的家族聚类
  - 评测关键点：域名识别的准确度 & 家族聚类的准确度
  - [注：未指定家族类数]
  - 提供流量为pcap格式数据包，大小为2.1GB（人造流量）
- **答题时间**
  - 跨度：4月9日 – 4月28日；采取不定期阶段性判分的策略；
- **提交形式**
  - 提交CSV文件：Domain，Family-Code

赛题设计（二）

- **DGA家族示例**
  - 选择部分具有代表性的DGA家族

家族名称	TLD	e.g.	SLD
bamital	[co.cc, cz.cc, info, org]	cd8f66549913a78c5a8004c82bcf6b01.info	Like md5 hash value; 26 domains per day
banjori	Same as seed domain	earnestnessbiophysicalohax.com	Only change the first 4 letters of the seed domain; 219
blackhole	[ru]	mkjdkbwuxcnuxtqd.ru	Fix length of 16, a-z, 2 domains per day
chinad	[com, org, net, biz, info, ru, cn]	qowhi81jvoid4j0m.biz	A fix length of 16, mix a-z and 0-9; 1000 domains per
conficker.a	[com, net, org, info, biz]	gfedo.info	A length of 5-11, a-z chars; 250 domains per day
cryptolocker	[com, net, biz, ru, org, co.uk, info]	nvjwoofansjbh.ru	A length of 12-15, a-y; 1000 domains per week
dyre	[cc, ws, to, in, hk, cn, tk, so]	l54c2e21e80ba5471be7a8402cffb98768.so	Fix length of 34, 1 char[a-z] + 33 characters from SHA
emotet	[eu]	grdawgrcwegpjaoo.eu	Fix length of 16, a-y; 96 new domains per day
gameover	[com, org, biz, net]	14dtuor1aubbmjhgup7915tlinc.net	A length of 20-28, mix a-z and 0-9; 1000 domains per
matsnu	[com]	mattermiss-type.com	Combined 2~3 words from two predefined dictionaries
murofet	[biz, info, org, net, com]	uqiqvqylwlhutwvh.info	A length of 8-16, a-z; 1020 domains per day
mydoom	[com, biz, us, net, org, ws, info, in]	wmhmqsqsqa.in	A length of 10, ["aehmnpqrsw"]; 51 domains per day
nymaim	[com, org, biz, net, info, ru, in, xyz, pw]	onrfza.info	A length of 5-12, a-z; 30|128 domains per day
padcrypt	[com, co.uk, de, org, net, eu, info, online, co, cc, website, tk, ga]	mdadbfcmnelbfbac.website	
proslikefan	[eu, biz, se, info, com, net, org, ru, in, name]	nuipkjqarq.in	A length of 6-13, a-z; 100 domains per day

赛题设计（二）

- 沙箱流量示例
  - 流量特征较为明显

7	140.771836	172.16.1.170	172.16.1.1	DNS	82	Standard query 0x9d6c	A 1iikzhymmebrhizl8x.biz
9	143.500379	172.16.1.170	172.16.1.1	DNS	81	Standard query 0xe31e	A sg8ehpgubx4zq3xjyc.ru
10	144.500631	172.16.1.170	172.16.1.1	DNS	81	Standard query 0xe31e	A sg8ehpgubx4zq3xjyc.ru
12	147.038546	172.16.1.170	172.16.1.1	DNS	81	Standard query 0x0f91	A q37sdffispwof8fqap.ru
13	148.038568	172.16.1.170	172.16.1.1	DNS	81	Standard query 0x0f91	A q37sdffispwof8fqap.ru
15	150.577958	172.16.1.170	172.16.1.1	DNS	81	Standard query 0x9be8	A yyt28ece7jrkvi3e4y.ru
17	153.527280	172.16.1.170	172.16.1.1	DNS	81	Standard query 0xd326	A 7wgho36ggv2gyxh57v.ru
19	156.472948	172.16.1.170	172.16.1.1	DNS	82	Standard query 0x5f51	A odhfnirc8pgjw83xd5.biz
21	159.014189	172.16.1.170	172.16.1.1	DNS	82	Standard query 0x229b	A 81faeqxyczkza5oyle.biz
23	210.382141	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xdb34	A 4c38p6d77oq6p2hn5m.biz
24	211.381586	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xdb34	A 4c38p6d77oq6p2hn5m.biz
26	213.923159	172.16.1.170	172.16.1.1	DNS	81	Standard query 0xc757	A ihqvt34azyi2dsxjf3.cn
28	216.497754	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xe4f5	A 7hruauae27422ct2k8.biz
30	217.027159	172.16.1.170	172.16.1.1	DNS	82	Standard query 0x9398	A 7hruauae27422ct2k8.biz
32	217.554961	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xcabe	A 7hruauae27422ct2k8.biz
34	219.279354	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xbe81	A o257q1nfs4xkmplzgb.net
35	220.279576	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xbe81	A o257q1nfs4xkmplzgb.net
36	221.279593	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xbe81	A o257q1nfs4xkmplzgb.net
37	223.279594	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xbe81	A o257q1nfs4xkmplzgb.net
38	227.279634	172.16.1.170	172.16.1.1	DNS	82	Standard query 0xbe81	A o257q1nfs4xkmplzgb.net
40	229.816655	172.16.1.170	172.16.1.1	DNS	81	Standard query 0x3bad	A x1gh45kfjqzddmboty.ru
42	232.717673	172.16.1.170	172.16.1.1	DNS	82	Standard query 0x40e1	A 5ezgmxvkqjpdybfz5a.biz
44	235.541732	172.16.1.170	172.16.1.1	DNS	81	Standard query 0x8374	A uclyngho5sfnpsiwpv.ru
45	236.541601	172.16.1.170	172.16.1.1	DNS	81	Standard query 0x8374	A uclyngho5sfnpsiwpv.ru
46	237.541585	172.16.1.170	172.16.1.1	DNS	81	Standard query 0x8374	A uclyngho5sfnpsiwpv.ru

赛题设计（二）

- 判分规则
  - 原则：惩罚误报 (同样)
  - 1) DGA域名识别准确性

$$Score_1 = max\left(\frac{1}{N} \cdot w \cdot \sum_{k=1}^{m}([D_k \in FTrue]), 0\right)$$

各项符号的含义：

$FTrue$：标准答案集

N：标准答案集（$FTrue$）的长度

m：选手提交的域名列表长度

D_k：选手提交的域名列表中，第 k 个域名

[]：其中为判断语句。语句成立时，值为 1；不成立时，值为-1

$max()$：取最大值函数

w：积分权重

赛题设置（二）

- 判分规则
  - 原则：惩罚误报(同样)
  - 2) DGA家族聚类准确性

$$Score_2 = \frac{1}{N} \cdot \sum_{k=1}^{m} \frac{count\left(F_{k \in F} \cap FTrue_{k \in FTrue}\right)}{count\left(F_{k \in F}\right) + count\left(FTrue_{k \in FTrue}\right) - count\left(F_{k \in F} \cap FTrue_{k \in FTrue}\right)}$$

各项符号的含义：

$count()$：计算集合长度的函数

N：标准答案集的长度

$FTrue_{k \in FTrue}$：标准答案中，第k个域名对应的家族集合

m：选手提交的域名列表长度

$F_{k \in F}$：选手提交的答案中，第k个域名对应的家族集合

赛题设置（二）

- 举个栗子
  - 有[1-6]共计6个域名。标准答案为两个家族：{1,2,3,4}，{5,6}
  - 参赛选手提交了三个家族聚类，{1,2}, {3,4}, {5,6}
- 评分步骤
  - 第一部分：全部正确，满分
  - 第二部分
    - 域名1得分为：$\frac{2}{2+4-2}=0.5$，域名2，3，4与域名1相同
    - 域名5得分为：$\frac{2}{2+2-2}=1$，域名6与域名5相同
    - 公式累加：$\frac{1}{6}*(0.5*4+1*2)=0.667$

赛题一思路

- **题目：DNS基础安全知识**
- **阿里云安全团队的Writeup**
  - 题中说明存在五种攻击方式，且提交的是DNS query的packet id，表明出题人自信已经100%吃透了这1kw数据包。因此本次五种攻击模式不会太复杂。
  - 每种攻击流量都是“干净”的 (可以用规则搞定答案全集，不存在模棱两可、特征模糊、人工难辨的情况)，猜测攻击包有可能是出题人自己造的。
  - eth层、frame层、UDP层、TCP层的特征高度统一，出题人没有留下漏洞，因此重点分析DNS层即可。

赛题一思路

- **题目：DNS基础安全知识**
- **特征工程**
  - 强烈依赖于对DNS攻击的专家知识
  - 常见和不太常见的DNS攻击方法都需要知道

三类DNS攻击

- 密集请求型：随机子域名DDoS、反射型DDoS。其特征为QPS高、时序特征强，一般能够可视化观察到波峰。
- 漏洞攻击型：针对DNS server的已知漏洞攻击。其特征为数量少、受DNS type影响，适合分类统计。如果批量PoC的话，则特征同1。
- 数据传输型：DNS Tunnel、Malware DGA、PoC中的DNS回显、SSRF重绑定等。其特征在于域名文本特征明显、适用于规则匹配。

IP维度：

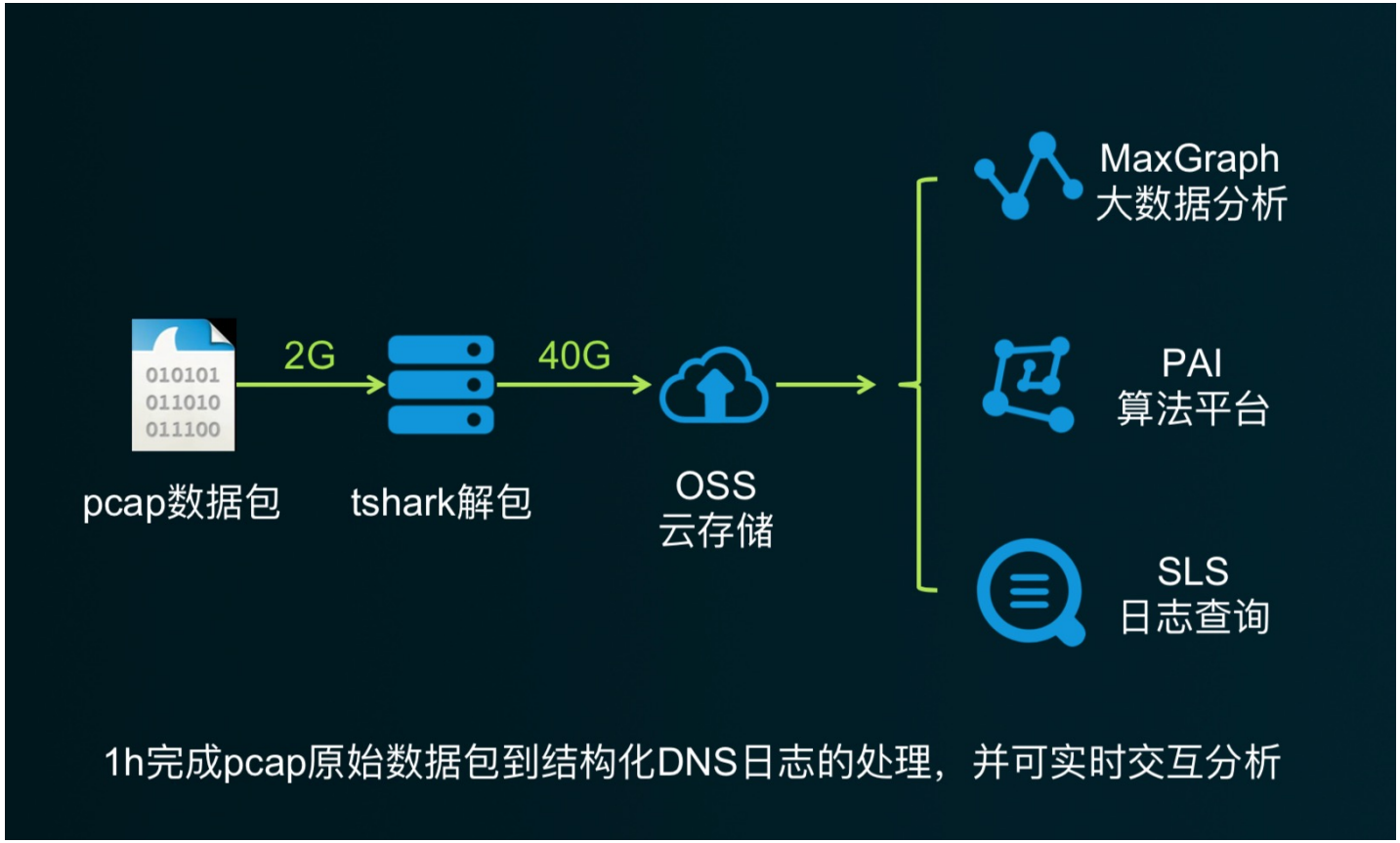
- DNS请求时序分布
- QPS min/max/avg
- QPS均值
- QPS波动性
- 连接成功率
- DNS响应率
- TCP报文占比
- 请求响应比
- 单域名平均访问次数
- 单目标高频访问
- 多级子域名变化率

DNS请求维度：

- DNS type时序分布
- DNS type源IP分布
- 长随机域名
- DNS Tunnel特征
- 部分DNS RCE
- 心跳包

赛题一思路

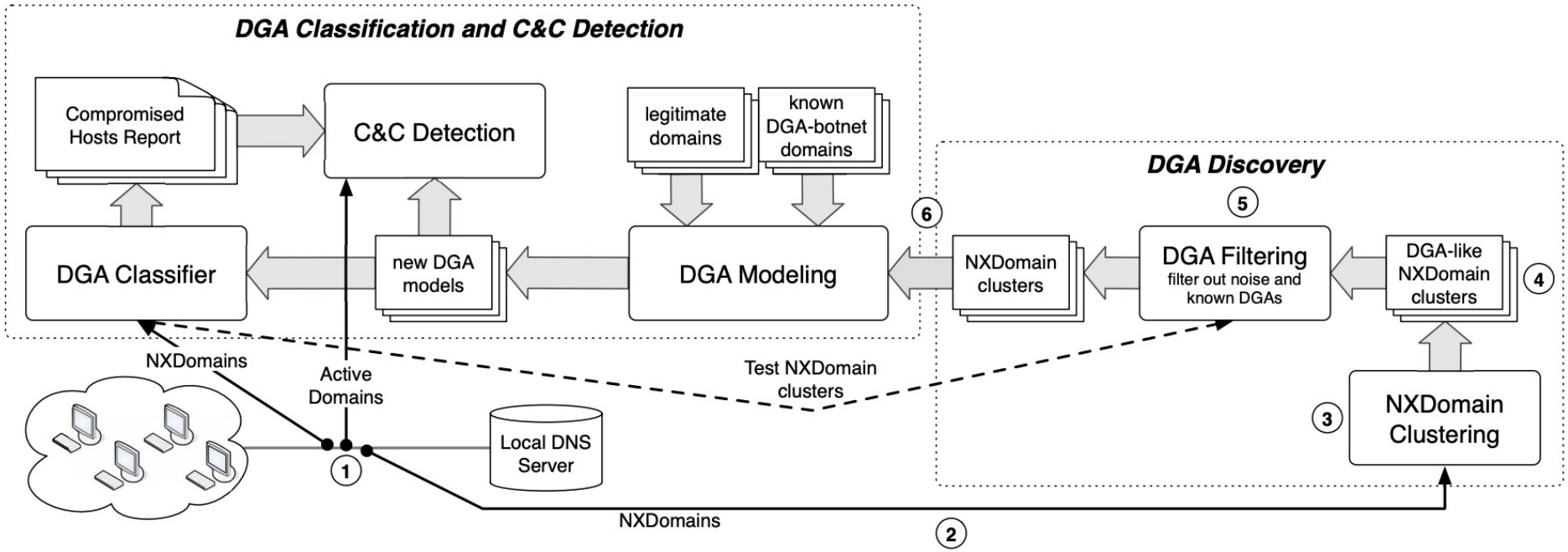
- 看了想打人的数据分析平台



序号	alive_seconds	qps_avg	qps_max	src_ip	d4_cnt
1	120	284.95	323	144.202.64.226	15382
2	1100	1.2081818181818182	15	52.73.105.200	587
3	1149	1.2297650130548303	18	112.198.179.24	581
4	3119	2.395639628085925	18	245.165.143.164	213
5	3483	1.5805340223944875	16	119.38.217.113	184
6	3477	1.6163359217716422	14	128.2.139.83	165
7	7798	2.154398563734291	12	31.187.146.83	156
8	8182	2.14421901735517	13	214.180.95.136	140
9	1582	2.2844500632111253	14	29.213.60.77	137
10	1881	1.1743753322700692	12	45.41.251.30	107
11	122	5.040983606557377	80	25.44.195.161	102

赛题二思路

- 题目：DGA域名家族聚类
- 学术界中一系列的常规操作
  - [USENIX 12, Manos Antonakakis]
  - 在参赛队伍中看理想与现实



赛题二思路

- **DGA Discovery**
  - **NXDomain Clustering**

    相似度标准：1）字符串统计维度；2）感染IP地址维度
  - **DGA Filtering**

    无监督聚类，噪音；已识别的DGA家族，热门流行域名
- **DGA Classification**
  - **DGA Modeling 将聚类结果进行多分类打标签**
  - **GroundTruth: 360 Netlab**

赛题二思路

- **阿里云团队：DGA域名发现 [二分类]**
  - [注：开始时并没有考虑时间特征，他们仅考虑了文本特征]
  - 特征维度

colname	feature_importance
dns_count_labels_avg	7.098254584511219e-7
dns_resp_name_len	0.486851015925180 43
dns_flags_rcode_min	0.09380669054958125
dns_qry_name_len	0.08286261820522951
dns_qry_name_consecutive_consonant	0.05165039925351584
dns_qry_name_count_digits	0.046638077601526357
dns_resp_type_min	0.03279234892125366
dns_qry_name_markov_p	0.03226292786968624
dns_flags_rcode_avg	0.030525270517365 04
dns_flags_recavail_min	0.0280585501594973 47
dns_qry_name_entropy	0.024847542112054 59
dns_flags_recavail_avg	0.016294232379689 43
dns_count_auth_rr_min	0.015600932601233563
dns_qry_name_digits_rate	0.009893560714853046
src_ip_cn	0.008147846032435051

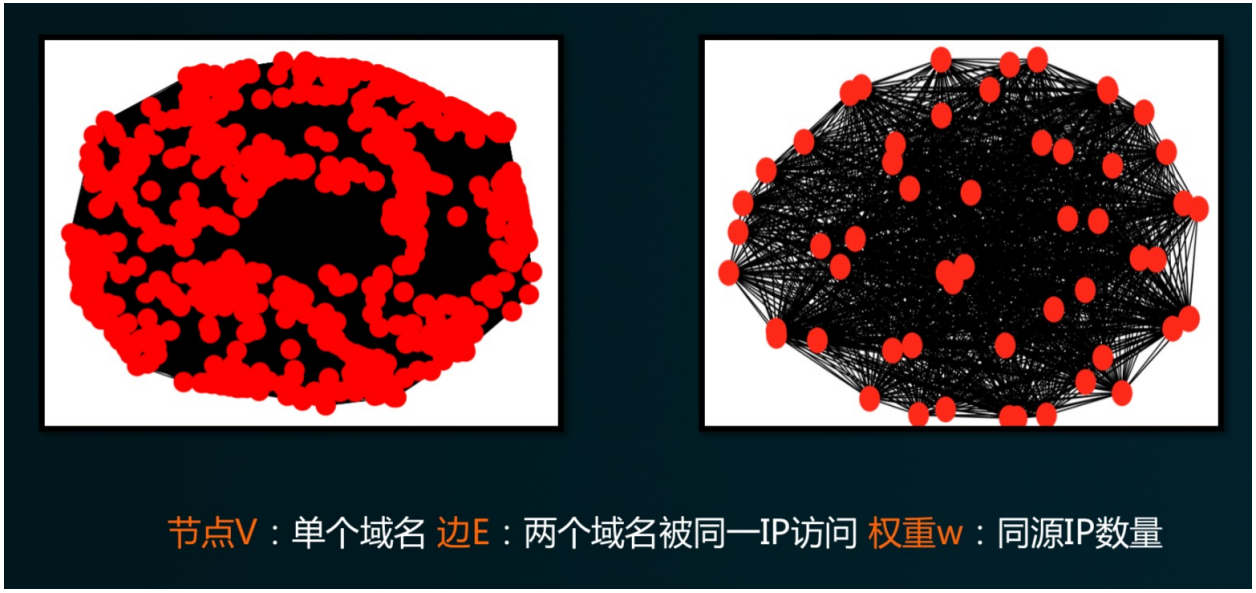
赛题二思路

- **DGA域名发现 [二分类]**
  - 遇到瓶颈：只能达到90%左右的recall预估
  - 最终调整：从感染主机的时间窗口视角出发，逐步将recall优化近100%

- **DGA家族聚类 [多分类]**
  - 社区发现的前提：域名降噪
  - 假设：同一个DGA家族的域名会被同一批源IP地址访问
  - 社区发现：定义图节点与边，源IP地址为图节点，共同访问同一个域名则建立一条边

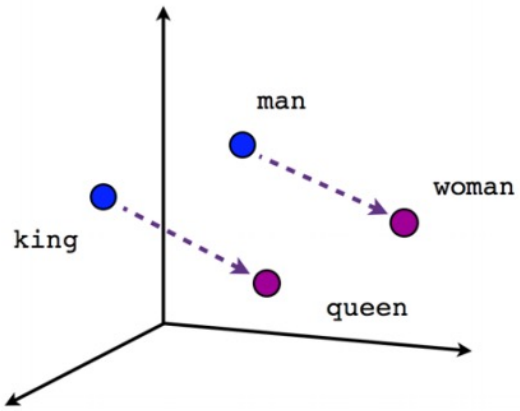
赛题二思路

- 社区合并
  - 初步得到54个家族，然后采取手工聚合
  - 社区合并特征包括
    - TLD分布 [biz info]
    - DNS SLD字符集 [a-zA-Z0-9]
    - DNS SLD子域名长度变化区间 [16, 32]

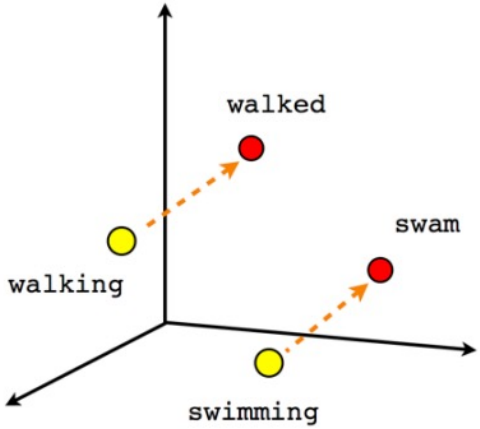


DGA域名家族聚类

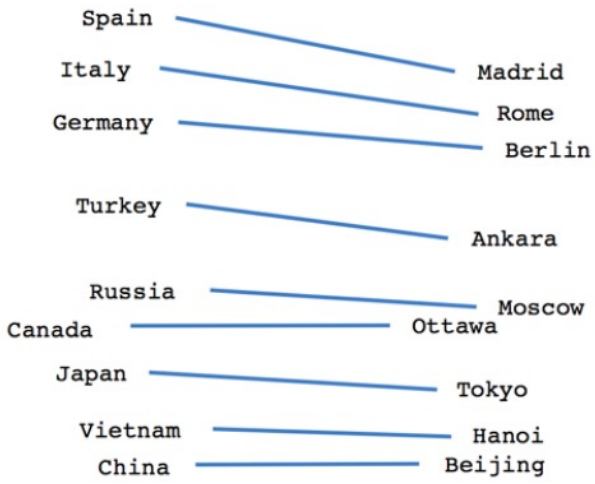
- 当然，我必然不是来说一些常规操作的人
  - 重磅干货：Netlab 梁锦津博士出品，Domain2Vec
  - 基本思想：单词与单词之间存在语义距离，域名之间也可能存在；根据DNS流量特征，DGA域名存在强伴生关系；



Male-Female



Verb tense



Country-Capital

DGA域名家族聚类

- **Talk is cheap, show me your code.**
  - 已知：ntp.gtpnet.ir与僵尸网络相关
  - 出发点：哪些域名与ntp.gtpnet.ir相关？

d2v

date	domain	ratio	count	rtype	source
20170508	ntp.gtpnet.ir	1.000000	144	1:0,1:2	d2v
20170508	dl.gtpnet.ir	0.710912	7	1:0	d2v
20170508	load.gtpnet.ir	0.641602	7	1:0	d2v
20170508	ntp2.eye4.cn	0.608029	2737	1:2,5:0	d2v
20170508	ntp.eye4.cn	0.600348	1199	5:0	d2v
20170508	s3.vstarcam.com	0.578202	12357	1:0,1:2	d2v
20170508	s2.eye4.cn	0.572022	11746	1:0,1:2	d2v
20170508	t.beibei.com.cn	0.515862	16	5:0	d2v
20170508	lanmai.xiaowm.com	0.502406	17	1:0	d2v
20170508	a2.zz06.net	0.502362	160	1:0,1:2	d2v

其它

- 明年的题，该怎么办？

- 我们可以一起愉快的玩耍吗？