文章编号:1674-6864(2019)05-0045-06 DOI: 10.16508/j.cnki.11-5866/n.2019.05.009

DGA恶意域名检测方法

蒋鸿玲,戴俊伟
(北京信息科技大学 信息管理院,北京 100192)

摘 要: 针对目前DGA(domain generation algorithm)恶意域名检测方法计算量较大、检测精确率不高等问题,提出了DGA恶意域名检测框架。首先对域名的字符统计特征和N-Gram模型特征进行分析,提取出区分度大的域名特征组合;然后利用正常域名和DGA恶意域名数据集训练不同的机器学习模型,如朴素贝叶斯、多层感知器和XGBoost(extreme gradient boosting)模型,再用训练好的模型检测恶意域名。实验结果表明,采用域名的N-Gram模型特征的精确率和召回率都优于统计特征,多层感知器的精确率较高,误报率较低,其AUC(area under curve)值高于朴素贝叶斯和XGBoost模型。

关 键 词: DGA;统计特征;N-Gram;朴素贝叶斯;多层感知;极端梯度
中图分类号: TP 393 **文献标志码:** A

DGA malicious domain name detection method

JIANG Hongling,DAI Junwei
(School of Information Management,Beijing Information Science & Technology University,Beijing 100192,China)

Abstract: To solve the problems of large computation and low detection accuracy of DGA (domain generation algorithm) malicious domain name detection method, a framework of DGA malicious domain name detection is proposed. First, the statistical features of domain names and N-Gram model features are analyzed, and the features of domain names with large discrimination are extracted. Then, different machine learning models, such as Naive Bayesian, Multilayer Perceptron and XGBoost (extreme gradient boosting) Model, are trained using normal domain names and DGA malicious domain names data set. Then malicious domain names are detected by the trained model. The experimental results show that the accuracy and recall rate of N-Gram model of domain name are better than statistical features. The accuracy rate of multi-layer perceptron is higher and the false alarm rate is lower. The AUC (area under curve) value of N-Gram model is higher than those of Naive Bayesian model and XGBoost model.

Keywords: DGA; statistical feature; N-Gram; naive Bayesian; multilayer perceptron; extreme gradient

0 引言

近年来恶意程序的数量呈现逐年递增的趋势,并且越来越高级和复杂[1]。因为域名系统(domain name system,DNS)在所有网络中都存在,并且通常不会被防火墙过滤[2],攻击者常使用DNS来隐藏其恶意行为,维护恶意网络自身的健壮。

恶意程序在感染了主机后,通常和远程的命令与控制服务器连接,攻击者可以直接控制命令与控制服务器。如在高级持续性威胁攻击(advanced persistent threat,APT)[3]和僵尸网络中,被感染的主机会通过与远程的C&C服务器(command and control server)连接,下载最新的恶意程序,或者获取恶意指令[4];信息窃取等恶意程序会将窃取的信息发送给远程服务器等;垃圾邮件依赖DNS重定向网页[5]。这些恶意程序常通过域名来访问远程服务

收稿日期: 2019-06-08
基金项目: 北京信息科技大学学校校科研基金(1925023)
第一作者简介: 蒋鸿玲,女,博士,讲师。

器，而不用服务器的 IP 地址[6]，因而域名在恶意行为中发挥着重要的作用[7]。

为了逃避检测，使得恶意网络更健壮，攻击者会采用 domain-flux 技术[8]。domain-flux 技术是指恶意程序采用域名生成算法(domain generation algorithm，DGA)，基于一个种子，如当前的日期，每天动态生成大量的域名，其中的一部分域名是被攻击者注册的有效域名，多个域名对应一个命令与控制服务器的 IP 地址[9]。被感染的主机查询大量自动生成的域名，并与其中少数几个建立连接。由于域名的数量很大，并且每天自动生成，很好地隐藏了攻击者的恶意网络。因而有效检测出 DGA 恶意域名，对发现恶意攻击具有重要的意义。

当前检测基于 DGA 攻击的方法可以分为 2 类。一类是通过分析网络流量，检测 DGA 特定行为的方法，如 Manos 等[10]发现被同一个恶意程序感染的主机会呈现出相似的 DNS 查询行为，产生相似的 NXDomain(域名不存在)响应，通过对相似行为聚类检测基于 DGA 的恶意程序; Reza 等[11]分析可疑的组行为和可疑的 DNS 失败查询，并用负面声誉系统来检测采用 domain-flux 技术的僵尸网络。

另一类是分析域名本身的特征。赵越[12]提出了基于语音和分组特性的 DGA 域名检测方法，着重考虑了 DGA 域名可读性较差以及分组较多的特性，结合 DGA 域名的短文本特性，提取域名的语音等方面特征，并根据这些特征使用随机森林分类器对域名集合进行分类; Stefano 等[13]提取 DGA 域名中有意义字符比例、N-Gram 字符串在字典中的占比等特征，检测 DGA 域名，但需要与海量的字典单词做匹配等分析，计算量较大。

当前通过分析网络流量的方法能够检测未知特征的 DGA 域名，但分析和处理网络流量计算量较大。分析域名字符特征的检测方法计算量较小，但采用哪些域名的字符特征能够较好地区分正常域名和恶意域名，尚没有明确的定论，检测精确率不高。本文通过分析正常域名与恶意域名的字符特征，对比不同特征组合的检测效果，并分析不同机器学习算法检测 DGA 域名的精确率，建立一个 DGA 域名检测框架，有效检测出 DGA 域名。

1 DGA 域名检测框架

本文提出的 DGA 域名检测框架如图 1 所示，主要包括数据集获取、域名特征提取、DGA 检测模型训练和域名检测 4 个部分。

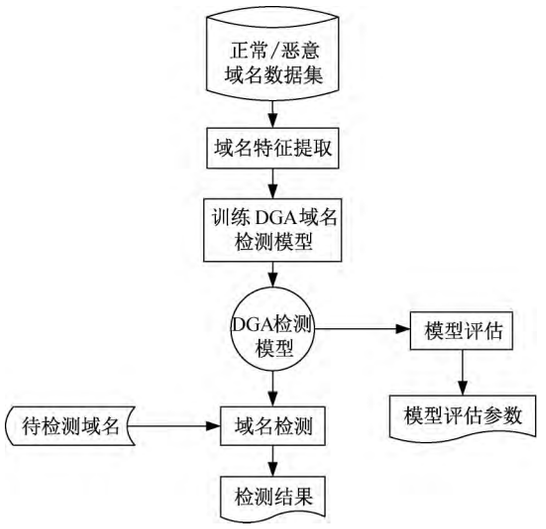

图 1 DGA 域名检测框架

1) 获取数据集。分别获取正常域名和恶意域名数据集，作为已标记的数据集。

2) 特征提取。从域名数据集中提取域名的特征。一类是域名字符的统计特征，如域名字符长度、字符随机性、唯一字符数、元音字母比例等; 另一类是域名的 N-Gram 模型特征。

3) 训练 DGA 域名检测模型。将训练数据集输入到机器学习模型当中进行模型训练。本文工作将采用多种检测模型，并对不同模型进行评估，对比检测效果的优劣。

4) 域名检测。用训练好的模型对域名进行检测，检测出正常的或者恶意的 DGA 域名。

2 DGA 恶意域名特征分析

域名特征提取是 DGA 域名检测的基础，特征选取的好坏直接影响 DGA 域名检测的效果。因此，本文对正常域名和恶意域名的特征进行了分析和对比，以提取出区分度大的特征组合。本文分别对域名字符的统计特征和域名的 N-Gram 模型特征进行了分析。其中，域名字符的统计特征包括: 域名字符长度、域名字符随机性、元音字母比例、唯一字符比例、顶级域名类型等。本文从正常和恶意域名数据集中随机抽取各 2000 条，共 4000 条域名数据进行正常和恶意域名的特征分析。

2.1 域名字符统计特征分析

1) 字符长度。正常域名为了便于用户记忆，不会选择过长的域名，而 DGA 恶意域名是由 DGA 算法随机生成的，不会考虑用户体验，并且为了注册时与现有的域名冲突，会使用较长的域名。

域名字符长度分布如图 2 所示。几乎所有的正常域名的字符长度都在 19 以内,并集中在 8 到 12 之间,仅仅只有少数域名达到了 19 以上。而恶意域名长度范围在 8 到 32 之间,并出现了 2 个高峰点,分别是 12 和 30,其中长度为 30 的恶意域名数量更多。从数据的分布上可以看出近 1/3 的恶意域名的长度在正常域名长度范围之内,但恶意域名长度普遍偏大。

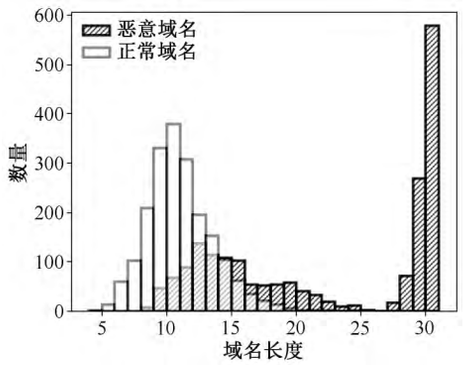

图 2　域名字符长度分布

2) 字符随机性。由于恶意域名是 DGA 算法随机生成的,其字符的随机性较大,混乱程度较高,而正常域名字符的随机性校对较小。域名字符的随机性通过计算字符的熵来判断:

$$H(d) = -\sum \lg(P(X_i)) \times P(X_i) \quad (1)$$

式中: d 为域名; X_i 为 d 中的某一个字符; $P(X_i)$ 为该字符出现的概率。

图 3 为域名字符的熵分布。正常域名字符的熵偏低,熵在 3 左右的正常域名较多; 而恶意域名字符的熵较高,熵在 4 以上的恶意域名较多。但正常和恶意域名字符的熵还是存在少量的交集。

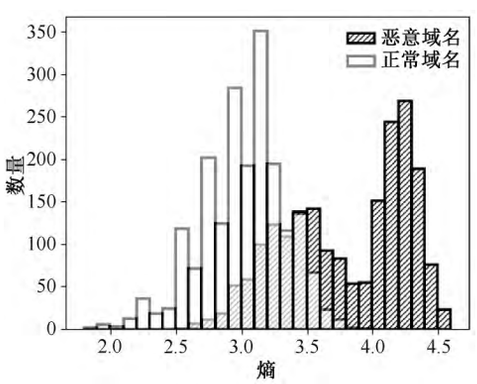

图 3　域名字符的熵分布

3) 元音字母比例分析。正常域名通常采用单词或名字的拼音,方便使用者记忆; 同时为了具有较好的可读性,正常域名往往会插入一定的元音字母,让域名读起来更顺口。恶意域名由于随机生成,不会考虑可读性,因而正常域名的元音字母的比例会比恶意域名的高。图 4 为元音字母比例分布图,正常域名的元音字母比例在 0.4 左右分布的较多,而恶意域名在 0.2 左右分布的较多。

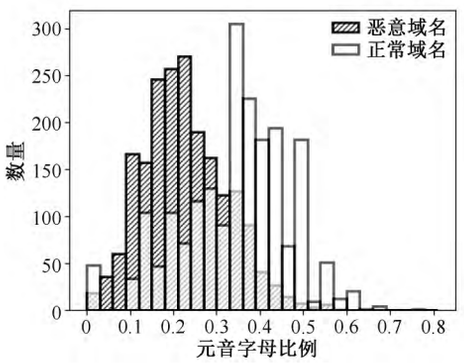

图 4　元音字母比例分布

4) 唯一字符数。DGA 域名有很大的随机性,其唯一字符数会较高。唯一字符是域名中不同字符的个数,如域名 baidu 的唯一字符 [b,a,i,d,u],唯一字符数为 5; 域名 google 唯一字符数有 [g,o,l,e],唯一字符数为 4。正常域名和恶意域名的唯一字符数分布情况如图 5 所示。正常域名和恶意域名的唯一字符数在 6 到 15 之间出现部分交集,唯一字符数量达到 15 以上后就没有出现正常域名了,都是恶意域名。

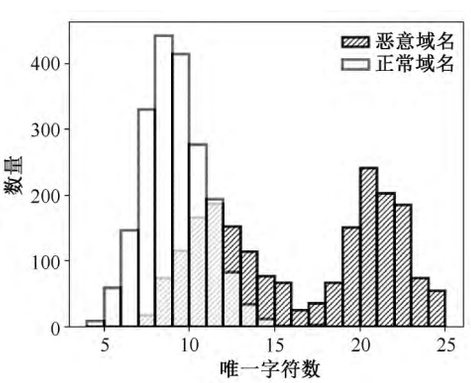

图 5　唯一字符数分布

5) 唯一字符比例。计算域名中唯一字符数与域名长度的比值,正常域名和恶意域名的唯一字符比例如图 6 所示。

从图 6 可知正常域名的比值比恶意域名的比值普遍要大一些。结合图 5 和图 2 可以知道虽然正常域名的唯一字符数普遍小于恶意域名,但是占域名

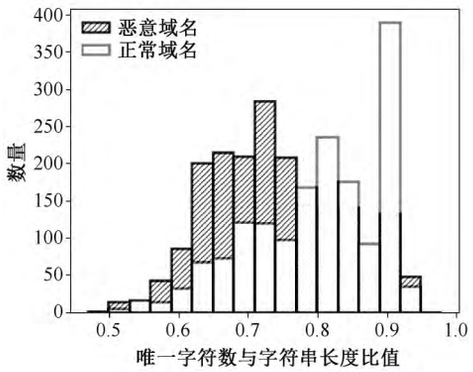

图 6 唯一字符比例分布

字符的比例却高一些。这是由于恶意域名的长度偏大导致。

6) 顶级域名分析。正常域名的顶级域名一般会使用常见的顶级域名,如.cn 和.com 等。恶意域名的顶级域名比较随意,攻击者会选择一些审核不严格的顶级域名,如.biz、.ru 等。本文选择了常用的顶级域名: ['cn', 'com', 'cc', 'net', 'org', 'gov', 'info'],分别统计正常和恶意域名中常用顶级域名和其它顶级域名的数量。正常域名中其顶级域名数为 1933,其他为 67; 恶意域名中其顶级域名数为 1342,其他为 658。大部分正常域名的顶级域名都在常用顶级域名范围内,只有少数个别的没有在其中。恶意域名中有近 2/3 的域名其顶级域名是常用顶级域名。

2.2 N-Gram 模型特征分析

为了区分正常域名和恶意域名,本文采用 N-Gram 模型对域名进行 N-Gram 建模,分别提取域名的 1-gram、2-gram 和 3-gram 特征,然后用机器学习算法对域名的 N-Gram 特征进行处理。如域名数据['baidu.com', 'google.com']经过 1-gram 处理后会获得词汇表 [*a*,*b*,*c*,*d*,*e*,*g*,*i*,*l*,*m*,*o*,*u*],再构建词向量,如表 1 所示。

表 1 1-gram 词向量示例

域名	*a*	*b*	*c*	*d*	*e*	*g*	*i*	*l*	*m*	*o*	*u*
baidu.com	1	1	1	1	0	0	1	0	1	1	1
google.com	0	0	1	0	1	2	0	1	1	3	0

3 DGA 域名检测方法

为了消除不同域名特征之间数据大小的影响,提取域名统计特征后,构建特征向量,并对特征向量进行标准化处理。本文采用 z-score 标准化方法:

$$x' = \frac{x - \mu}{\sigma} \tag{2}$$

式中: x' 为标准化后的数据; x 为原始数据; μ 为计算的平均值; σ 为标准差。处理后数据符合标准正态分布,即均值为 0,标准差为 1。对于域名的统计特征和 N-Gram 模型特征,分别使用机器学习算法进行 DGA 域名检测模型的训练。本文采用的机器学习模型有朴素贝叶斯、多层感知器和极端梯度模型。

4 实验与分析

4.1 数据集

实验的数据集包括正常域名和恶意域名两部分。正常域名从 Alexa[14] 上获取排名靠前的域名。恶意域名取自 360 提供的开源 DGA 域名[15]。实验分别取 Alexa 的前 2000 条域名作为正常域名,从 360 的 DGA 域名中随机选取 2000 条作为恶意域名。模型的训练数据占 3/4,测试数据占 1/4。

4.2 实验环境

本文的实验环境如表 2 所示。

表 2 实验环境

开发环境	参数
处理器	Intel(R) Core(TM) i7-4700HQ CPU @ 2.40 GHz
内存	4 GB
操作系统	Windows 7 旗舰版
IDE	Pycharm
开发语言	Python
第三方包	seaborn、numpy、matplotlib、pandas、math、scipy、sklearn、keras

4.3 模型评价标准

本文评价 DGA 检测模型的标准为精确率 P、召回率 R 和误报率 F:

$$P = \frac{T_P}{T_P + F_P} \tag{3}$$

$$R = \frac{T_P}{T_P + F_N} \tag{4}$$

$$F = \frac{F_P}{F_P + T_、} \tag{5}$$

式中: T_P 为被模型检测为恶意域名并且检测正确的样本数量; F_P 为被模型检测为恶意域名但检测错误的样本数量; F_N 为被模型检测为正常域名但检测错误的样本数量; T_N 为被模型检测为正常样本并且检测正确的样本数量。

4.4 特征组合评估

本文提取了 4 组特征: F_1 代表 2-gram 特征; F_2 代表 1-gram、2-gram 特征组合; F_3 代表 1-gram、2-

gram、3-gram 特征组合; F_4 代表域名字符统计特征组合(包括字符长度、字符熵、元音字母比例、唯一字符数、唯一字符比例、是否是常用顶级域名) 。本文分别用朴素贝叶斯、多层感知器和极端梯度模型检测域名。为了达到较好的效果,本文采用 10 折交叉验证方法进行训练。不同特征组合下的精确率和召回率分别如图 7、图 8 所示。

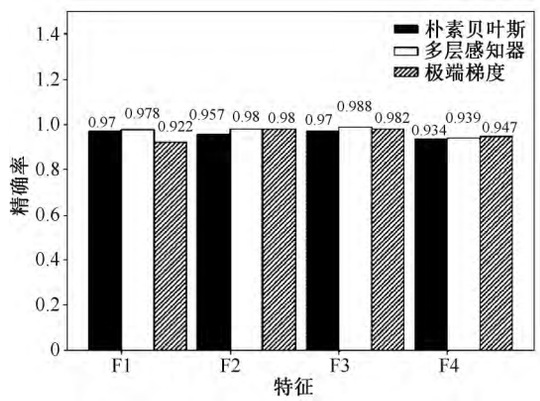

图 7　不同特征组合下的精确率

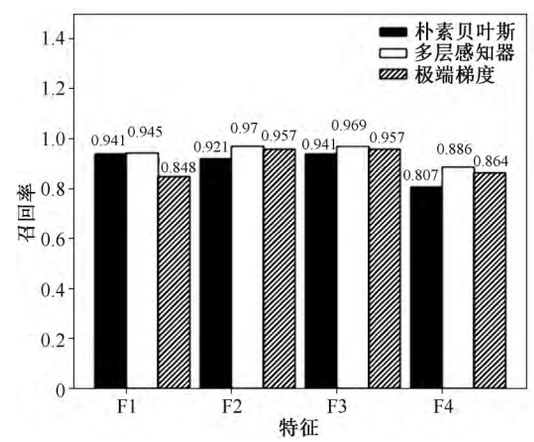

图 8　不同特征组合下的召回率

实验发现,N-Gram 模型特征比统计特征的精确率和召回率都较高。N-gram 模型中 F_3(即 1-gram、2-gram、3-gram) 特征组合的效果最佳。

4.5　模型效果评估

本文采用 ROC 曲线(receiver operating characteristic) 和 AUC(area under curve) 值作为域名检测模型的评价指标。根据上述实验,F_3 特征组合的效果最好,因而本文选用 F_3 特征组合,对比朴素贝叶斯、多层感知器和极端梯度模型,不同模型的 ROC 曲线如图 9 所示。其中,真正率(true positive rate) 为召回率,其计算方式如式(4) 所示,假正率(false positive rate) 计算方式如式(5) 所示。

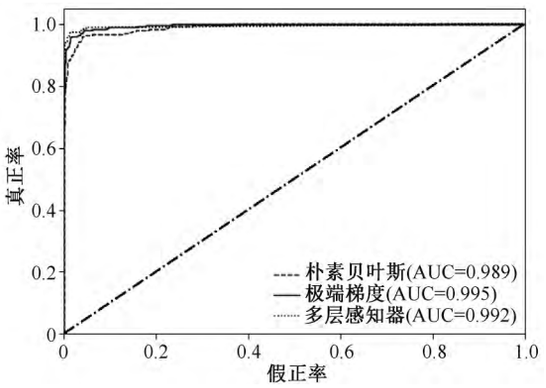

图 9　不同模型的 ROC 曲线

从图 9 可以看出,朴素贝叶斯、多层感知器和极端梯度模型的 AUC 值都在 98%以上,其中极端梯度模型的效果最好,其 AUC 值为 99.5%。

4.6　讨论

选取当前 2 个典型的检测方法(文献 [11] 和文献 [13] 所述方法) ,分别从计算量、隐私保护、是否需要相似行为方面,与本文方法进行了比较,如表 3 所示,其中√表示方法实现了该指标,×表示未实现该指标。

表 3　不同方法比较

指标	文献 [11]	文献 [13]	本文方法
计算量小	×	×	√
隐私保护	×	√	√
不需要相似行为	×	√	√

文献 [11] 是分析 DNS 流量的方法,需要采集 DNS 网络流量,并进行解析,然后分析 DNS 流量。该文基于假设: 如果是同一个 DGA 产生的恶意域名,它们的流量特征具有相似性。采集、解析和分析 DNS 流量的方法,需要较多的计算量,并且,由于直接采集网络流量,会存在隐私的问题。此外,如果被测网络中没有同一个 DGA 算法生成的域名,则不会出现相似行为,就很难检测出恶意域名。文献 [13] 是分析 DGA 域名本身的特征,提取 DGA 域名中有意义的字符比例,则需要与海量的字典单词做匹配,增大了计算量。本文的方法不需要分析 DNS 流量,避免了隐私问题和大量的计算。本文方法直接分析 DGA 域名的特征,不需要和字典匹配,也不需要被测网络中有同一个 DGA 生成的域名。综上所述,本文方法在计算量、隐私保护方面具有优势,并且不需要相似性行为也可以检测恶意域名。

5 结束语

本文目的是分析域名的特征，包括域名字符的统计特征和 N-Gram 模型特征，提取出能够区分正常域名与恶意域名的特征，以提升恶意域名的检测精确率。本文分析了 DGA 域名的特征，比较了 4 种特征组合，最终选取 1-gram、2-gram、3-gram 作为域名特征。分别评估了朴素贝叶斯、多层感知器和极端梯度模型的检测效果，实验结果表明多层感知器模型的 AUC 值最高。本文工作主要依赖于域名的字符特征，下一步工作将结合域名对应的 IP 地址，挖掘域名与域名、域名与 IP 地址之间的关联关系，并设计相应的算法检测恶意域名。

参考文献:

[1] Shun Tobiyama, Yukiko Yamaguchi, Hajime Shimada, et al. Malware detection with deep neural network using process behavior [C] // IEEE, Proceedings International Computer Software and Applications Conference. 2016: 577-582.

[2] Matija Stevanovic, Jens Myrup Pedersen, Alessandro D'Alconzo, et al. A method for identifying compromised clients based on DNS traffic analysis [J]. International Journal of Information Security, 2017, 16(2): 115-132.

[3] Shi Yong, Chen Gong, Li Juntao. Malicious domain name detection based on extreme machine learning [J]. Neural Processing Letters, 2017, 48(3): 1347-1357.

[4] Dilara Acaralia, Muttukrishnan Rajarajana, Nikos komninosa, et al. Survey of approaches and features for the identification of HTTP-based botnet traffic [J]. Journal of Network and Computer Applications, 2016, 76: 1-15.

[5] Leyla Bilge, Sevil Sen, Davide Balzarotti, et al. EXPOSURE: A passive DNS analysis service to detect and report malicious domains [J]. ACM Transactions on Information and System Security, 2014, 16(4): 1-28.

[6] Lee Jehyun, Lee, Heej. GMAD: Graph-based malware activity detection by DNS traffic analysis [J]. Computer Communications, 2014, 49: 33-47.

[7] Zang XiaoDong, Gong Jian, Mo ShaoHuang, et al. Identifying fast-flux botnet with AGD names at the upper DNS hierarchy [J]. IEEE Access, 2018, 6: 69713-69727.

[8] Zhauniarovich Yury, Khalil Issa, Yu Ting, et al. A survey on malicious domains detection through DNS data analysis [J]. ACM Computiong Surveys, 2018, 51(4): 1-35.

[9] 臧小东，龚俭，胡晓艳. 基于 AGD 的恶意域名检测 [J]. 通信学报, 2018, 39(7): 15-25.

[10] Manos Antonakakis, Roberto Perdisci, Yacin Nadji, et al. From throw-away traffic to bots: detecting the rise of dga-based malware [C] // In Proceedings of the 21st USENIX Security Symposium, 2012.

[11] Reza Sharifnya, Mahdi Abadi. DFBotKiller: Domain-flux botnet detection based on the history of group activities and failures in DNS traffic [J]. Digital Investigation, 2015, 12: 15-26.

[12] 赵越. 基于 DNS 流量特征的僵尸网络检测方法研究 [D]. 天津: 天津大学, 2016.

[13] Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, et al. Phoenix: DGA-based botnet tracking and intelligence [J]. In Lecture Notes in Computer Science, 2014, 8550: 192 - 211.

[14] Alexa [DB/OL]. (2019-04-15) [2019-05-30] https://www.alexa.com/.

[15] Netlab OpenData Project. DGA Domain List [DB/OL]. (2019-04-10) [2019-05-25] http://data.netlab.360.com/feeds/dga/dga.txt.