DOI：10.7652/xjtuxb201308010

域名请求行为特征与构成特征相结合的域名变换检测

张永斌，陆寅，张艳宁
（西北工业大学计算机学院，710029，西安）

摘要：针对僵尸网络为避免域名黑名单封堵而广泛采用域名变换技术的问题，提出一种域名请求行为特征与域名构成特征相结合的僵尸网络检测方法。该方法通过支持向量机(SVM)分类器对网络中主机解析失败的域名进行分析，提取出可疑感染主机；通过新域名聚类分析，将请求同一组新域名的主机集合作为检测对象，分析请求主机集合是否由可疑感染主机构成，提取出僵尸网络当前使用的域名集合以及命令与控制(Command and Control，C&C)服务器使用的IP地址集合。实验结果表明：训练后SVM分类器可达98.5%以上的准确率；经对ISP域名服务器监测，系统可准确提取出感染主机和C&C服务器的IP地址。

关键词：网络安全；僵尸网络；域名；域名变换
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2013)08-0054-07

Detecting Domain Flux Through Patterns of Domain Names' Alphanumeric Characters and Querying Behavior of Hosts

ZHANG Yongbin，LU Yin，ZHANG Yanning
(School of Computer Science，Northwestern Polytechnical University，Xi'an 710029，China)

Abstract：The technique of domain flux has been used by many botnets to avoid being blocked by domain blacklists. A new technique is proposed to detect botnets by analyzing the patterns inherent to domains that comprise alphanumeric characters and query behavior of hosts. The method analyzes failed domain queries through support vector machine (SVM) to identify suspicious compromised hosts. Clustering analyses are then performed to generate new successful domains and the groups of hosts that query these domains，and to examine if these host groups are composed of compromised hosts. Then，the command and control (C&C) domains and related IP addresses used by botnets are detected. Experimental results show that the accuracy of SVM prediction reaches more than 98.5% after training，and that the system can accurately detect compromised hosts and IP of C&C servers when DNS traffic from the ISP is monitored.

Keywords：network security；botnet；domain name；domain flux

域名解析服务(DNS)是互联网体系结构中重要的基础服务之一，通过DNS将抽象的IP地址映射为易于记忆的域名，可使互联网用户更加方便地访问各种网络资源。由于DNS自身缺少恶意行为检

收稿日期：2012-10-12。 作者简介：张永斌(1976—)，男，博士生；张艳宁(通信作者)，女，教授，博士生导师。 基金项目：国家自然科学基金资助项目(60903126，60872145)。
网络出版时间：2013-04-28 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20130428.0908.002.html

测能力，所以常常被利用进行各种恶意活动。例如：Bot 程序通过 DNS 服务器获取命令与控制(C&C)服务器的 IP 地址，与 C&C 服务器建立连接。虽然针对这些恶意请求行为，可以通过域名黑名单快速发现感染主机，并阻断其与攻击者之间的联系，但现在的僵尸网络广泛采取了 Domain Flux 技术，通过特定的域名产生算法(Domain Generation Algorithms，DGA)定期生成大量新域名进行请求，来对抗域名黑名单的封堵，如 Conficker[1]、Kraken[2]、Torpig[3]等。以 Conficker 为例，Conficker.C 以 24 h 为周期，采用网络时间作为域名产生种子，保证所有感染主机每个周期生成相同的域名，一个周期内生成长达 5 万条的随机域名列表。由于每个周期生成的域名不同，并且数量巨大，Domain Flux 技术有效地干扰了域名黑名单的检测和维护。因此，为了防范僵尸网络的攻击行为，针对 DGA 生成域名的检测与封堵成为网络安全研究的重要课题。

本文通过分析网络中主机的域名请求行为，实现对采用 Domain Flux 技术的僵尸网络的检测，并可提取出感染主机集合、恶意域名集合以及 C&C 服务器 IP 地址集合，同时利用 C&C 服务器 IP 地址信息，建立 IP 黑名单，对感染主机与 C&C 服务器之间的通信进行封堵。

1　相关研究

Yadav 等人针对 DGA 生成域名的文法统计规律特点，对同一 IP 地址集合或同一域名二级域上的所有域名的字符分布进行统计，采用机器学习方法对程序生成域名进行识别，并对比域名的 K-L 距离、编辑距离、Jaccard 系数分别作为特征向量的识别效果[4]。Yadav 等人在之后的研究中，通过对解析成功的域名与失败的域名进行时间相关分析、信息熵分析，检测网络中存在的域名变换行为，并提取出 C&C 服务器对应的 IP 地址[5]。Stalmans 等人采用单个域名的字符分布，以全变差距离作为特征向量，通过朴素贝叶斯分类算法对算法产生域名进行识别[6]。Jiang 等人通过对解析失败的域名请求进行分析，建立请求主机与解析失败域名之间的关系图，通过图分解的方法将 DNS 关系图分解为子图并进行研究，发现不同的子图结构往往代表不同的恶意行为[7]。Hao 等人从顶级域名服务器进行监测分析，对比了合法域名与恶意域名的请求行为的不同[8]。文献[9-11]利用机器学习的方法，分别提出了 3 种不同的域名信誉系统，对恶意域名进行检测，其中 Notos、Exposure 主要通过旁路报文捕获的方式对网络中主机域名请求流量或 DNS 服务器流量进行监测，Kopis 主要是对顶级域名服务器、授权服务器进行监测，与 Notos、Exposure 系统相比，Kopis 可以从全球视角对域名的请求情况进行监测与分析。

以上研究主要通过 DGA 算法生成域名与正常域名在字符组成分布上的不同，对 DGA 生成域名进行检测，然而国内很多域名采用汉语拼音的缩写或拼音缩写与数字结合的方式，这些域名与 DGA 算法生成域名在字符组成分布上十分相似，如果只通过域名的字符构成特征进行检测，存在较高的误报率。本文通过对 ISP 的 DNS 服务器进行长期监测与分析，发现感染主机为获取 C&C 服务器的 IP 地址，在域名请求行为上具有明显的组行为特征，即会请求大量新域名与失效域名，系统通过采集网络中主机的域名请求信息，将请求同一组新域名的主机集合作为检测对象，利用域名的字符构成特征，通过 SVM 分类器对集合内主机请求的失效域名(域名解析结果为域名不存在)进行检测，有效提取出感染主机集合、恶意域名集合和 C&C 服务器使用的 IP 地址集合。本文所采用的算法主要从感染主机的域名请求行为特征出发，并结合域名的字符构成特征进行检测，与文献[4-6]所使用的方法相比，可有效降低误报率，与文献[7]提出的方法相比算法复杂度低，同时可以快速提取出 C&C 服务器地址集合，以便建立 IP 黑名单对恶意流量进行封堵。

2　Domain Flux 特征及检测方法

2.1　Domain Flux 原理与行为特征

为提高整个感染网络的鲁棒性和抗封堵能力，Botnet 每个周期通过特定的 DGA 算法生成大量的域名。与正常域名相比，这类算法生成的域名在字符构成上存在一定的差异。同时，由于生成域名数量巨大，Botnet 的控制者仅仅只会注册其中的一个子集，从而导致感染主机需要不断进行尝试才能连接到 C&C 服务器，在 DNS 请求行为规律上，感染主机与正常主机的请求行为存在模式差异，具有明显的组行为特性。

2.1.1　*域名构成特征*　为了便于理解与记忆，合法注册的域名常常采取英文单词或知名词组的缩写，而 DGA 算法在生成域名时无需考虑这方面的问题，并且为了避免在新域名注册时发生冲突，DGA 算法生成的恶意域名往往是一些字母、数字的随机

组合，与正常域名在字符概率分布上有着显著不同。图 1 显示了合法域名的字符组成分布与 Conficker、Kraken 算法生成域名的字符组成分布(统计分析时不考虑域名的顶级域)。可以看到，正常域名字符组成分布概率曲线符合英文文法统计规律，DGA 生成域名则与正常域名存在较大差异，如 Conficker 以当前的格林威治时间作为随机种子生成当前域名列表，从而导致域名的字符分布概率趋向平均。

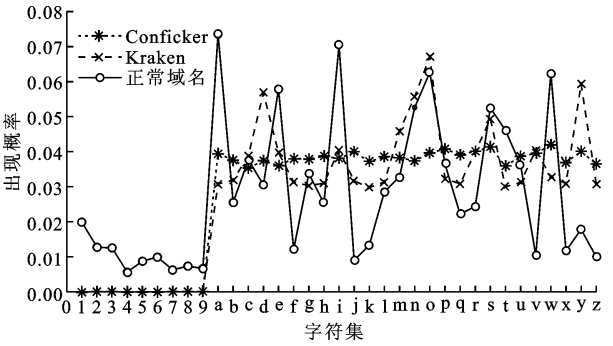

图 1 域名组成字符概率分布

2.1.2 *感染主机域名请求行为规律* Bot 程序一般以 24 h 为一个周期，每个周期均会生成大量新域名，出于经济成本的考虑，僵尸网络的控制者不可能每天注册全部生成域名。在理论上，控制者只需根据算法每天事先注册一个域名，便可实现对整个网络的控制，因此在实际使用中，控制者往往从中抽取一个子集完成注册。对于 Bot 程序，为实现与 C&C 服务器建立连接，每个周期必须生成相同的域名集合，然后按照一定规则抽取进行尝试请求，直至获取 C&C 服务器的 IP 地址。与正常主机相比，其行为模式具有显著特征，主要体现在以下两个方面。

(1)感染主机请求大量新域名，其中多数解析失败。感染主机通常采取随机请求域名生成列表中的域名进行 DNS 解析，因为如果采取了可预测的顺序进行请求，当研究人员获取恶意样本后，根据 DGA 算法，事先抢注相关域名，便可获得整个 Botnet 的控制权，对于 Botnet 的控制者来讲，这是需要极力避免的。现假设 Bot 程序的 DGA 算法每天生成长度为 N 的域名列表，Botnet 控制者每天注册其中的 M 个域名。感染主机随机从域名列表中抽取域名进行请求，则感染主机第 1 次访问到注册域名的概率为 $p_1 = M/N$，第 i 次访问到的概率为 $p_i = \left(1-\sum_{j=1}^{i-1} p_j\right)(M/(N-(i-1)))$，则 Bot 程序成功连接 C&C 服务器的平均请求域名数量为

$$E(l) = \sum_{i=1}^{N-M+1} p_i i = (N+1)/(M+1) \quad (1)$$

其中 DNS 请求解析成功的域名数量为 1，解析失败的域名数量为 $E(l)-1$。由此可以看到，感染主机在域名请求行为上会请求大量新域名，并存在一定数量的域名解析失败的行为特征。

(2)感染主机域名请求行为具有明显的组行为特性。当网络中存在多个感染相同 Bot 程序的主机时，感染主机的域名请求行为具有明显的组行为特征：首先，感染主机在域名请求行为上都会请求大量新域名，并存在一定数量解析失败的域名请求行为；其次，由于 Botnet 控制者所拥有的 IP 资源有限，那些解析成功的新域名往往指向相同的 IP 地址；最后，这些解析成功的新域名只有感染主机才会请求。

2.2 检测方法

针对感染主机域名请求行为具有以上特征，系统从失效域名与新域名两个方面对采用 Domain Flux 技术的僵尸网络进行检测，具体如下。

(1)通过失效域名分析，提取可疑感染主机。正常主机通常不会请求失效域名，但由于一些原因也会造成该行为，如访问一些已弃用的网站或 P2P 资源，而这些域名的字符构成与感染主机请求的失效域名存在一定差异，系统通过支持向量机(SVM)分类器对主机请求的失效域名进行分析，判断一个主机请求的失效域名是否主要由算法生成域名构成，从而提取出可能被感染的可疑主机。

目前，很多 CDN 网络、色情网站也采取了类似 DGA 算法的相关技术生成服务器的域名，同时国内很多域名采用汉语拼音的缩写或拼音缩写与数字结合的方式，这些域名与 DGA 算法生成域名在字符组成分布上十分相似，但这些域名往往能够正常解析，因此系统通过 SVM 分类器对失效域名进行检测，有利于降低检测误报。同时，由于正常主机请求失效域名数量相对较少，根据式(1)可知，感染主机请求失效域名数量的期望为 $E(l)-1$，因此可通过失效域名数量及 SVM 分类器检测为 DGA 算法生成域名所占的比例，进一步限制由 SVM 分类器对单个域名的误判所造成的对可疑感染主机的误判。

(2)通过新域名分析，获取 C&C 服务器 IP 地址。由于 DGA 算法生成域名往往指向相同的 IP 地址，系统通过新域名解析返回的 IP 地址对新域名进行聚类分析，将新域名划分为不同的域名集合，分析每个新域名集合的请求主机组是否主要由失效域名分析所提取的可疑感染主机构成，从而提取出最

终感染主机集合、恶意域名集合、C&C 服务器 IP 地址集合。由于 Bot 程序具有很强的传染能力，在 ISP 网络内往往存在几十台甚至上百台感染主机，其组行为特征更为明显，因此可通过请求主机组的规模以及可疑感染主机所占的比例，进一步规避由于对单个感染主机的误判所造成的对整个检测结果的误报。

3　算法设计与实现

整个检测算法分为数据预处理、感染主机分析、组行为分析、C&C 服务器 IP 地址集合可信度计算 4 个处理过程。其中数据预处理过程主要是通过域名白名单对可信、合法的域名进行过滤，降低后续处理过程的数据分析量，同时对网络中成功解析的域名进行采集、保存，建立完善的域名数据库和检测基线，以判断一个解析成功的域名网络中主机是否请求过，即是否为新域名，通过该过程系统提取出所有请求新域名的主机及请求的新域名。感染主机分析过程主要是通过 SVM 分类器对主机请求的失效域名进行分析，提取出所有可疑感染主机。组行为分析过程主要是对新域名进行聚类，将新域名划分为不同的域名集合，利用感染主机分析过程的结果，对每个新域名集合的请求主机组进行分析，以便提取出感染主机集合、恶意域名集合、C&C 服务器 IP 地址集合。由于 C&C 服务器 IP 地址集合对应的所有域名与感染主机请求的失效域名源自于同一 DGA 算法，可信度计算过程实质是通过已建立的 SVM 分类器对 C&C 服务器 IP 地址集合上的所有域名进行分析，来计算 DGA 生成域名所占的比例。通过可信度计算，可以有效排除 DGA 算法生成域名与已有域名冲突时造成的误判。

3.1　数据预处理

系统通过白名单对知名域名、CDN 网络上的域名进行过滤，同时为了判断出一个域名网络中主机是否请求过，在检测之前，系统首先对网络中所有主机成功请求过的域名进行收集，通过一段时间的学习后，形成一个相对完善的域名库，以便建立一个检测基线。在检测阶段，如果网络中主机请求的域名不在域名基线库中，则认为主机请求了一个新域名。由于互联网资源访问服从幂律分布，通过一段时间学习后，每天出现的新域名数量相对有限，同时产生新域名请求的主机数量也相对较少。

3.2　感染主机分析

由于 DGA 算法生成域名与正常域名在字符组成分布上存在相对明显的差异，系统采用 SVM 文本分类技术，通过 bigram 方法对每个域名的字符构成进行分析（分析时不考虑注册域名的顶级域），以每个 bigram 的频率作为分类特征，采用线性 SVM 分类器，判断主机请求的每个失效域名是否为 DGA 算法生成域名。系统默认周期为 1 d，在每个检测周期，系统首先对语法错误的失效域名（如包含非法字符的失效域名、使用不存在的顶级域的失效域名）进行过滤，并通过 IP 地址对每个主机进行标识，将请求主机集合记为 $H=\{h_1,h_2,\cdots,h_k\}$。由于是对运营商的 DNS 服务器进行监测，NAT 技术、代理技术在互联网建设中广泛使用，因此一个主机域名的请求行为可能是多个主机的混合行为。设主机 h_i 请求的失效域名的数量为 n_i，其中 SVM 分类器检测为 DGA 算法生成域名的数量为 m_i，则可疑感染主机判定条件为 $m_i/n_i\geqslant\mu$ 或 $m_i\geqslant M$，其中 μ、M 为检测阈值。在实际实验时，选取 $\mu=0.6$、$M=10$。

3.3　组行为分析

经过数据预处理后，将每个检测周期内主机成功解析的新域名集合记为 $D=\{d_1,d_2,\cdots,d_n\}$，R_i 为域名 d_i 解析返回的所有 IP 地址集合。由于感染主机通常随机访问域名生成列表中的域名，从而导致感染主机成功连接 C&C 服务器所使用的域名不一定相同，但这些新生域名往往都指向相同的 IP 地址。定义域名间关系图 G，图中节点 i 为域名 d_i，对 G 中任意两个节点 d_i、d_j，如果 $R_i\cap R_j\neq\varnothing$，则认为两个节点间存在边，否则不存在边。对于 G 中任意一个连通子图 G_k 称为域名关联集合 C_k，C_k 中所有域名对应的 IP 地址集合记为 I_k，请求主机集合记为 Q_k（Q_k 中任意主机至少请求过 C_k 中一个域名）。通过连通子图特性可知，对于 C_k 中任意一个域名 d_i，其满足如下关系：$\forall d_i\in C_k$，则 $\exists d_j\in C_k$，$i\neq j$，$R_i\cap R_j\neq\varnothing$。通过对 G 的连通子图进行分析，可有效地将网络中请求新域名的主机划分为不同的请求组，以便进一步分析每个请求组内主机的构成。为得到 G 中所有连通子图，可通过由下至上的层次聚类方法对 D 中域名进行聚类分析，聚类使用的距离函数定义如下

$$\mathrm{dist}(C_i,C_j)=\begin{cases}0, & I_i\cap I_j\neq\varnothing\\ +\infty, & I_i\cap I_j=\varnothing\end{cases}\tag{2}$$

对于任意域名关联集合 C_k，如果请求 C_k 的主机组 Q_k 中大量主机为可疑感染主机，则 C_k 为 Botnet 使用的域名集合的可能性很高。据此，定义可疑度算子

$$S_k = |M_k| / |Q_k|$$

式中：M_k 为 Q_k 中由感染主机分析过程判断为可疑感染主机的主机集合。由于 Bot 程序具有很强的传染能力，在一个网络内往往存在多台感染主机，因此判定 C_k 中域名为 Botnet 使用的域名集合条件为 $|Q_k| \geqslant \gamma$ 且 $S_k \geqslant \delta$，其中 γ 为 Q_k 中主机数量阈值，δ 为可疑度 S_k 阈值。经实验，取 $\gamma=10$、$\delta=0.75$ 具有较好的验证效果，可以有效排除小型网站、P2P 应用等使用的新生域名。

3.4 IP 地址可信度计算

当 C_k 判定为恶意域名集合时，I_k 不一定是 C&C 服务器的 IP 地址。例如：DGA 算法生成域名与互联网已注册域名冲突时，则域名对应的 IP 地址并不是 C&C 服务器的 IP 地址。如果 I_k 为 C&C 服务器的 IP 地址，则所有解析返回的 IP 地址属于 I_k 的域名集合 P_k 中大量域名应为 DGA 算法生成域名。通过 3.2 节中构造的 SVM 分类器对 P_k 进行检测，恶意地址的可信度定义为 $p_k=|O_k|/|P_k|$，O_k 为 P_k 中判断是 DGA 算法生成的域名集合。经实际测试，当 p_k 大于 0.8 时，可认为 I_k 为 C&C 服务器所在的恶意 IP 地址。

4 实验验证

实验数据主要源于某地市 ISP 的 DNS 服务器集群，在上网高峰期时段，该 DNS 服务器集群接收 DNS 请求平均 1.2 万次/s，服务近 20 万用户。系统采集了 45 d 的域名请求信息，并建立域名基线数据库，其中包含 869 万个域名、1 900 多万条域名与 IP 地址对应信息。基线建立后，随机选取了连续 5 d的监测数据进行检测与分析，取得了较为满意的结果。

4.1 数据集分析

由于实验分析数据过多，故仅以第 5 d 的监测数据为例进行说明。

4.1.1 域名白名单过滤效果 系统通过域名白名单可滤掉 90%左右的域名请求(参见图 2)，在凌晨 0～5 点之间，用户单位小时内域名请求次数是全天最低的一个时间区段，此时白名单的过滤率也相对较低。分析造成原因是：在该时间段内上网用户数量较少，因此用户个性化行为特征更为明显。

4.1.2 失效域名与新域名分析 系统对每小时内请求失效域名的 IP 数量进行监测，发现每小时内请求失效域名 IP 数量占该小时内请求 IP 地址总量的 5%～20%(参见图 3)。在当天请求失效域名的 IP 地址总数量为 64 579 个，失效域名的数量为 145 593个，其中大量主机只请求了一个失效域名(累积分布曲线参见图 4)，近 90%的主机请求失效域名数量不超过 5 个，经 SVM 分类器检测，符合 2.2 节中检测条件的 IP 地址仅有 381 个。

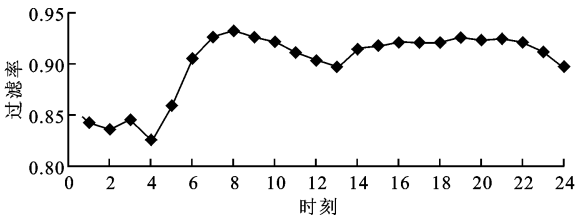

图 2 白名单过滤率

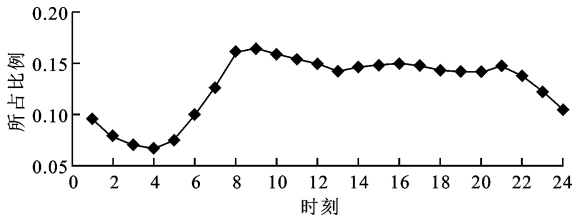

图 3 请求失效域名主机所占比例

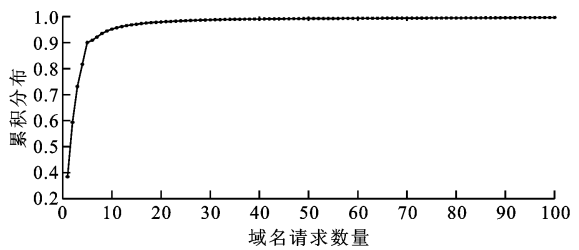

图 4 主机请求失效域名数量的累积分布

当天请求新域名的 IP 地址数量为 150 981 个，新域名的数量为 119 427 个，其中当天新出现域名的数量为 25 877 个。从域名请求主机数量上看，大量域名只有一个 IP 请求，经聚类分析后，得到 34 558组域名集合，其中 28 463 组仅有一个 IP 访问。符合 3.3 节中检测阈值 $\gamma \geqslant 10$ 条件的域名组仅有626 个。

4.2 分类器的训练与测试

每天主机请求的失效域名集合中包含很多已弃用或注销的合法域名，如 data. vrs. sohu. com、nccpr. p2p. baofeng. net、btrouter. sandai. net 等。同时，很多 P2P 种子文件中包含一些已经不可用的资源服务器信息，也会造成正常主机请求大量失效域名的现象。由于这些域名与正常域名在字符分布上十分相近，同时每天失效域名的数量巨大，包含大量不知名域名，很难对这部分域名进行标注。因此，在实际训练中，通过白名单提取出一些正常域名作

为反例训练样本，具体主要通过以下两个方法进行提取：①通过 Alexa 网站排名数据库[12]，在域名基线数据库中提取属于排名前 1 万的域名；②对收集到的域名信息手动分析，提取出了 30 个 CDN 网络，再对 CDN 网络上的域名进行提取。在通过这两种方法提取的域名中，随机抽取了 1 万个域名作为正常域名样本。恶意域名样本主要包含 Conficker（A、B、C 三种变种）、Kraken、Bobax、Simda、BankPatch、Zeus（两个不同变种）生成的域名，其中 Conficker 域名通过工具 Downatool2 产生[13]，Kraken 的域名主要源于 Damballa 公司发布的数据[4]，其余恶意软件的域名主要是通过将恶意软件样本植入到虚拟机中分析获得。

为检验 SVM 分类器的准确率及通用性，分类器训练了两次，并用未参与训练的样本进行测试。在第 1 次样本训练时，恶意域名样本中只选取了 Conficker. A、Conficker. B 两种样本，第 2 次样本训练时，在原有恶意样本中加入了识别效果不好的 Kraken、Bobax、BankPatch 样本各 100 个，具体结果参见表 1。表 1 测试结果表明，SVM 分类器具有很高的准确率，对比两次测试结果可知，SVM 分类器对域名分析部分中“随机部分”的长度十分敏感，即算法生成部分，如果该部分长度占域名分析部分较大比例时，即使训练时，未包含该类样本也可准确识别。例如 Zeus 两个不同变种、Simda 等，如果算法生成部分所占比例较小，只需加入少量样本进行训练，便可迅速提高检测的准确率，如 Kraken、Bobax、BankPatch 等。

表 1　SVM 分类器的检测准确率

类别	域名样例	准确率/%	
		第 1 次	第 2 次
Conficker. A	zhckf. info，rgxxkjzn. com，vqquqnjf. biz …	98. 60	98. 60
Conficker. B	efsmqizcqjo. cn，sissq. info，vygcctncgl. org …	99. 20	99. 20
Conficker. C	cuzgq. com. fj，llemuny. nf，yrkcl. com. ng，oicjkneed. kn …	98. 30	98. 50
Kraken	bltjhzqp. dyndns. org，cffxugijxn. yi. org …	32. 63	100. 00
Zeus 变种 1	zxxldfmgeqclzozhytqinmjzx. info，zxylwcyhpfeojlbpvgxhsweiqw. biz …	100. 00	100. 00
Zeus 变种 2	a47avm29ewlujypqi15htoslsc39ird50bsgx. com，a47e21prn60pxnsatj26awatjukqjseri35dy. biz …	99. 80	99. 80
Simda	puzugeq. info，puvuteq. info，pujyxyg. info …	99. 40	99. 60
Bobax	vmcmvsaxro. 2mydns. net，xrpxmxira. widescreenhd. tv …	25. 77	99. 20
BankPatch	blmk. sinkhole. dk，kwmi. sinkhole. dk …	0. 00	100. 00
10 万个正常域名		99. 43	98. 50

4.3　实验结果

完成基线检查数据库和分类器构造后，系统选取了连续 5 个工作日的域名请求信息。由于实验分析数据过多，在文中主要讨论 $|Q_k|$ 分别取阈值 5 和 10 两种情况，其他条件为 $\mu=0.6$，$M=10$，$S_k\geqslant 0.6$。

（1）当 $|Q_k|$ 取阈值为 10 时，共提取出 11 组可疑域名集合，解析返回地址对应 3 组 IP 地址集合，其中 2 组 IP 地址集合连续 5 d 都出现，可信度达 0. 984 以上，1 组只出现了 1 d，可信度只有 0. 1（参见表 2）。对每个检测周内请求可疑域名集合的主机集合进行分析，发现基本为同一组主机，即这 11 组可疑域名集合是由同一 Botnet 产生的，经确认为

表 2　Conficker 生成域名分析结果

域名	IP 地址	时间/d	可信度
uskrtr. cn，pnyoedtj. cn，eibmdfzhaks. cn …	221. 8. 69. 25	5	0. 984
wqmkyfdee. org，jobgp. info，xeeecu. info，papgrsw. biz，vncvocbhhu. org …	143. 215. 130. 33，143. 215. 143. 11，149. 20. 56. 32，149. 20. 56. 33，149. 20. 56. 34	5	0. 993
roish. com	74. 117. 116. 65	1	0. 100

Conficker 生成的域名。在监测期间,请求这些域名的主机最多时达 467 个不同 IP 地址,感染主机规模十分庞大。对可信度较低的 IP 地址集合进行分析发现,该域名对应的网站是合法网站,判断被提取出的原因是 DGA 算法生成域名与已有域名冲突造成的。

(2)当$|Q_k|$阈值为 5 时,除 11 组 Conficker 对应的域名,还提取出 8 组可疑域名集合,可信度均为 0,经分析这 8 组域名均为误报,造成原因是部分感染 Conficker 的主机请求了相同的新域名。

5 结束语

本文主要通过分析主机解析成功及失败的域名,对采用 Domain Flux 技术的 Botnet 进行检测,通过实际环境的验证,可以有效地检测出感染主机集合及 C&C 服务器使用的 IP 地址集合。在未来研究中,将进一步完善算法,以便提取出未知类型的 DGA 生成域名。

参考文献:

[1] LEDER F, WERNER T. Know your enemy: containing conficker [EB/OL]. [2011-03-05]. http://www.honeynet.org/papers/conficker.

[2] ROYAL P. On the kraken and bobax botnets [EB/OL]. [2010-09-10]. http://www.damballa.com/downloads/r_pubs/Kraken_Response.pdf.

[3] STONE-GROSS B, COVA M, CAVALLARO L. Your botnet is my botnet: analysis of a botnet takeover [C]//Proceedings of the 16th ACM Conference on Computer and Communications Security. New York, USA: ACM, 2009:635-647.

[4] YADAV S, REDDY A, REDDY A. Detecting algorithmically generated malicious domain names [C]// Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. New York, USA: ACM, 2010:48-61.

[5] YADAV S, REDDY A. Security and privacy in communication networks [M]. Berlin, Germany: Springer, 2012:446-459.

[6] STALMANS E, IRWIN B. A framework for DNS based detection and mitigation of malware infections on a network [C]//Proceedings of the IEEE Information Security South Africa. Pretoria, South Africa: ISSA, 2011:1-8.

[7] JIANG N, CAO J, JIN Y. Identifying suspicious activities through DNS failure graph analysis [C]// Proceedings of the Eighteenth IEEE International Conference on Network Protocols. Kyoto, Japan: ICNP, 2010:144-153.

[8] HAO S, FEAMSTER N, PANDRANGI R. An internet-wide view into DNS lookup patterns [EB/OL]. [2010-06-16]. http://www.verisigninc.com/assets/whitepaper-dns-lookup-patterns.pdf.

[9] ANTONAKAKIS M, PERDISCI R, DAGON D. Building a dynamic reputation system for DNS [C]// The Proceedings of 19th USENIX Security Symposium. Berkeley, California, USA: USENIX Association, 2010:273-289.

[10] ANTONAKAKIS M, PERDISCI R, LEE W. Detecting malware domains at the upper DNS hierarchy [C]//Proceedings of the 20th USENIX Security Symposium. Berkeley, California, USA: USENIX Association, 2011:27-27.

[11] BILGE L, KIRDA E, KRUEGEL C. Exposure: finding malicious domains using passive DNS analysis [C]//Proceedings of the 18th Annual Network & Distributed System Security Conference. San Diego, USA: NDSS, 2011:1-17.

[12] Alexa Com. The Web information company [EB/OL]. [2008-05-18]. http://www.alexa.com/topsites.

[13] LEDER F, WERNER T. Containing conficker tools and infos [EB/OL]. [2011-06-08]. http://net.cs.uni-bonn.de/wg/cs/applications/containing-conficker.

[本刊相关文献链接]

脱立恒,倪宏,刘学. 一种网络冗余流量消除算法. 2013, 47(4):22-27. [doi:10.7652/xjtuxb201304005]

夏秦,王志文,卢柯. 入侵检测系统利用信息熵检测网络攻击的方法. 2013,47(2):14-19. [doi:10.7652/xjtuxb201302003]

伍文,孟相如,马志强,等. 模块化动态博弈的网络可生存性态势跟踪方法. 2012,46(12):18-23. [doi:10.7652/xjtuxb201212004]

杨柳静,秦涛,王晨旭. 应用交互式网络流模型的高速网络异常行为检测与控制方法. 2012, 46(6):58-65. [doi:10.7652/xjtuxb201206011]

夏秦,王志文,刘璐. 基于域名共现行为的僵尸网络行为追踪. 2012,46(4):7-12. [doi:10.7652/xjtuxb201204002]

胡鹤,胡昌振,姚淑萍. 应用部分马尔科夫博弈的网络安全主动响应决策模型. 2011,45(4):18-24. [doi:10.7652/xjtuxb201104004]

(编辑 荆树蓉 武红江)