**中图分类号：TP311.5**

**论文编号：10006ZY1506221**



硕 士 学 位 论 文

**恶意域名检测技术研究**

作者姓名 王文博

学科专业 计算机技术

指导教师 兰雨晴 周渊

培养院系 计算机学院

**Research on malicious domain name detection technology**

A Dissertation Submitted for the Degree of Master

**Candidate：Wenbo Wang**

**Supervisor：Yuqing Lan**

School of Computer Science and Engineering

Beihang University, Beijing, China

**中图分类号：TP311.5**

**论文编号：10006ZY1506221**

硕 士 学 位 论 文

**恶意域名检测技术研究**

* 利用PDNS检测恶意域名

作者姓名 王文博 申请学位级别 工程硕士

指导教师姓名 兰雨晴 职 称 教授

学科专业 计算机技术 研究方向 网络空间安全

学习时间自 2015 年 09 月 01 日 起至 2018 年 03 月 04 日止

论文提交日期 2018 年 03 月 07 日 论文答辩日期 2018 年 03 月 04 日

学位授予单位 北京航空航天大学 学位授予日期 2018 年 月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名：        日期： 年 月 日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

# 摘 要

这是摘要。

关键字：a；b；b；d

# **Abstract**

This is abstract.

**Key words:** a; b; b; d

目 录

[第一章 绪论 1](#_Toc503784159)

[1.1 研究背景 1](#_Toc503784160)

[1.1.1 恶意域名数量巨大 1](#_Toc503784161)

[1.1.2 恶意域名危害增大 1](#_Toc503784162)

[1.2 问题的提出 4](#_Toc503784163)

[1.3 论文的主要内容 5](#_Toc503784164)

[1.4 论文的组织结构 5](#_Toc503784165)

[第二章 国内外研究现状分析 7](#_Toc503784166)

[2.1 信誉系统相关研究 7](#_Toc503784167)

[2.2 利用机器学习的恶意域名检测 8](#_Toc503784168)

[2.3 恶意域名特征研究 14](#_Toc503784169)

[2.3.1 针对DGA的相关研究 14](#_Toc503784170)

[2.3.2 针对色情域名的研究 17](#_Toc503784171)

[2.4 威胁情报平台相关应用 17](#_Toc503784172)

[2.5 本章小结 19](#_Toc503784173)

[第三章 利用PDNS检测恶意域名的算法研究 20](#_Toc503784174)

[3.1 PDNS数据介绍 20](#_Toc503784175)

[3.2 恶意域名特征介绍 20](#_Toc503784176)

[3.2.1 域名字符特征 20](#_Toc503784177)

[3.2.2 域名访问特征 20](#_Toc503784178)

[3.3 快速提取恶意域名相关流量的方案 20](#_Toc503784179)

[3.3.1 针对DNS放大攻击相关域名的提取 20](#_Toc503784180)

[3.3.2 针对随机子域名相关域名的提取 20](#_Toc503784181)

[3.3.3 针对DGA域名的提取 21](#_Toc503784182)

[3.4 恶意域名分类算法方案 21](#_Toc503784183)

[3.4.1 DGA域名检测算法 21](#_Toc503784184)

[3.4.2 色情域名检测算法 21](#_Toc503784185)

[3.5 本章小结 21](#_Toc503784186)

[第四章 恶意域名检测应用原型系统设计与实现 22](#_Toc503784187)

[4.1 需求分析 22](#_Toc503784188)

[4.1.1 PDNS预处理评分模型 22](#_Toc503784189)

[4.1.2 域名特征提取模块 22](#_Toc503784190)

[4.1.2 恶意域名分类模型模块 22](#_Toc503784191)

[4.2 总体设计 22](#_Toc503784192)

[4.3 功能实现 22](#_Toc503784193)

[4.3.1 特征提取 22](#_Toc503784194)

[4.3.2 参数选择 22](#_Toc503784195)

[4.3.3 流量监控 22](#_Toc503784196)

[4.3.4 恶意域名检测 22](#_Toc503784197)

[4.4 本章小结 22](#_Toc503784198)

[第五章 实验结果与分析 23](#_Toc503784199)

[5.1 总体情况 23](#_Toc503784200)

[5.2 参数 23](#_Toc503784201)

[5.3 恶意域名检测结果 23](#_Toc503784202)

[5.3 本章小结 23](#_Toc503784203)

[总结与展望 24](#_Toc503784204)

[研究工作总结 24](#_Toc503784205)

[未来工作展望 24](#_Toc503784206)

[参考文献 25](#_Toc503784207)

[攻读硕士学位期间取得的学术成果 27](#_Toc503784208)

[致谢 28](#_Toc503784209)

图 目

[图1 test](#_Toc475557138) 1

[图2 test](#_Toc475557138) 1

表 目

[表1 test 1](#_Toc475386069)

[表2 test 2](#_Toc475386070)

# 第一章 绪论

## 1.1 研究背景

### 1.1.1 恶意域名数量巨大

IP地址是由IP协议提供的数字型统一地址标识，作为一种逻辑地址来定义一台设备在网络之中的位置，网络设备逐渐增多IP地址的记忆困难显现出来，保罗·莫卡派乔斯 (Paul Mockapetris) 在1983年的第882和在南加州大学里资讯科学研究院所提出的883号因特网标准草案中提出DNS的架构，提议将其改进为分布式和动态的数据库域名系统，也就是我们今天所用的域名系统的雏形。从1985年Symbolics公司注册的第一个.com域名到如今仅中国域名总数增长至3698万个[1]，域名产业飞速发展，随之而来的安全问题也越来越多的暴露出来。高级持续性威胁常态化，移动互联网黑色产业链已经成熟，巨大的利益促使大量人进行相关活动。以僵尸网络控制端（通常使用域名来进行联系）为例，2012年木马和僵尸网络控制端数量高达36万余个，随着检测技术和安全意识的提高，至2014年数量有明显降低，仍有10万个左右的僵尸网络控制端，并且该数量稳定至今[3]。

### 1.1.2 恶意域名危害增大

我们将当前存在恶意行为或者被恶意使用的域名都视为恶意域名，这其中就包含了DGA、DNS放大攻击、钓鱼域名等等。正如1.1.1节中所述，尽管随着互联网安全监管的加强，恶意域名数量仍然庞大，并且恶意域名相关的技术在与检测技术的对抗之中不断进步，造成的安全威胁更加巨大。2017年活跃的WannaCry、Mirai等恶意程序就是最好的佐证。

WannaCry是一种勒索木马，感染主机后利用微软系统漏洞EternalBlue（永恒之蓝），获取系统权限，将硬盘中的文件加密进行勒索。2017年5月第一波爆发，感染了全球一百多个国家和地区超过三十万台主机，造成经济损失达百亿美元。根据对样本的观察以及之后研究人员对WannaCry源码的分析，我们了解到WannaCry是否在感染主机后进行文件的加密处理与域名“*www.iuqerfsodp9ifjaposdfjhgosurijfaewrwergwea.com*”（该域名为第一次攻击时使用的域名，之后使用其他域名）休戚相关，WannaCry会首先尝试访问该域名，如果无法被正常访问则会继续产生恶意行为。这和以往的恶意软件很不一样，大多数恶意程序会通过域名或其他手段与攻击者产生连接，根据攻击者指令产生攻击行为，而WannaCry恰恰相反，域名的成功连接作为攻击停止的信号来使用。这里的域名就属于恶意域名的范畴之中。

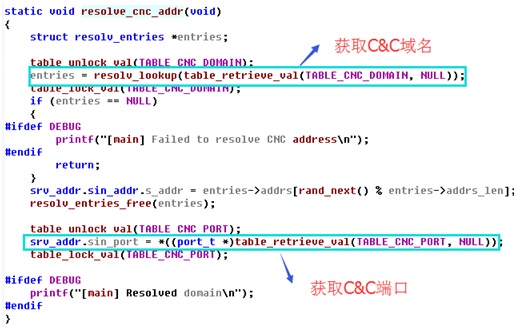


图1-1

Mirai是一类针对物联网设备的僵尸网络木马，可以由一个感染设备对其他可感染设备进行SYN扫描与探测，扩散能力极强。Mirai所感染的肉鸡通过域名与攻击者所控制的C&C服务器产生连接。如图1-1所示，最初的Mirai使用的域名还不是自动生成的，而是简单存在一个列表“*TABLE\_CNC\_DOMAIN*”中。被感染设备会遍历这个域名列表，攻击者只需要选择其中一个域名注册使用，即可控制所有的被感染设备。在2016年10月美国爆发的大规模网络瘫痪事件中，黑客正是利用Mirai控制大量的物联网设备针对Dyn域名服务器发起的DDOS攻击。如今物联网设备或者家庭智能设备数量不断的增长，黑客也将注意力更多的放在了这些设备上。在之后出现的诸多Mirai变种中，使用了DGA（域名生成算法）代替了原本的固定域名列表，那么使用DGA有什么好处呢。

传统僵尸网络使用固定IP或者域名与C&C服务器建立连接，隐蔽性很差，极易被发现。后来出现的例如Nugache，Storm，Waledac，Zeus等基于P2P的僵尸网络虽然具有较好的鲁棒性与稳定性，但实现难度和维护成本较高。如今大部分活跃的僵尸网络都采用了DGA，依赖于集中的C&C服务器，相比于前两者具有简单易行，兼顾稳定性与隐蔽性的优点。诸如Locky[4]、GameOver、Rovnix等新型木马均使用了DGA来获取与C&C服务器连接的域名。DGA作为一种随机算法，输出为域名，我们将算法输入称为种子（例如数字常量、当前时间、Tiwwter动态等），按照种子产生类型以及算法类型，可以对DGA进行分类。如果这个种子与时间有关，称之为TD（time-dependent，时间相关），反之，称为TI（time-independent，时间无关），如果种子可以估计（例如日期），称为D（Determinism，可估计），反之（例如欧洲央行每天外汇参考利率），称为N（Non-determinism，不可估计）。域名产生模式分别有A（Arithmetic，算法类）、W（Wordlist，单词表类）、H（Hashing，哈希类）、P（Permutation，置换类）四种。进行排列组合理论上有16类DGA，实际上只出现了TDD-A, TID-A,TDD-W , TDD-H, TDN-A , TID-P 这六种类型[5]。可见DGA算法千变万化，是黑客的一大利器。

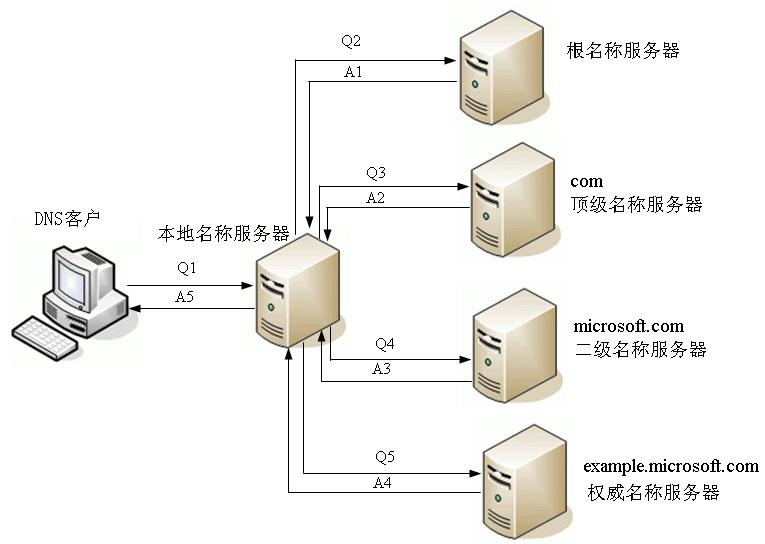


图1-2

上文介绍了僵尸网络在建立和连接上如何使用了域名，攻击者利用这些僵尸网络可以很轻松的发起DDos攻击（Distributed denial-of-service，分布式拒绝服务攻击）。在2015年，DDoS攻击峰值流量不断上升，甚至出现了1T超大流量攻击事件，全年的攻击总流量接近28万Tbytes[6] ，常用手法有UDP攻击、SYN攻击、ICMP包攻击等。在各种类型的DDos攻击中，有两种涉及到了恶意域名，一是利用DNS解析的反射放大攻击，在利用流量来实施的DDoS攻击中，只要有可以利用来进行放大数据量的协议，都有可能被恶意使用。正常的DNS查询是从原IP地址向DNS服务器（递归或者权威），大小相对固定（70字节左右），也就是攻击成本相对固定，DNS返回数据由于请求域名和类型的不同，数据大小从几百到几千字节（查询“*www.baidu.com.*”的返回数据为302字节），放大了四倍以上。由于利用A记录或者Cname记录放大效果并不十分明显，Akamai研究人员在2014年发现了利用TXT记录来进行的DNS放大攻击，攻击者使用名为DNS Flooder的工具，从guessinfosys.com获得TXT记录，攻击峰值高达4.3Gbps[7]。二是随机子域名攻击，这是一种专门针对域名服务器的DDos攻击。例如攻击者的利用的域名为“*example.cn*”，那么攻击者会利用僵尸网络中的主机产生一系列无意义的域名头部，与目标域名组合之后形成诸如“*iygyaid.example.cn*”，“*aduh3uh.example.cn*”之类的域名，每一台主机都会对这些域名进行DNS请求，如图1-2所示，由于二级域名都是合法的，导致在域名的递归查询过程中，这些DNS都会进行相同的二级域名服务器直到权威域名服务器，在僵尸网络足够庞大的情况下，这些递归DNS服务器就会因为处理这些NXDomain（Non-exist Domain，不存在域名）而陷入瘫痪。

恶意域名中钓鱼域名也是不可忽视的一部分，钓鱼网站（phishing page）一直以来都是一个难以规避的网络问题，同时他也是一种社会工程学的攻击方式。通常攻击者对恶意页面进行伪装，使得页面看起来像是银行网站或是其他可以获取关键隐私信息的网站，受害者往往由于受到邮件或者短信的欺骗，通过其中的链接导向到钓鱼网站，或者由于误植域名导向到钓鱼网站。而受害者对于网站的真伪缺少足够的判断和辨识的能力，导致银行账户密码等隐私信息被攻击者窃取。还有一个需要提及的是色情域名，由于文化以及法律法规的差异，各国对色情域名的定义大相径庭，在这里我们将他视为一类恶意域名来研究。由于网络的普及以及缺乏对网络内容的合理分级，导致网络中大量非法的负能量的内容可以轻易的被不适宜的人群所接触，尤其是青少年人群对一些内容的辨识度不够，很容易误入歧途，造成不良的社会影响。除了以上提及的恶意域名种类之外，还包含有垃圾邮件域名、域名阴影的域名等等，每一类域名都不断延伸成一个完整的产业链，对互联网环境都有着巨大威胁。

## 1.2 问题的提出

各类恶意域名带给互联网的安全威胁巨大，那么一个行之有效的方法来检测和识别出这些恶意域名就很有必要了，现阶段针对恶意域名的检测技术主要面向三个方向。一是域名信誉系统（DRS，Domain Reputation System），简要来说DRS就是一个给域名打分的系统，域名的良性与恶性都由分数来决定，我们熟知的黑名单也可以看作是一种最简单的信誉系统，黑名单上的域名的分数统一为0分。建立黑名单是一种切实有效的防御手段，国内外很多安全厂商和团队都长期维护着黑名单，可以说这是误报率最低同时也是检测速度最快的一种防御措施。但黑名单也有着不可避免的缺点，将数据信息进行人工或者程序的审核，确认为恶意的数据信息加入黑名单，因此维护和更新的成本相对较高，而且黑名单的建立必然在恶意行为发生之后，相对滞后。

二是逆向工程，逆向工程对于恶意软件防御方面而言是一种很常规的手段，Khaled Yakdan等人的论文中正是对现在活跃的大部分DGA使用了逆向工程[8]，在实际使用中可以做到FP值为0，他们历时数年的细致工作完美体现了逆向工程精准的特质，但是逆向工程的缺点也暴露无遗，那就是太耗费人力和时间，一旦出现可能是新的DGA算法，就需要专业的逆向工程师来处理、验证，这对于如今层出不穷的新型木马而言是远远不够的。

三是机器学习，这其中也包含了深度学习的部分。如今信息安全领域仍然是一个富数据，穷分析的领域。2016乌云白帽大会上，phunter做了报告《What can you get from 100 billion DNS queries, each day, in real time?》，可以想见当我们有1000亿条实时数据，如果仍然使用蜜罐、逆向工程，周期长度以及工程量都是无法承受的。如何做到又快又精确地对给定域名进行分析，那么机器学习就是不二之选。特征的选择体现了人类的经验，而把这些量化特征交给机器，相当于机器利用人类的经验来完成这些重复性的工作。

有幸在研究生学习期间在国家互联网应急中心实习，获得了中心提供的PDNS（Passive DNS，被动DNS）数据，该数据为广东省电信部门的DNS解析数据，数据具体详情见 节。由上述检测技术可见无论哪一种方法都存在着一定缺陷，本论文希望能够探讨如何利用PDNS数据的高效快速并且尽可能准确的检测出相关的恶意域名。综上所述，利用PDNS的恶意域名检测技术有如下几个问题需要解决：

1. 基于PDNS数据可以检测出哪几类恶意域名；
2. 采用怎样的算法来进行检测；
3. 如何快速和准确的来检测恶意域名。

## 1.3 论文的主要内容

## 1.4 论文的组织结构

第一章：绪论，简单介绍了本文的研究背景和相关概念，提出了本文研究的问题。

第二章：

总结与展望：对本文进行总结，并对下一步的研究工作进行了展望。

# 第二章 国内外研究现状分析

本章主要对国内外学者在恶意域名检测技术的研究现状进行介绍和分析。对于恶意域名检测技术的研究主要分为逆向工程、信誉系统、机器学习三类。逆向工程具有误报率低、召回率高的优点，但是耗时耗力、对攻击反应不及时；信誉系统具有检测速度快、误报率低的优点，但是召回率较高，更新滞后；机器学习方法具有高效、便捷、实时性高的优点，相对前两者精度略有差距。在本章中我们主要关注信誉系统与机器学习。

## 2.1 信誉系统相关研究

佐治亚理工学院的M.Antonakakis、P.Roberto、W.Lee等人是最早一批深入研究PDNS数据的研究人员，他们先后对DNS缓存投毒、域名信誉系统、DGA、僵尸网络等方向都做出了很高的贡献。在2010年建立的Notos[9]，是一个动态的综合性的信誉系统，首次提出了针对域名的信誉系统，利用被动DNS数据来输出信誉分数，而于此之前主要针对IP的信誉系统完全不同。Notos使用了网络和地区的特征，通过配置信息、使用情况和域名管理情况等，能够学习良性和恶性的域名分别是如何工作的，并对每一个新域名计算一个信誉得分。对于一个域名，如果它与恶意活动(例如僵尸网络、垃圾邮件、恶意软件传播等等)有关，就给他赋予一个低的信誉值，图2-1是Notos系统的整体流程。

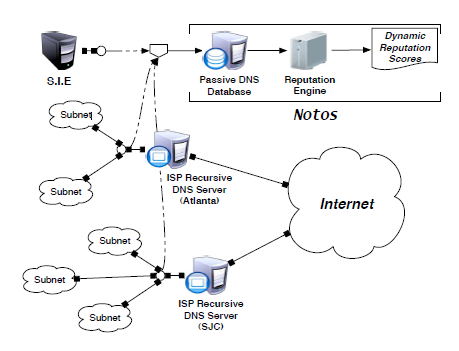


图2-1

首先利用被动DNS数据构建三组特征：

1. 基于网络的特征：第一组的统计特征是从RHIPs的集合中提取的。计算如下特征：与d历史相关的IP地址的总数量、它地理位置的多样性、它们所在的不同的自治系统的数量等等。

2. 基于域的特征：第二组特征是从RHDNs集合提取的。计算在RHDNs里域名的平均长度、不同TLD的个数、不同字符出现的频率等等。

3. 基于证据的特征：最后一组特征包括与d相关的不同恶意软件样本的数量、与d曾经指向的IP相关的恶意软件样本的数量等等。

Notos的信誉引擎有两个运行模式：离线训练模式和在线分类模式。在离线模式下，Notos使用在知识库中收集到的数据来训练信誉引擎，知识库即一组已知的恶意和合法的域名及其相关的IP地址。在之后的在线模式中，对于每一个新域名d，Notos询问信誉引擎从而为d计算一个信誉值。动态信誉评分系统的基础是从成功解析的域名A记录中获得的历史或“被动的”信息。作者使用了来自两个ISP的采集节点的DNS流量，一个位于美国东海岸(亚特兰大)，另一个位于美国西海岸(圣荷西)。另外他们也汇总了SIE项目中不同网络的DNS流量。数据库收集到了从2009年7月19号至2009年9月24号这68天的解析记录共27377461条。结果Notos可以以96.8%的精度和0.38%的误报率发现恶意域名，并且它比黑名单方法更快。

## 2.2 利用机器学习的恶意域名检测

2011年Manos Antonakakis等人构建了一个叫做Kopis的恶意域名监测系统[10]，可以通过对DNS查询解析模式的分析得到恶意域名，与EXPOSURE和Notos这些依赖本地递归DNS服务器的系统相比，Kopis使用的是上级DNS数据。在此基础上他们在2012年又提出了一个新颖的检测系统Pleiades[11]，首先使用了不存在域名（NXDOMAIN）响应。NXDOMAIN是指像域名服务器提出解析请求，但是无法解析得到对应IP。对于使用DGA的被控主机大多数查询都会返回一个NXDomain，而同一僵尸网络控制下(使用相同的DGA算法)的被控主机会产生有类似特征的NXDomain的流量。Pleiades聚类和分类算法的结合，充分利用了机器学习的优势，将有着相似字符特征和相似访问特征的域名聚集在一起。分类算法用来将这些生成好的聚簇分配到已知的DGA模型中去。如果一个聚簇不能被分到已知的DGA中去，那么意味着可能出现了新的DGA变种或DGA家族。

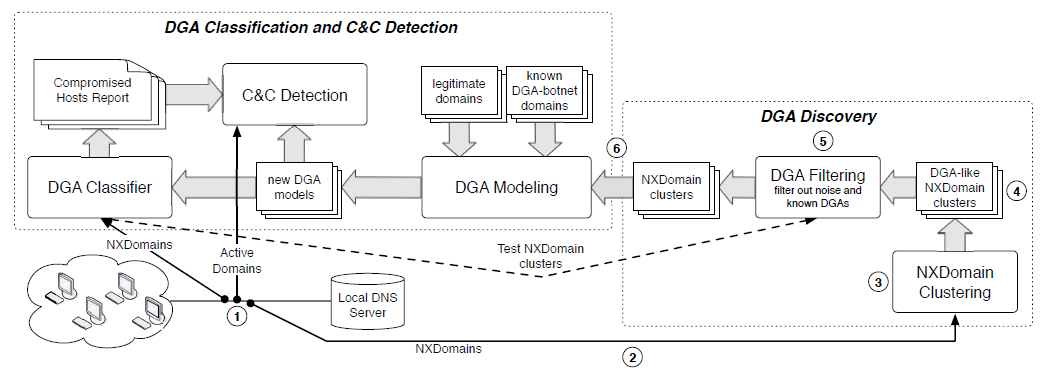
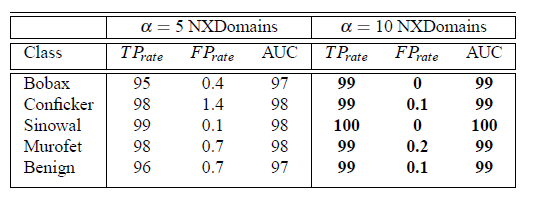
图2-2

图2-2是Pleiades的整体流程。DGA发现模块分析了未成功的DNS解析流量，它部署在DNS服务器的下游。在一定时间段内网络产生的所有NXDomains都会被收集在内。接着，收集到的NXDomains会根据以下两个相似准则分别被聚类：(1)域名字符串所具有的相似的统计特征(例如相似的长度、随机性和有关字符的频率分布等)；(2)域名被一组相同或相近的IP访问。对NXDomain聚类的主要目的是将那些使用同一种域名生成算法的主机和域名聚类出来。由于这个聚类过程是无监督的，所以聚类出来的某些NXDomain簇里可能包含那些偶然错误的域名(例如由于拼写或者配置错误导致的NXDomains)。因此需要对这些数据进行过滤。作者使用有监督的DGA分类器去修整这些聚类簇，修正的对象包括已经发现并建模的DGA产生的域名和那些与合法的域名相近的域名。DGA发现模块最终的输出是NXDomains的聚簇集，其中的每一个集合都代表了已知或者未知DGA产生的域名。每当发现一个新的DGA，就使用监督学习方法为这样的域名建立一个模型，该模型用来描述新DGA产生的域名“长什么样”。为此建立了两种不同的统计模型：(1)一个为肉鸡 产生的一组NXDomains赋予DGA标签(如DGA-Conficker.C)的多类统计分类器；(2)基于隐马尔科夫模型判断 查询过的可能由一个DGA产生的活跃的单一C&C域名的类型。DGA分类模块的工作过程如下。与DGA发现模块相似，作者监测DNS递归服务器中由每个主机产生的NXDomains流量。给定一个主机产生的NXDomains的子集，然后提取与其字符串有关的一系列统计特征。接着将这些特征作为特征向量传入分类器函数，分类器会输出这些NXDomains是否由一个已知的DGA产生。如果一个主机产生的NXDomains被贴上了相关的DGA标签，那么这个主机也极有可能被相关的僵尸网络所控制。一旦获得了那些感染主机的列表，就可以进行更深一步的检测。之前所有步骤都是围绕无效域名NXDomains展开的，下一步则把注意力集中到这些感染主机访问的活跃域名上。其目标是确定哪些由DGA算法生成的域名最后解析成了IP地址，即识别僵尸网络的C&C服务器。为了实现这一目标，作者将感染主机访问过的所有可疑活跃域名都收集起来。接着对这些域名输入之前训练好的隐马尔科夫模型中，由它来决定单一活跃域名是由已知DGA产生的或是未知DGA产生的。使用隐马尔科夫模型而不是分类器做判断的原因是需要对单一域名进行检测。DGA分类器不适合单一域名的检测，因为它根据一个感染主机产生的一组无效域名进而判断这组无效域名是哪种DGA产生的。

作者使用Bobax, Sinowal, Conficker-A, Conficker-B,Conficker-C和Murofet产生的NXDomains来引导分类器。在两种模式中测试分类器。第一种模式使用Conficker的一个“超类”，它由Conficker-A, Conficker-B和Conficker-C的相同数目的样本组成。另一种模式则将Conficker的每个变种作为不同的类别看待。从每个类的域名中，作者随机选取了大小为α的3000个集合。其使用的α值有2、5、10和30。这是建立不同的训练集的过程，目的就是通过实际证明哪个α值可以给出DGA模型之间的最佳的分割。

表2-1



Pleiades从2010年11月的第一天就开始对NXDomains流量进行聚类。作者使用已知的DGA作为正例和一组Alexa域名作为反例来引导DGA建模过程。通过检查NXDomains与从恶意程序库中提取到的NXDomains的重叠关系来发现恶意程序家族。另外他们也在威胁情报公司的帮助下手动地检查了聚簇，检测出每个DGA变种有着平均32个感染主机，这些主机横跨全州的ISP网络。通过十几个月的实验，我们证明了Pleiades可以达到一个很高的精确度。此外，Pleiades部署在大型ISP网络下的这15个月来，它可以发现6个属于已知恶意程序家族的DGA和6个之前从未报出过的DGA。

Pleiades有一些局限。例如，一旦发现一个新的DGA，Pleiades可以很准确地为它建立统计模型，它可以知道这种DGA产生的域名“长什么样”，但它不能通过学习重现它的域名生成算法。因此，Pleiades会产生一定程度的假正和假负。但是上表的数据表明，Pleiades可以建立一个很准确的DGA分类器模型，当α=10时假正和假负都很低。C&C检测模块可以判断单一活跃域名的DGA种类，在大多数情况下表现很好。但是有一些情况下基于HMM的分类有些问题。作者认为这样的原因是HMM只考虑了域名的单一字符序列。总体来说，Pleiades聚类和分类算法的结合，充分利用了机器学习的优势，将有着相似字符特征和相似访问特征的域名聚集在一起，分类算法用来将这些生成好的聚簇分配到已知的DGA模型中去，是这个领域非常重要的一篇文章。

除开佐治亚理工学院的这些人，其他地区的研究者和组织也对这个领域做出了极大的贡献。Perdisci 等提出 FluxBuster 系统[12], 这是一个专门针对速变域名检测的系统，将域名的IP变迁情况引入特征集, 共9组13个特征, 采用聚类算法来识别速变域名。除了针对某一种域名的检测系统，更多的是具有广泛适用性的系统。

2011年L.Bilge等人建立了EXPOSURE系统[13]，一个检测多种恶意域名的系统。该系统从DNS数据中分别基于时间、基于DNS响应、基于生存时间值（TTL）、基于域名提取了这4类共计15种特征，使用J48决策树训练分类器。相比于Notos，Exposure有着更完善的特征选取，也弥补了无法检测一个IP地址只被恶意使用一次的恶意域名。相比于之前文章都只能在DNS数据中找到特定种类的恶意域名，EXPOSURE对恶意域名的检测更加广泛。

2013年Terry Nelms，Roberto Perdisci等人构建的ExecScent[14]是一个旨在从真实的企业网络流量中挖掘新的、从未出现过的C&C域名的系统，同时ExecScent也是第一个使用自适应C&C流量模型的系统。ExecScent从已知的C&C通信样本中自动地学习控制协议模板(CPT)，并且这些CPT会匹配它们所部署的网络流量。ExecScent构建的这种自适应模板从部署模板的网络流量中学习，这种“自适应”的方法使得ExecScent在保持一个很高检出率的同时极大地降低了误报率。ExecScent自动地寻找不同恶意样本的C&C协议间的共同的特点，接着将这些共性编码进CPT集合中。每个模板都标有恶意程序家族的名称或者与该C&C流量相关的犯罪组织(如果有的话)。一旦将CPT部署在网络的边缘，任何与该模板匹配的HTTP(S)流量就会被认为是C&C流量。与该流量相关的域名就会被标记为C&C域名，并且会归于恶意程序家族或者与该CPT关联的组织。图3-3展示了ExecScent生成和标记CPT的整体过程。

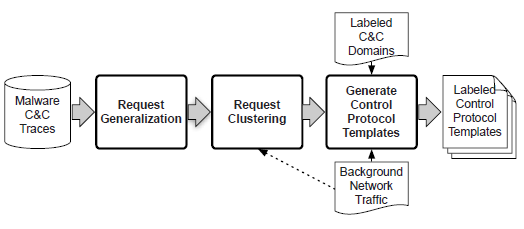


图3-3 ExecScent系统概述

给定一个大的恶意软件生成的网络痕迹，首先重构每个恶意样本的HTTP请求。接着，进行一个“请求泛化”的过程，这里将请求的一些参数(如URL参数值)替换为它们的类型和长度。当恶意程序的请求被聚类后，我们进行模板学习过程，在这里我们推导出CPT。本质上讲，一个CPT高度概括了聚在一起的HTTP请求，并且记录了一系列重要的特征如URL结构、请求头的集合、每个恶意软件访问过的IP地址等等。此外，这些模板与恶意程序家族标签一一关联。在真正部署模板前，作者先把CPT放在应该部署到的网络流量中观察一段时间。特别是对于模板的每一个组件(如泛化的URL路径、user-agent字符串、请求头集合等等)，计算各个组件在部署网络中出现的频率。在背景流量中“活跃”的那些组件会在该网络中得到一个低的“匹配信任”。另一方面，在流量中那些不经常(甚至是从来没)出现的组件会得到较高的信任。部署后，如果一个HTTP请求被CPT以很高的相似性和特异性所匹配，那么它就会被标记为C&C请求。这就是说，一个请求与CPT描述得很接近并且匹配的CPT的组件在特定的部署网络中有着很高的特异性(即不经常出现)。

2016年B.Rahbarinia创新性的提出了一种基于行为的系统Segugio[15]，Segugio在大的ISP网络中通过追踪被恶意软件感染主机的DNS请求行为来高效的发现新增的malware-control域名。相比于Notos和EXPOSURE，前两者建立的都是domain-IP映射关系模型（使用域名字符串的特征、域名承载的恶意内容等信息）而没有利用本地DNS服务器下游的主机请求行为，Segugio通过监测ISP网络用户的DNS请求行为，重在精确的追踪新增的“malware-only”域名。相比于Kopis，他的做法和本文有相似性（Kopis用请求者散度、请求者画像等信息），但Kopis利用权威或TLD服务器的数据，这种数据难以获得（需要与大的DNS区域运营商紧密合作）。Segugio不用关心顶级域名，通过监测本地ISP流量（在ISP使用者和他们的本地DNS解析器之间的DNS流量）。因此Segugio可以依靠ISP网络权限独立开发，不需要与外部的DNS运营商合作。

国家互联网计算机网络应急技术处理协调中心张雪松等人提出了对算法生成恶意域名的实时检测[17]。图3-5是算法的处理流程，主要包括下面几部分：数据采集及预处理、已知算法生成恶意域名的解析及访问关系构建、新增域名与已知域名的关联分析、未知域名的跟踪分析。数据采集与预处理主要是通过镜像DNS流量，实时过滤并提取新增域名，利用域名白名单机制降低待处理的数据量。系统从DNS流量中提取域名，保存到域名数据库，并通过一段时间的数据收集建立域名的检测基线，依靠基线提取新增域名作为系统分析对象；收集并整理已公布和确认的算法生成恶意域名，作为关联分析的基础。将已确认恶意域名按照域名结构及解析内容进行分组，在内存中建立域名分组与解析IP集合分组与访问客户端集合的对应关系，用于后续与新增域名间的关联分析；新增域名与已知域名分组进行关联后，如确定域名属于某一分组，输出并更新域名分组信息，否则对新增域名持续跟踪，收集域名的解析及访问数据，按照周期分析域名的活跃情况，判定域名是否为算法生成的恶意域名。

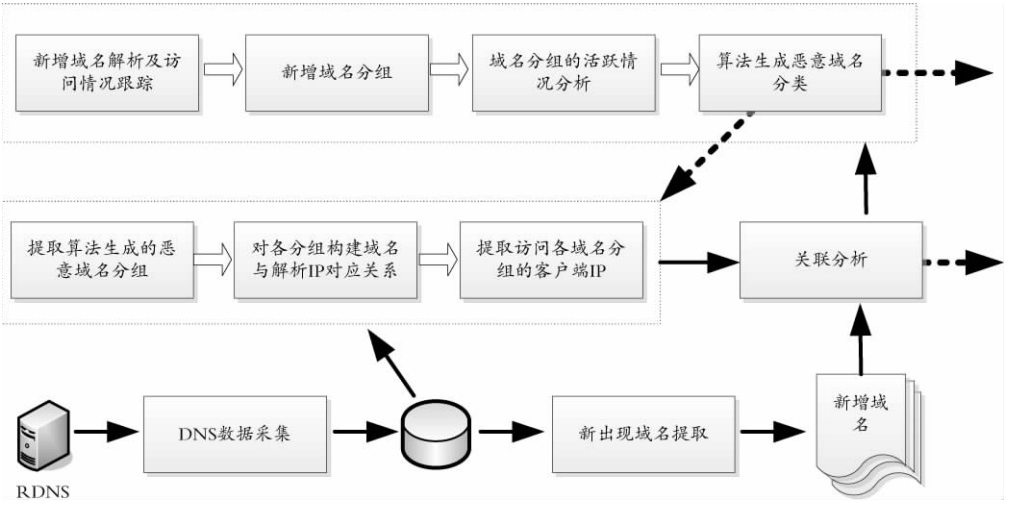


图3-5 域名实时检测流程

除了这些机构，一些高校对此也有所研究，东南大学张维维等人针对黑名单在维护和更新上存在开销大和及时性差，且攻击者常常使用算法自动生成大量的随机域名来躲避检测的不足以及实时检测开销过大的缺陷，设计了轻量级的检测算法来快速锁定监测目标，以便有针对性地使用更为复杂和更为准确的检测算法[18]。轻量级算法需要在有限的系统资源和计算时间内，尽可能多地检测出可疑域名因此，算法设计优先考虑空间开销和计算复杂度，而检测精度可以由更为复杂和更为准确的后续算法去保证。

文章依据最小的语言学单位—词素，设计启发式字符串切割算法快速切分域名，并在二层域名标签聚类的基础上，通过统计域名所含词素的比重、均长和频率分布熵等特征测度，应用有监督的机器学习方法检测恶意服务使用的非法域名。 为验证算法的可行性，作者基于统一的标准域名集，比较词素特征与已有n元组频率分布以及单词特征的检测能力。实验结果表明：词素特征能够有效地刻画域名字面的语言学特征，与n元组频率分布特征相比，可以提高检测准确率，降低假阳性，有效抵挡攻击者借助事前相应特征统计的逃避策略以及借助字典或Kwyjibo工具的随机域名生成策略；与单词特征相比，在保证检测准确率的同时，较小的词素集可以保证算法具有较低的计算复杂度和存储开销。最后，实际应用该轻量级算法对中国教育科研网江苏省网边界采集到的域名集进行检测，结果表明：该算法具有较高的检测准确率(87.2%)、较低的内存开销(80.14MB的临时内存,2.71MB的常驻内存开销) 和计算复杂度(运行时间196.1s)。

## 2.3 恶意域名特征研究

### 2.3.1 针对DGA的相关研究

D.Plohmann，F.Fkie等人针对DGA做了大量细致的工作[16]，他们对43种DGA恶意软件家族和变种进行一个综合性的研究，针对DGA提出一个分类学方法，并用它对所研究的DGA进行分类与比较。并重现了这些算法，预先计算所有可能的AGD，覆盖了大部分已知的活跃DGA，以过去八年总计一千八百万DGA域名的注册状态，来证实预计算得到的域名确实是是可靠的。对于botmaster的域名注册策略也提出了见解。

DGA作为一种随机算法，输出为域名，作者将算法输入称为种子（例如数字常量、当前时间、Tiwwter动态等），按照种子产生类型以及算法类型，可以对DGA进行分类。如果这个种子与时间有关，称之为TD（time-dependent，时间相关），反之，称为TI（time-independent，时间无关），如果种子可以估计（例如日期），称为D（Determinism，可估计），反之（例如欧洲央行每天外汇参考利率），称为N（Non-determinism，不可估计）。域名产生模式分别有A（Arithmetic，算法类）、W（Wordlist，单词表类）、H（Hashing，哈希类）、P（Permutation，置换类）四种。进行排列组合理论上有16类DGA，实际上只出现了TDD-A, TID-A,TDD-W , TDD-H, TDN-A , TID-P 这六种类型。

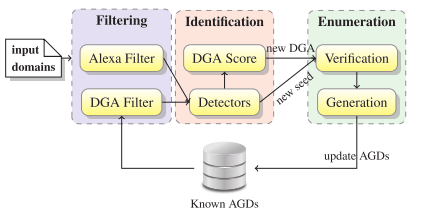


图3-4 DGA收集过程

图3-4为作者的DGA域名收集过程，实际工作中首先通过恶意软件分析报告和blog，定义了最初的22个DGA家族。然后利用Alexa top-ranked10000的域名和已知AGD（由DGA算法产生的域名）进行过滤。接下来是识别的过程，构建探测器用来检测一个给定域名是否符合已知DGA的输出，捕捉产生域名的最大最小长度，DGA字母表，已知TLD等字符特征。如果域名模式相同，标为使用新种子的已知DGA；如果不一致或者域名数量不足，通过计算一系列数值，例如n-gram，熵，长度等，来判定是否是新的DGA种类。当检测出来是一个可能的新DGA算法，人工进行逆向工程来验证。

国内也有很多机构活跃在恶意域名的检测领域，中国科学院信息工程研究所有大量与此相关的文章。张洋等人基于多元属性特征的恶意域名研究[28]针对的是域名域名的伪装特点和跳变特点，针对性从多个角度地提取一些特征进行恶意域名检测。 如“taobao.xxx.com”，可以从域名词法角度提取属性特征，也可以通过其网络属性来进一步丰富属性特征，如生存时间值(Time-To-Live，TTL)。表3-2与表3-3是文章所使用的11个词法特征和9个网络特征：

表3-2 词法特征

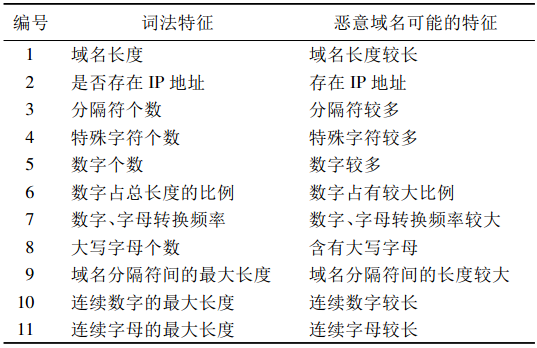


表3-3 网络特征



文章使用随机森林分类器对恶意域名和正常域名进行分类。随机森林分类器具有诸多优势: 能够处理高维数据，能够处理多种格式的数据，能够处理缺失的特征，能够平衡误差，不会出现过拟合的现象等。鉴于此，随机森林分类器非常适合对域名进行分类。随机森林分类器作为基于机器学习的一种集成学习方法，结合了多个决策树的分类效果，最终通过投票方式选出得票最多的类别作为最终的分类。实验结果表明，该方法的准确率、召回率99.8%，具有较好的检测效果。

张永斌等人认为相同周期内产生的僵尸网络域名大多具有相同字符特征，被感染主机有相同行为特征，提出了他们的一个检测算法[29]。整个检测算法分为：数据预处理、失效域名聚类分析、新域名聚类分析、恶意域名提取4个处理过程。其中数据处理过程主要是判断一个解析成功的域名是否是新域名，并将网络中主机请求的新域名及失效域名信息进行保存，以便后续分析。失效域名聚类分析过程主要是通过请求失效域名的主机IP地址信息对失效域名进行聚类分析，提取 出可疑失效域名集合、潜在感染主机集合。新域名聚类分析过程主要是通过新域名解析返回的IP地址对新域名进行聚类分析，利用聚类结果将请求主机划分多个请求组。恶意域名提取过程主要是通过失效域名聚类结果，对请求每个新域名集合的主机组进行分析，判断组内主机在域名请求行为上是否具有组行为特征，提取出最终的恶意域名、感染主机、C&C服务器IP地址集合，并对分析结果进行评估。

### 2.3.2 针对色情域名的研究

当前针对色情域名的研究大多数集中在针对网页内容的研究上，Luh等人基于内容实现了对网站的分级[22]，苏贵阳等人基于页面汉字文本实现了中文色情网站识别[23]。除此之外还有一些研究人员通过网络访问日志来进行色情网站的识别，曹建勋等人基于日志中所记载的用户行为的挖掘，验证了用户访问色情网站与普通网站时的行为确实具有明显的差异[24]。他们从日志中提取了特定时间段网站的访问率、搜索引擎以及网站内部网页访问率、色情词关联度、URL特定字符关联度。他们最后使用了朴素贝叶斯分类，五折交叉验证，最终准确率和召回率都达到96%以上。

## 2.4 威胁情报平台相关应用

在学术领域Notos、Pleiades、Kopis、EXPOURE等一些列经典的系统都已在上文中仔细介绍过了，这一节的重点放在国内外工程与商业上的应用成果。

Nominum是为网络运营商提供综合型用户、网络和安全解决方案的全球领先提供商[31]。Nominum所建设的N2平台每天处理的DNS查询量数量级达到万亿次，实现利用DNS数据的应用的快速开发与无缝集成。在海量数据的支持下，始终活跃在抗击网络攻击的最前线，同时提供着最为权威的安全报告，对当前流行的、危害高的网络威胁（例如Ddos、随机子域名攻击等）都有着最为细致的分析。图4-1是他们的DNS数据分析系统。

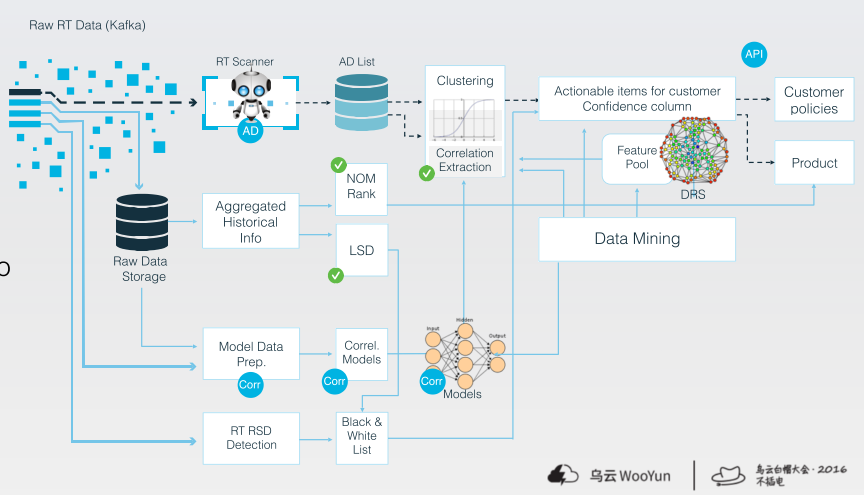


图4-1 NominumDNS数据分析系统

微步在线（ThreatBook）是一家从事专业威胁情报的公司，所发布的威胁分析平台VirusBook（www.virusbook.cn），也是国内首款综合性威胁情报分析平台。随着定向攻击及APT的日益泛滥，业界清醒的认识到：单纯的防御方式很难消除面临的关键风险，安全检测和响应的重要性重新被发现，作为贯穿检测－定位－决策－行动全过程的威胁分析已成为当今的热点。现阶段安全分析师所使用的分析平台主要是国外的站点，如VirusTotal、DomanTools、PassiveTotal及一系列的开源情报站点。但是这些站点还存在着一些问题：首先，这些站点大多只提供了分析所需要的某一方面的信息，并不能完全覆盖整个分析过程，导致分析师需要到多个站点进行查询，并手动进行关联分析，这样的工作无疑是枯燥、低效的，严重影响了分析师能力的发挥；其次由于网络的问题，某些重要网站往往不能正常使用，使工作难以进行。而微步的威胁分析平台做到一个站点完成鉴别、定性、溯源、追踪等多重任务，弥补了国内市场的不足。

这里需要介绍一下360网络安全实验室所建立的一个威胁情报项目[32]，公布了所有的检测数据并且做到每天更新，主要分为以下四个部分：

* 针对全自动攻击工具Exploit Kit的行为检测；
* 对20个DGA家族和超过170个种子的实时追踪；
* 对任一个IP、域名或者MD5值的恶意软件关联；
* 当下流行的恶意代码扫描。

## 2.5 本章小结

# 第三章 利用PDNS检测恶意域名的算法研究

本章主要介绍利用PDNS检测恶意域名的方案研究与设计。

## 3.1 PDNS数据介绍

被动DNS（PassiveDNS）数据是网络安全领域最为常用的资源之一，该数据收集技术由Florian Weimer于2004年提出[19]，主要目的是将DNS流量转换为易于访问的格式。递归域名服务器会响应其接收到的来自其它域名服务器的请求，对响应进行记录并将日志数据复制到中央数据库当中，因此记录这些数据就掌握了DNS解析的历史动态。具体而言，各类可联网终端设备（包含移动终端、台式机等）在需要进行对域名进行访问时，就产生了DNS查询，DNS服务器通过不断递归的方式向上查找，并返回结果，那么所有使用这个DNS服务器的用户都会通过某一层的递归服务器，并在上面留下相应的记录。通过对交换机、路由器端口的配置，将流向或者流出递归服务器的DNS流量实时的拷贝到一台专门的服务器上

Zdrnja等人首先说明了如何利用被动DNS从域名中获得安全信息[20]，2008年Plonka等人提出了Treetop[21]，它可以弹性地管理逐渐增长的被动DNS数据，并在同时关联域和网络属性。他们的聚簇区域是基于不同的种类的网络。Treetop依据是否符合各种DNS RFC标准以及解析结果来区分DNS流量并提供部分安全信息。

表 所示为被动DNS记录中保留的字段，在分析和实验之中主要使用到的字段有客户端IP、请求域名、请求类型、服务器响应标识等。域名请求类型主要有A、TXT、CNAME、ANY、MX等。返回的域名类型和请求类型的不同之处在于没有ANY的返回结果，因为客户端发出ANY的查询请求后，递归服务器会将该域名能收集到的各种类型记录分别返回，这也是为什么在DNS放大攻击中攻击者通常会使用ANY记录查询发起攻击。另一个需要的字段是服务器响应标识，当Rcode为0时，表示没有发生异常，返回正确结果；当Rcode为1时，表示服务器收到的请求格式错误；当Rcode为2时，表示递归服务器出现错误；当Rcode为3时，表示递归服务器无法找到域名，这部分域名即NxDomain；当Rcode为4和5时，分别表示服务器无法解析和拒绝解析。

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 备注 |
| SIP | ipv4\_addr | 客户端IP，字符串 |
| DIP | ipv4\_addr | 递归IP，字符串 |
| DNSID | int | DNS请求编号 |
| DOMAIN | char(256) | 请求域名 |
| QTYPE | char(10) | 请求域名类型，字符串 |
| SCOUNT | int | 请求报文重复次数 |
| SRATE | Int | 该时间间隔下报文采样率（如10:1） |
| DIR | int | 匹配标识（请求/应答/双向） |
| RD | Int | 递归请求标识 |
| AA | Int | 权威请求标识 |
| TTL | int | 应答的首个RR的TTL |
| TIME | datetime | 当前获得报文的时间 |
| PCODE | char(10) | 节点编码（编码参考附录） |
| QLEN | Int | 请求Ip包长 |
| RLEN | Int | 应答IP包长 |
| RRTYPE | char(10) | 应答的首个RR的类型 |
| RCODE | Int | 服务器响应标识 |
| VALUE | char(256) | 应答的首个RR的解析值 |

被动DNS资源记录保留的是递归服务器与客户端之间的DNS查询记录，具体分为R2C（Recursive Server to Client）资源记录和C2R（Client to Recursive Server）资源记录，两者最明显的区别在于是否包含RRTYPE字段，我在实验中所使用的为广东电信提供的R2C资源记录。

## 3.2 恶意域名特征介绍与分析方案

本节主要讲述域名的特征提取，这里只针对所有域名的二级域名。例如域名“*test.example1.com.cn*”和“*test.example2.com*”，这里“*.com.cn*”和“*.com*”被称为顶级域名，而“*example1.com.cn*”和“*example2.com*”被称为二级域名。大部分恶意域名都可以在二级域名这一级进行检测，例如DGA算法产生的域名，一般会从顶级域名选择一个或者多个，与生成有意义或无意义的字符串拼接成一个二级域名。其他的恶意域名也类似，攻击者可以直接注册二级域名来使用。其他情况例如随机子域名攻击，攻击者会利用合法二级域名构建子域名，这些子域名多是三级的甚至是四级的，例如“*test1.example2.com*”和“*test2.test1.example2.com*”，他们仍然具有相同的二级域名“*example2.com*”，因此我们也会统计一定时间下一个二级域名的不同的子域名数量。

### 3.2.1 域名字符特征

一个合法域名*d*包含了数字、字母、“*.*”、“*\_*”，例如“*www.example.com*”。最右边的部分被称为*TLD*（top-level domain，顶级域名），例如“*.com*”。最右边的两个部分被称为*2LD*（second-level domain，二级域名），例如“*.example.com*”。同样的，最右边三个部分被称为*3LD*（third-level domain，三级域名），例如“*www.example.com*”。构成域名字符特征的种类很多，这里主要选取三类特征，一个域名可以得到12个N-gram特征、2个熵的特征、3个其他统计特征，共17个字符特征。

**N-gram特征：**这里首先将白名单中的域名进行处理，把每个*d*的*SLD*去除*TLD*部分。然后在头部加上字符“*^*”，尾部加上字符“*$*”，例如合法域名“*www.example.com*”则变为“*^example$*”。将白名单中每一个域名做上述处理后用作语料库，分别计算N-gram的频率值，其中n的取值为1、2、3、4。当n取值为1时，不考虑起始字符“*^*”和结尾字符“*$*”，则“*^example$*”处理为[‘e’, ‘x’, ‘a’, ‘m’, ‘p’, ‘l’, ‘e’]。当n取值为2时，“*^example$*”处理为[‘^e’, ‘ex’, ‘xa’, ‘am’, ‘mp’, ‘pl’, ‘le’, ‘e$’]。由此获得域名每个子字符串的N-gram概率，例如一个长度为7的字符串，我可以得到7个1-gram频率值，8个2-gram频率值，7个3-gram频率值和6个4-gram频率值。分别计算每组频率值的均值，中位数和标准差，最后一个域名可以得到12个N-gram特征。

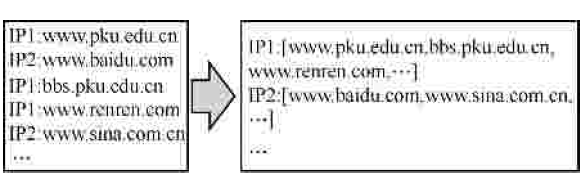
**熵的特征：**给定一个域名，将每个域名*d*的*SLD*去除*TLD*部分之后，按照公式1来计算熵值*H(d)*，其中*p(c)*为二级域名中每一个字符的出现概率，熵值可以表现出一个域名构成的随机程度。然后计算*H(d)*除以字符串长度l，得到字符平均熵，这里我们每一个域名可以获得两个熵相关的特征。

**其他统计特征：**给定一个域名，将每个域名*d*的*SLD*去除*TLD*部分之后，统计其长度*L*，及其中字母数量*La*和数字数量*Ln*。由于“*.com*”顶级域名下恶意域名数量和其它顶级域名下恶意域名数量的比例差别很大[11]（监管原因和价格原因），因此还需要记录该域名的顶级域名部分是否是“*.com*”。这一部分统计以上三个特征。

### 3.2.2 域名访问特征

在被动DNS中记录了每一个IP对域名的请求，接下来讨论一下如何记录IP和域名的访问关系特征，在Antonakakis等人的经典论文“从丢弃的流量中检测僵尸网络”中，他们使用一个二部图来记录相关特征[11]。感染了同一种DGA的恶意软件的主机有着很大的可能生成互相重叠的NXDomains集合。另一方面，其他“非DGA”的NXDomains则不会被多个主机查询。例如，在一段时间内多个用户同时犯同样的拼写错误是不太可能的。他们由此构建了一个稀疏矩阵M，其中行代表一段时间内查询过两个以上NXDomains的主机，列代表NXDomains，并对M进行谱聚类将相近主机共同查询过的NXDomains聚簇在一起。再统计聚簇的特征，做训练和分类使用。这样做的缺点是无法获取单个域名的特征，而且过程十分复杂，无法有效的将域名访问特征体现出来，因此我参考周昌令等人的文章将域名映射到向量空间之中[25]，在保留域名关联的同时，将出现达到一定次数的域名转化为向量，同时也避免了使用诸如one-hot编码导致向量过长，便于进一步的计算。这里将产生一个200维的词向量。除此之外，流量计数相关统计特征。。。。。。

**Word2vec特征：**word2vec是2013年由谷歌开发出一个词向量工具[26]，该算法既可以通过上下文来推测目标词语构建向量（CBOW模式），也可以通过词来推测上下文以构建向量（Skip-Gram模式），本文使用Skip-Gram模式。在经过3.3.1节的基本数据清洗之后，将每一个IP一定时间下的DNS查询按照时间戳的先后顺序并提取其SLD串联起来，相同的SLD可以多次重复的出现，如图3-3构建被域名查询列表[25]。这样构成的每一个IP的被域名查询序列看成是一句话，在本文中仅考虑所有的A记录查询。然后使用基于gensim的Word2Vec模块训练词向量。

****

**访问统计相关特征：**（qn,qan,qtn,子域名数量、nxdr）

|  |  |  |
| --- | --- | --- |
| 特征 | 数据类型 | 备注 |
| qac |  | 该域名被查询A记录的总次数 |
| qcnamec |  | 该域名被查询CNAME记录的总次数 |
| qtxtc |  | 该域名被查询TXT记录的总次数 |
| qmxc |  | 该域名被查询MX记录的总次数 |
| qnsc |  | 该域名被查询NS记录的总次数 |
| qdnskeyc |  | 该域名被查询DNSKEY记录的总次数 |
| qanyc |  | 该域名被ANY查询的总次数 |
| qsoac |  | 该域名被查询SOA记录的总次数 |
| qspfc |  | 该域名被查询SPF记录的总次数 |
| qsrvc |  | 该域名被查询SRV记录的总次数 |
| qdsc |  | 该域名被查询DS记录的总次数 |
| qnaptrc |  | 该域名被查询NAPTR记录的总次数 |
| success |  | 该域名A记录或AAAA记录查询成功的总次数 |
| wrongqueryc |  | 该域名A记录或AAAA记录查询不存在的次数 |
| wrongserverc |  | 该域名A记录或AAAA记录查询格式错误总次数 |
| nonexistc |  | 该域名A记录或AAAA记录查询不存在总次数 |
| sipc |  | 不相同的请求IP总个数 |
| secdomainc |  | 该域名下被查询的不相同子域名总个数 |
| qac\_n |  | 第n小时该域名A记录被查询的次数 |
| qcnamec\_n |  | 第n小时该域名CNAME记录被查询的次数 |
| qtxtc\_n |  | 第n小时该域名TXT记录被查询的次数 |
| qmxc\_n |  | 第n小时该域名MX记录被查询的次数 |
| qnsc\_n |  | 第n小时该域名NS记录被查询的次数 |
| qanyc\_n |  | 第n小时该域名ANY记录被查询的次数 |
| qsrvc\_n |  | 第n小时该域名SRV记录被查询的次数 |
| success\_n |  | 第n小时该域名A记录记录查询成功的次数 |
| nonexistc\_n |  | 第n小时该域名A记录记录查询不存在的次数 |
| sipc\_n |  | 第n小时不相同的请求IP个数 |
| secdomainc\_n |  | 第n小时该域名下被查询的不相同子域名个数 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

备注：由于数据的原因，其中n为13至23共11个小时

除此之外，还有一些由上述统计特征衍生出来的特征，

|  |  |  |
| --- | --- | --- |
| 特征 | 数据类型 | 备注 |
| DNSamplification\_score |  | DNS放大攻击得分（公式1） |
| PRSD\_score |  | 随机子域名攻击得分（公式2） |
| qar |  | 该域名A记录查询数量占比 |
| qcnamer |  | 该域名CNAME记录查询数量占比 |
| qtxtr |  | 该域名TXT记录查询数量占比 |
| qsuccessr |  | 该域名成功查询数量占比 |
| qnonexistr |  | 该域名不存在查询数量占比 |
| qacperip |  | 平均每个ip的A记录查询量 |
| vp\_n |  | 第n小时该域名查询量占总查询比例 |
| qac\_avg |  | 该域名A记录每小时查询数平均值 |
| qac\_sd |  | 该域名A记录每小时查询数方差 |
| successc\_avg |  | 该域名每小时被成功查询数平均值 |
| successc\_sd |  | 该域名每小时被成功查询数方差 |
| nonexistc\_avg |  | 该域名每小时被查询不存在平均值 |
| nonexistc\_sd |  | 该域名每小时被查询不存在方差 |
| vp\_sd |  | 该域名每小时查询数占比方差 |
|  |  |  |
|  |  |  |
|  |  |  |

### 3.2.3 特征分析方案

现有特征较为庞大，数据类型和表现形式上均有所不同，拟针对不同类型的特征给出不同的特征分析方案。

对于3.2.2节中的域名访问特征，由原始的统计特征衍生出了一系列特征，针对这些衍生变量我需要探索是否和原本的统计变量存在简单的线性相关，因此需要计算这些衍生变量和原本统计变量之间的简单相关系数，如公式 所示，其中*Cov(X,Y)*表示X和Y的协方差，*Var(X)*表示X的方差。

进一步的对于这些域名访问特征拟通过信息值来判断特征的优劣，IV（Information Value，信息价值）是一种对自变量预测能力的量化衡量指标，

## 3.3 快速提取恶意域名相关流量的方案

### 3.3.1 数据预处理

快速提取恶意域名相关流量的方案对采集来的原始数据进行三层预处理。分别是数据清洗、数据过滤和特征提取。首先对所有数据进行清洗，一个合法的域名只包含26个英文字母（包括大写和小写）、数字、中划线和用来分割成每一段的点。通过构建正则表达式很容易可以将这些数据清理干净。还有一部分是无顶级域名的域名，产生这种情况的原因比较多，有配置或者人为的各种原因。

其次，进行数据的过滤操作。第一部分先过滤反向解析域名，域名反向解析是指从IP地址到域名的映射，其主要应用于邮件服务器阻止垃圾邮件。为了实现逆向域名解析，因此有一个专门DNS服务器负责反向解析，返回数据包较小，不适合用作放大攻击，因此恶意流量中没有出现利用反向解析记录进行攻击的行为。第二部分要过滤配置错误产生的域名，这类域名极为常见，这其中以“*.local*”,“*.localhost*”为 后缀的域名出现最多。第三部分过滤国际化域名，国际化域名是指非英语国家推广本国语言域名系统的一个总称，使用punycode编码。

### 3.3.2 针对DNS放大攻击相关域名的提取

这一部分主要目的是要将放大攻击中作为跳板的那部分域名找出来。攻击者想要利用这些域名，必然是利用其TXT记录或者ANY查询返回该域名所有资源记录。我们获取ANY类型请求比例qar和TXT类型请求比例qtr，带入公式2中，β为我们设定的一个参数，当qar+qtr<=β时，结果为0，当qar+qtr>β时qar+qtr与s1成正相关，同时设定阈值α，其中s1>α，我们认定为疑似放大攻击的流量。这里α取值为0.1，β取值为0.05，阈值的取值将在下一章仔细讨论。

### 3.3.3 针对随机子域名相关域名的提取

这一部分目的是将随机子域名攻击所使用的那部分域名找出来，攻击者会在二级域名下伪随机的生成大量子域名，这些域名都是不存在的。因此我们用sdc与nxdr相乘来表示被恶意使用的可能性大小，而这个值范围较大，我们使用公式3将结果变到0到1之间。这里我们的θ取值为0.3.

### 3.3.4 针对DGA域名的提取

域名生成算法产生的域名也是二级域名的部分，所以这部分的检测关注针对流量中的二级域名，我们模型使用黑白名单训练，数据来源在4.1节中介绍。将名单中每个域名的二级域名提取出来，分别计算其长度、熵值、n-gram，其中n=2,3,4。分类器我们选择随机森林，训练每棵树时，从全部训练样本（样本数为N）中选取一个可能有重复的大小同样为N的数据集进行训练（即bootstrap取样）。随机森林的训练速度快，并且可以平衡误差。

## 3.4 恶意域名分类算法方案

本节针对。。。。

### 3.4.1 DGA域名检测算法

### 3.4.2 色情域名检测算法

## 3.5 本章小结

# 第四章 恶意域名检测应用原型系统设计与实现

本章是主要根据。。。。。。并设计实现了原型程序。

## 4.1 需求分析

### 4.1.1 PDNS预处理评分模型

### 4.1.2 域名特征提取模块

### 4.1.2 恶意域名分类模型模块

## 4.2 总体设计

## 4.3 功能实现

### 4.3.1 特征提取

### 4.3.2 参数选择

### 4.3.3 流量监控

在获取了原始的被动DNS之后，需要一个切实有效的可视化方案。这里我选择使用Influxdb + Grafana构建一个流量监控系统。Influxdb是一个开源的分布式时序、时间和指标数据库，使用go语言编写，无需外部依赖。它有三大特性：一是时间序列，因此与时间相关的函数使用非常的灵活（诸如最大、最小、求和等）；二是度量，可以对实时大量数据进行计算；三是事件，它支持任意的事件数据。同时Influxdb还具有无结构（无模式）、扩展性强、支持大量统计函数、原生的HTTP支持、强大的类SQL语法等一系列图片包含 电子产品, 陈列

已生成极高可信度的说明优点。我选择Influxdb作为数据库也正是基于此。

图2-2 流量监控系统

Grafana 是一个开箱即用的可视化工具，具有功能齐全的度量仪表盘和图形编辑器，有灵活丰富的图形化选项，可以混合多种风格，支持多个数据源特点。

### 4.3.4 恶意域名检测

## 4.4 本章小结

。。。

# 

# 第五章 实验结果与分析

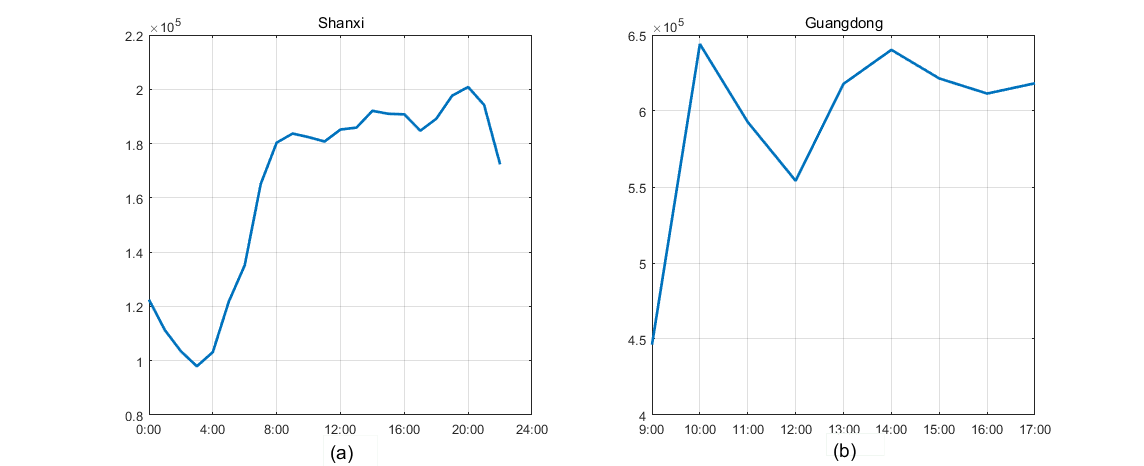
本章主要对按照前文所述设计的原型系统所产生的实验结果进行分析。

## 5.1 总体情况

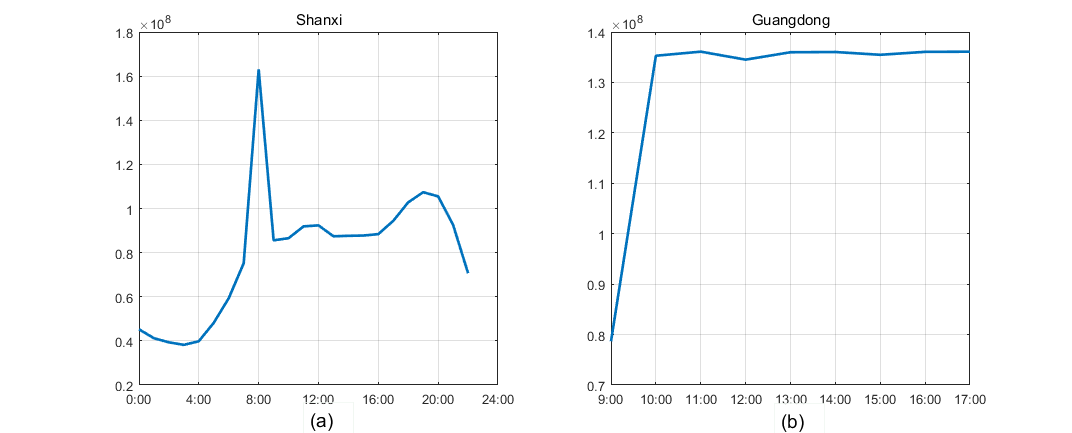
## 5.2 样本分析

### 5.2.1 恶意域名快速提取样本分析

我们使用了来自CNCERT/CC提供的数据，包括了山西省的中国电信PDNS数据和广东电信的PDNS数据。其中山西省数据为2015年10月15日 23个小时的数据，如图所示，DNS记录的总量接近20亿条，每小时不重复的二级域名数量在10万到20万之间，共标注DDOS相关恶意域名101个，DGA相关域名322个。广东省的数据为2017.4.14 9点到16点共计9个小时的数据，如图所示，DNS记录总量达11亿余条，每小时不重复的二级域名数量达到60万左右，共标注DDOS相关恶意域名163个，DGA相关域名265个。

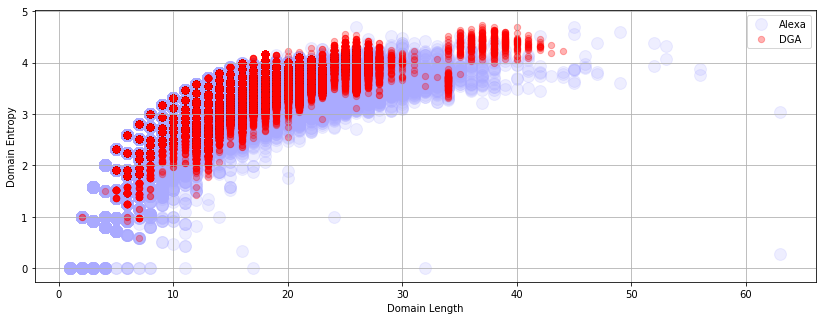


图：山西和广东不重复二级域名数量



图：山西和广东DNS记录数量

在进行DGA流量的提取中，我们使用alexatop100万的域名列表作为白名单，黑名单我们将从360安全实验室下载的DGA黑名单作为黑名单，包含1037304条二级域名。图展示了黑名单与白名单中的二级域名长度和熵值之间的关系，在长度一致的时候，dga的域名往往与有更大的熵值。



图：域名熵值随着域名长度的变化

### 5.2.2 DGA域名检测样本分析

由于原始数据时间跨度有限，因此无法在原始被动DNS数据中进行实验和验证，因此只保留了二级域名出现次数大于五次的域名。我们将DGA黑名单中出现的域名定为正样本，黑名单中未出现的域名视为负样本，正负样本比例大概为1：20。正负样本过于不均衡，因此使用了下采样的方法，对负样本进行了抽样，保留所有正样本，负样本每天抽取25000个，使正负样本比例达到1：5左右。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 20171127 | 20171128 | 20171129 | 合计 |
| 正样本 | 5211 | 6072 | 5451 | 16734 |
| 负样本 | 25000 | 25000 | 25000 | 75000 |
| 总样本 | 30211 | 31072 | 30451 | 91734 |
| 正样本比例 | 17.249% | 19.542% | 17.901% | 18.242% |

## 5.3 特征分析

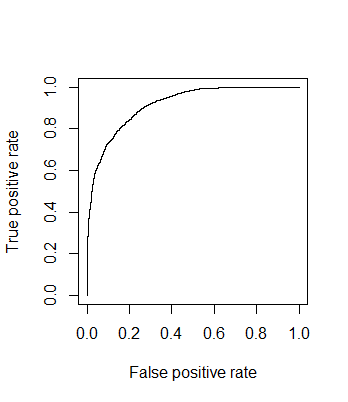
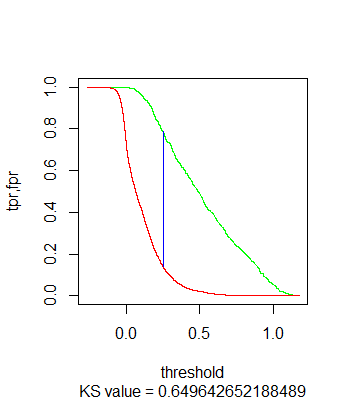
### 5.3.1 DGA域名检测特征分析

这部分内容是对各类特征进行初步的可用性分析，针对每一个单类特征进行分析，使用AUC、信息值等指标来对特征的好坏进行一个评估，评估使用2017年11月27日数据。

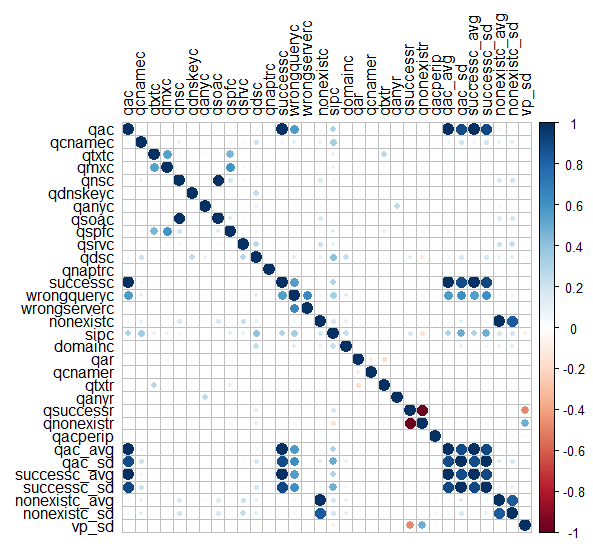
**Word2Vec特征分析：**使用gensim训练好的Word2Vec模型，随机挑选域名“*078mvrxcg4j3b49b.net*”，该域名属于Chinad家族木马产生的域名，表 所示为余弦相似度最高的十个域名。明显可见，虽然没有使用域名的字符特征训练余名，而只是利用访问序列训练出来的向量，具有很高的实用性。

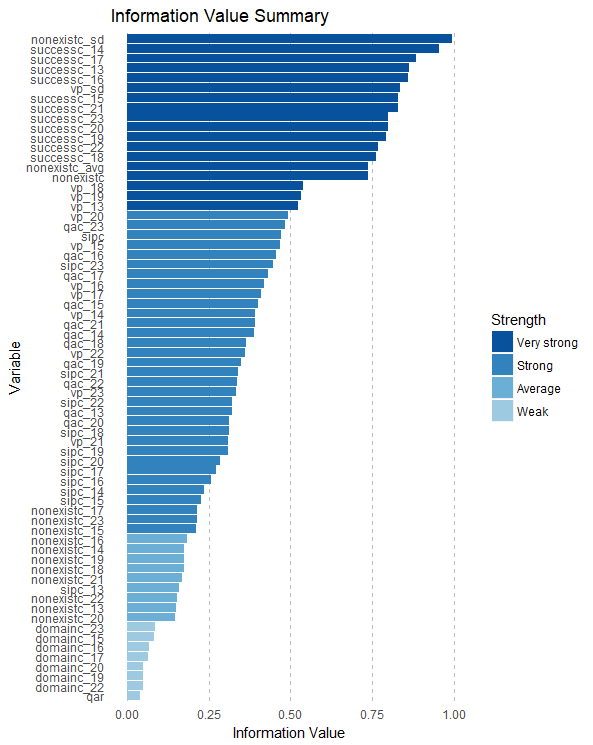
|  |  |  |
| --- | --- | --- |
| 域名 | DGA家族 | 余弦相似度 |
| wbopgyg26s4jtdjb.net | Chinad | 0.9648110270500183 |
| o1cw6qug9ixwhwlo.info | Chinad | 0.9643963575363159 |
| 60hwibc80hgdd853.net | Chinad | 0.964003324508667 |
| 18o67n7qrnvoca9k.info | Chinad | 0.9635913968086243 |
| x9se3frkdzvhob2q.biz | Chinad | 0.9630709290504456 |
| rcsw95z947xdx09q.cn | Chinad | 0.9628362059593201 |
| ait1kf8sae7cdjhm.info | Chinad | 0.9628314971923828 |
| 7l8zc789bucj50ky.org | Chinad | 0.9625819325447083 |
| 1uw34d1ywa6pf3gg.org | Chinad | 0.962489664554596 |
| nexd7g076ppa6n6w.ru | Chinad | 0.9624853730201721 |

进一步的将向量作为特征放入xgboost模型中，粗略调参，查看效果。这里xgboost具体参数设置为nrounds = 500,max\_depth = 6,eta = 0.02。使用五折交叉验证，将27日的向量数据分随机分为五份，其中四份用于训练，一份用于验证。图 所示为验证集的ROC（Receiver Operating Characteristic）曲线，其中横坐标为FPR（false positive rate），纵坐标为TPR（true positive rate），这样不必设定阈值就可以看出这个分类器的性能，曲线下方的面积为AUC（Area Under Curve）。AUC是二值分类器非常重要的评价指标之一，他表示从样本验证集的正负样本中各取一个样本，其中正样本的模型预测结果数值大于负样本的模型结果预测数值的概率，如果是完全随机分布的话，AUC的值应该约为0.5，这里的计算结果为0.915，可以看出效果极好。图 所示为验证集的K-S曲线，横坐标为阈值，纵坐标为在不同阈值下的TPR和FPR，ks数值为TPR与FPR差值最大的地方，即图中的蓝线，K-S曲线可以很好的表示模型对正负样本的区分能力。这里K-S的值为0.649，说明该模型在使用Word2Vec特征的情况下具有良好的正负样本分类能力。

**访问记录特征：**首先计算衍生变量和原本变量的简单相关系数，避免衍生变量与原本变量具有过强的相关性，造成特征的冗余。图 所示为相关系数的展示图，总体来说相关性较为稀疏，有部分变量具有明显的相关性，例如qnonexistr（域名查询不存在比例）和qsuccesser（域名查询成功占比）具有明显的负相关，qac（域名A记录查询数量）和qsuccesser具有较为明显的正相关。





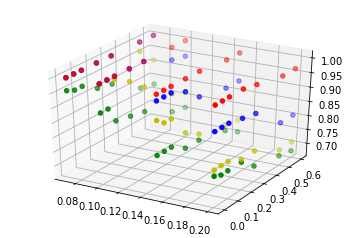
## 5.2 参数的讨论

### 5.2.1 针对DNS放大攻击提取相关参数

我们先3.3.2节中的参数进行讨论，这里的参数涉及到了三个，分别是时间间隔、判断域名是否可疑的阈值α、公式1中的参数β。其中时间间隔取值为10、30、60、120，α取值为0、0.05、0.1、0.2、0.3、0.5、0.6，β的取值为0.07、0.1、0.15、0.2，如图所示，z轴代表召回率，当时间间隔为10时，召回率始终在0.98以上，当时间间隔为120时，召回率最高为0.94，都无法满意的对α与β取值。表1展示了在不同参数的取值与可疑域名数量之间的关系。可以看到时间间隔选为30和时间间隔为60所能达到的效果差别不大，而前者的执行次数为后者的两倍，因此我们将时间间隔定为60分钟，为了获得尽可能小的域名数量，我们将α设定为0.1，β设定为0.05。

Table 1: When the recall rate is 100%, the relationship between α, β and the number of suspected domain names

|  |  |  |  |
| --- | --- | --- | --- |
| α | β | 30 mins | 60 mins |
| 0.07 | 0.05 | 784 | 764 |
| 0.07 | 0.1 | 779 | 758 |
| 0.07 | 0.2 | 768 | 749 |
| 0.07 | 0.3 | 762 | 738 |
| 0.1 | 0 | 751 | 731 |
| 0.1 | 0.05 | 750 | 730 |



### 5.2.2 针对随机子域名提取相关参数

另一个需要讨论的参数是3.3.3节中公式3中的θ，θ的值越小，这个函数的曲线越平滑。如图3所示，描述了θ的取值和提取出来的域名的数量之间的关系。其中当θ为0.1时，召回率为50%。当θ取值大于0.2时，召回率达到100%。从图中可以看出θ的取值域名数量保持正相关，为避免过拟合，我们将θ取值为0.3。

## 5.3 恶意域名检测结果

## 5.3 本章小结

# 总结与展望

。。。

## 研究工作总结

。。。。

## 未来工作展望

。。。。

# 参考文献

1. 第38次中国互联网络发展状况统计报告. CNNIC，2016.
2. Bilge, L., Kirda, E., Kruegel, C., Balduzzi, M., & Antipolis, S. (2011). EXPOSURE : Finding Malicious Domains Using Passive DNS Analysis. Ndss, 1–17.
3. 2016中国互联网网络安全报告, 2017.
4. Vissers, T., Joosen, W., & Nikiforakis, N. (2015). Parking Sensors-PPT. Proceedings 2015 Network and Distributed System Security Symposium, (February), 8–11.
5. Plohmann Daniel,Fkie Fraunhofer,Yakdan Khaled,Klatt, Michael.A Comprehensive Measurement Study of Domain Generating Malware, USENIX Security Symposium,2016.
6. 2015DDoS威胁报告.绿盟科技.2015.
7. Attackers are increasingly leveraging large Domain Name System (DNS) TXT records in an effort to amplify the impact of their distributed denial-of-service (DDoS) attacks, Akamai's Prolexic Security Engineering and Research Team (PLXsert) warned on Tuesday,http://www.securityweek.com/large-dns-text-records-used-amplify-ddos-attacks-akamai.
8. Plohmann, D., Fkie, F., Yakdan, K., Klatt, M., Bader, J., Gerhards-padilla, E., … Bader, J. (2016). A Comprehensive Measurement Study of Domain Generating Malware. USENIX Security Symposium.
9. Antonakakis, M., Perdisci, R., Dagon, D., Lee, W., & Feamster, N. (2010). Building a Dynamic Reputation System for DNS. USENIX Security’10: Proceedings of the 19th USENIX Conference on Security, 1–17.
10. Antonakakis, M., Perdisci, R., Lee, W., Ii, N. V., & Dagon, D. (2011). Kopis:Detecting Malware Domains at the Upper DNS Hierarchy. USENIX Security Symposium., 11, 1–16.
11. Antonakakis, M., & Perdisci, R. (2012). From throw-away traffic to bots: detecting the rise of DGA-based malware. Proceedings of the 21st USENIX Security Symposium, 16.
12. Perdisci R, Corona I, Giacinto G. Early detection of malicious flux networks via large-scale passive DNS traffic analysis. IEEE Transactions on Dependable and Secure Computing, 2012, 9(5): 714–726
13. Bilge, L., Kirda, E., Kruegel, C., Balduzzi, M., & Antipolis, S. (2011). EXPOSURE : Finding Malicious Domains Using Passive DNS Analysis. Ndss, 1–17.
14. Nelms, T., Perdisci, R., & Ahamad, M. (2013). ExecScent: Mining for New C&C Domains in Live Networks with Adaptive Control Protocol Templates. Proceedings of 22nd USENIX Security Symposium, 589–604.
15. Rahbarinia, B. (2016). Segugio:Efficient and Accurate Behavior-Based Tracking of Malware-Control Domains in Large ISP Networks, 19(2).
16. Plohmann, D., Fkie, F., Yakdan, K., Klatt, M., Bader, J., Gerhards-padilla, E., Bader, J. (2016). A Comprehensive Measurement Study of Domain Generating Malware. USENIX Security Symposium.
17. 张雪松,徐小琳,李青山. 算法生成恶意域名的实时检测[J]. 现代电信科技,2013,07:3-8.
18. 张维维,龚俭,刘茜,刘尚东,胡晓艳. 基于词素特征的轻量级域名检测算法[J]. 软件学报,2016,09:2348-2364.
19. F. Weimer. Passive DNS replication. In Proceedings of FIRST Conference on Computer Security Incident, Hand ling, Singapore, 2005.
20. B. Zdrnja, N. Brownlee, and D. Wessels. Passive mon- itoring of DNS anomalies. In Proceedings of DIMVA Conference, 2007.
21. D. Plonka and P. Barford. Context-aware clustering of DNS query traffic. In Proceedings of the 8th IMC, Vou- liagmeni, Greece, 2008. ACM.
22. Lee L, Luh C. Generation of pornographic blacklist and its Incremental update using an inverse chi-square based method[J]. Information Processing and Management, 2008, 44 (5):1698-1706
23. Su Guiyang, Li Jianhua, Ma Yinghong, et al. A KNN algorithm on Chinese erotic text filtering [J]. Journal of Shanghai Jiaotong University, 2004 , 38: 86-79
24. 曹建勋,刘奕群,岑荣伟,马少平,茹立云.基于用户行为的色情网站识别[J].计算机研究与发展,2013,50(02):430-436.
25. 周昌令,栾兴龙,肖建国.基于深度学习的域名查询行为向量空间嵌入[J].通信学报, 2016, 37(3):165-174.
26. Tomas Mikolov. Word2vec project[EB/OL]. [2014-09-18]. https://code.google.com/p/ word2vec/.
27. MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. c2013:3111-3119.

# 攻读硕士学位期间取得的学术成果

。。。。

# 致谢

。。。。