算法生成恶意域名的实时检测*

张雪松　国家计算机网络应急技术处理协调中心工程师
徐小琳　国家计算机网络应急技术处理协调中心高级工程师
李青山　北京大学网络与软件安全保障教育部重点实验室博士研究生

摘要：当前对算法生成域名技术的检测，检测所用时间周期过长，无法对算法生成的恶意域名进行快速检测。针对此问题，本文基于新增域名与已分类恶意域名之间的关联关系，提出一种算法生成域名的实时检测方法，并在某省运营商DNS服务器机房部署本系统，实验验证本检测方法。实验表明与已有方法相比，本方法能够快速筛选用于恶意网络行为的算法生成域名。但本方法需要消耗大量的计算资源和内存资源，需要在后续的工作中研究解决。

关键词：域名生成算法，僵尸网络，算法生成域名，域名变换

Abstra ct：The current detection of algorithmically generated malicious domain has long detecting cycle, and has no sufficient use of known and published malicious domain. Aimed at this problem, we propose a real time detect method based on relationship between new domain name and known malicious domains combined with live pattern of algorithmically generated domains. The experiment result shows that the method can quickly and effectively filter out algorithmically generated domains used by malware, but this method consumes a large amount of computation and memory resources that needs to be solved in the following work.

Keywords: domain generation algorithms;botnet; algorithmically generated domain;domain-flux

1 引言

域名系统是当前互联网重要的基础设施之一，大量的网络服务依赖于域名服务来开展。由于域名系统并不对依托于其开展的服务行为进行检测，很多恶意网络活动也利用域名服务来完成。

近些年，大量的恶意软件引入域名算法生成技术，利用特定的域名生成算法（DGA：Domain Generation Algorithms），生成大量域名用于自身的组织和控制，以提升自身的生存能力，延长系统的生存时间。不同的恶意软件，由于域名生成算法不同，对外会显现出不同的利用特征。以 Conficker[1-2]为例，Conficker.C 在一天内，利用网络时间作为种子，随机生成 5 万多个域名，分布于一百多个顶级域名上。而像 2012 年公布的恶意利用软件 Blackhole，其以所在主机的时间戳作为种子，每 12 个小时仅生成一个新的二级域名。

域名算法生成技术也被称为域名变换技术（Domain Flux）[3]，通过算法生成大量域名，并注册其中的部分域名，自从 2004 年被引入到僵尸网络，迄今已被众多恶意软件所使用，外在表现为指向一个或同一组 IP 资源的域名不

* 基金项目：国家 242 信息安全计划基金资助项目（242-2010A009）。

断变化。当前存在多种针对域名算法生成技术的检测方法，这些方法从多个角度进行检测，如针对生成域名的字符分布、域名的客户端访问、域名流量特征等。但是，由于算法生成的域名变化非常迅速，这些方法对新增域名的实时性检测不够关注，检测周期相对较长，无法快速检测新增域名。

本文基于新增域名与已分类恶意域名之间的关联关系，提出一种算法生成域名的实时检测方法，用于快速筛选用于恶意网络行为的算法生成的域名。

2 相关研究

针对域名算法生成技术的研究，早期主要基于逆向和蜜罐技术。基于逆向工程，可以深入了解恶意样本采用的域名生成算法以及对域名的使用机制，并可在此基础上采取有针对性的措施。如 Santa Barbara[4]等人对 torpig 样本进行逆向分析后，得到样本所使用的域名生成算法。他们通过提前抢注域名，并部署假的 C&C 服务器，成功控制了 torpig 大约 10 天的时间。

由于逆向需要消耗大量的人力资源，并且分析周期较长，逆向的应用受到很大的限制。后续针对域名算法生成技术的检测，主要集中在两个方面，一是域名的结构及字符特征分布，另一方面是主机在使用域名时产生的域名流量特征。

文献 5 提出一种方法，用于在 DNS 流量中检测域名变化行为。针对算法生成域名与正常域名的字符分布的不同，分别利用域名的 K-L 距离、编辑距离、Jaccard 系数分别作为特征向量的识别效果，采用机器学习方法对算法生成域名进行识别。

一些恶意的网络行为会引起大量的的 DNS 请求解析失败，如基于 fast-flux 的 web 服务、木马或僵尸网络等恶意软件连接与控制等[6]。因此，文献 7 通过采集失败的 DNS 请求流量，抽取主机与解析失败域名之间的关系图。利用扩展的 tNMF[8]算法，对关系图进行递归分解，抽取其中的稠密子图，分类并寻找这些子图与各类恶意活动之间的对应关系。

Damballa 的 Antonakakis 等人认为，同一个 botnet 内，各个 bot 会产生一致的不存在域名访问流量，可以利用不存在域名 (Non-Existent 类型) 流量数据检测随机生成的域名，采用分类与聚类相结合的方式，对大型的不存在域名集合进行分析[9]。

当前针对算法生成域名的检测，主要通过收集整理一段时间内的域名，并对域名进行分组、分类和聚类分析后，确定域名及分组的属性。而基于算法生成的恶意域名，其域名出现和更新的速度很快，现有算法无法及时对新出现域名进行快速分析。

3 算法设计与实现

如图 1 所示，整个检测算法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。

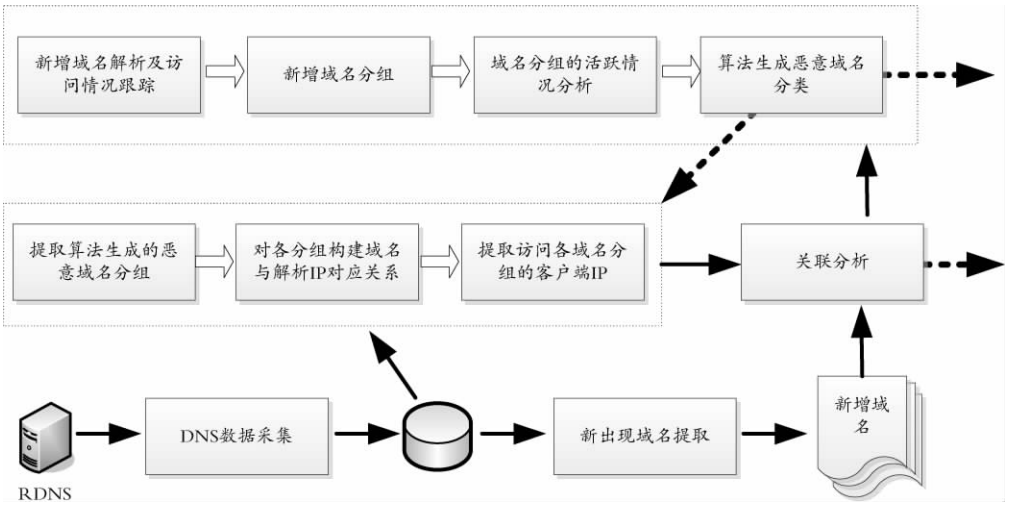

图 1 算法处理流程

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。

3.1 数据采集及预处理

在整个域名系统中，域名缓存服务器（RDNS）是沟通用户与权威服务器的桥梁，整个系统大部分的请求及解析流量都会经过它。由于数据流量过大，且大部分流量为正常使用流量，需要对数据进行过滤。利用白名单机制对知名域名进行过滤，建立域名数据库，经过一段时间的学习后，建立域名检测基线，以划分并提取新出现域名。

3.2 已知域名的解析及访问关系构建

提取已知算法生成的恶意域名，建立相关域名的解析及访问对于关系，以便于后续与新增域名间的关联分析，确定新增域名与已知域名的从属关系。

本课题中已知恶意域名数据来源于两个方面，一是国内外相关研究机构或组织发布的域名黑名单列表，另一个则是利用本课题前期检测并验证的域名列表。

依赖采集到的缓存服务器镜像数据，提取已知域名数据的解析及访问资源记录，在内存中建立域名或域名分组解析及访问对应关系。将一个已知域名分组用 D 表示，则 $D-\{d_1,d_2...,d_n\}$ ，每个域名分组包含至少一个域名。提取分组内各域名的解析 IP，合并为集合 $P_D-?_iP_{di}$。提取访问域名分组的客户端 IP，建立域名分组的访问客户端 IP 集合 $Q_D-?_iP_{di}$。

3.3 域名关联分析

恶意软件利用域名算法生成技术，主要用于达到两个目的：一是通过生成大量的域名迷惑安全系统，提高安全系统检测和分析的代价；二是利用域名系统进行资源的访问与控制。典型的一些应用场景包括 被感染主机定位 botnet 的命令控制服务器，被控主机连接恶意软件加载站点，恶意软件访问存储偷取数据的服务器等。

由同一算法生成的域名，其在域名解析、域名访问及域名的构成等方面存在一致性，可依靠此关联关系，对新出现域名进行检测。这些关联关系主要包括三个方面 域名结构、域名解析数据集合以及域名的访问客户端集合。

由同一恶意软件生成的域名，由于采用同一生成算法，这些域名具有相同的域名结构及字符分布。以某一算法生成的恶意域名分组为例，其域名的前 3 级的域名前缀相同，第 4 级域名呈现相同的长度和字符分布，如表 1 所示。

表 1 某算法生成域名

域名
aa847932bb37c7ce6e3f815fa3c1d850128deff9.107.vfrtg.com
c9693728a25e6142e836d207e5f82fd36943fafa.107.vfrtg.com
566e5d445a5a201daec96b105f72b448f002c3ff.107.vfrtg.com
fc5609a875f259b7ae186d9e5a5527960b757bf9.107.vfrtg.com
42a3d5d5469a5c5909506a2628ce812fbecfbbff.107.vfrtg.com
a14a14ae1b694cd9f9d7f517e4f42a9e5c1d815c.107.vfrtg.com
42a3d5d5469a5c5909506a2628ce812fbecfbbff.107.vfrtg.com

域名间的相似程度可以采用编辑距离来描述。编辑距离（Edit distance）用于描述两个字符串之间，由一个转成另一个所需的最少编辑操作次数。如果新增域名与集合内各域名的编辑距离数值非常接近，则可认为此新增域名与集合内域名具有相同的字符分布。

将某一已知的算法生成恶意域名分组记为 $D=\{d_1,d_2,...,d_n\}$，新出现域名记为 d，则从下面几个条件确定 $d \in D$。

$$\left.\begin{array}{l} t(d) \in t(D) \\ len(d)=\dfrac{\sum_{i=1}^{n} len(d_i)}{n} \\ di(d,d_i)-di(d,d_j)<\delta(i \neq j) \end{array}\right\} \Rightarrow d \in D \quad (1)$$

t(d)表示取域名或域名集合的域名前缀，公式 1 中条件 1 表示新增域名的前缀包含在已有域名分组的前缀集合内。len(d)为取域名的字符串长度，新增域名的长度需要与已知域名集合内各域名的平均长度一致。di(m,n)用于得到两个字符串间的编辑距离。如果新增域名与已知域名集合都是由同一算法生成，则新增域名与集合内任意域名的编辑距离会在一个固定的数值上下波动。判定域名字符串的相似程度，也可采用 K-L 距离（Kullback-Leibler)或 Jaccard 系数等，考虑到系统的实时性检测需求，采用计算相对简单的编辑距离。

同一恶意软件生成的域名，由于拥有相同的系统资源，这些域名的解析 IP 会指向同一 IP 集合。虽然由于实际部署的限制，无法获取或捕获到所有的解析 IP，但可以依据捕获的部分数据计算域名间解析 IP 的重合程度，进而确定域名与已知恶意域名分组的关系。

从图 2 可以清晰得看到域名与解析 IP 间的关系。

对于新增域名 d，如果其解析 IP 集合 P_d 包含于 D 的域名解析 IP 集合 P_D 中，则可认为新增域名 d 属于 D，即

$$P_d \in P_D \Rightarrow d \in D \quad (2)$$

如果新增域名有少量的解析 IP 不属于 P_D，则需要依靠解析 IP 交集的情况对域名与域名集合的关系进行判定。如果新增域名解析得到的 IP 绝大部分属于 P_D，则可认为新增域名属于已知域名分组 D，即：

$$\frac{|P_d \cap P_D|}{|P_d|} \geq \alpha \Rightarrow d \in D \quad (3)$$

同一恶意软件生成和使用的域名，由受控网络内的所有感染主机使用。通过提取建立域名分组域访问客户端 IP 集合，按照上述公式 3 对比新增域名与已知域名分组的客户端 IP 集合，来确定新增域名是否属于已知域名分组。

上述三个关联特征相对独立，如果仅依赖某一关联特征，会造成域名检测的漏报。如果生成的恶意域名分布在多个二级域名上，则仅依靠域名的结构特征无法有效检测出新增的二级域名。基于这种考虑，当新增域名符合其中任意一个关联特征，则认为域名属于已知域名分组。

3.4 未关联域名的分类

如果无法通过关联分析，确定新增属于已确认的恶意域名分组，系统将对新增域名进行持续跟踪，收集新增域名的相关访问及解析数据，并按周期，对新增域名进行分组，基于分组的域名活跃特征，确定新增域名或分组是否属于恶意的算法生成域名。

同一恶意软件生成的域名，其使用特点决定，每个域名的活跃周期只占恶意软件整个生存周期的一部分。对于某一域名，当其使用周期过后，即使还能够解析到域名数据，但不再有客户端使用它。算法生成域名的使用如图 3 所示。

以 5 分钟为周期分割时间 T_d，对属于恶意软件的每一个域名，提取其被请求的时间集合，其中，$T_{d1} \cup T_{d2} \cup ... \cup T_{d3}$ 为此恶意软件的活动周期。按照下述条件分类域名分组是否为算法生成的恶意域名：按照时间顺序对各域名出

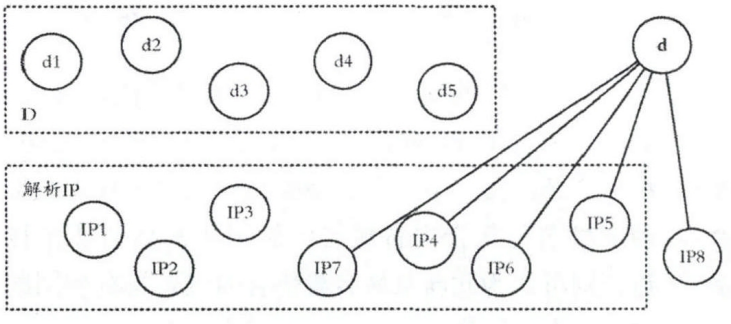

图 2 域名间解析 IP 对应关系图

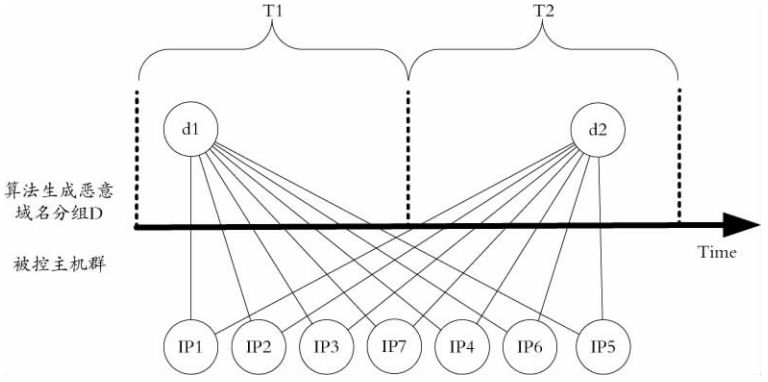

图 3 算法生成域名的使用方式

现的时间集合进行排序，并计算相邻集合间的交集程度：$\frac{|T_{d1} \cap T_{d2}|}{|T_{d1} \cap T_{d2}|} \leq \beta$，当域名分组内域名的交集都小于设定的 β 时，认为域名分组 D 为算法生成。

4 实验及分析

系统部署在某省运营商的缓存服务器侧，通过镜像方式获取 DNS 数据。系统的部署结构如图 4 所示。

首先，系统收集整理多个安全组织和机构公布的算法生成恶意域名列表，作为初始外部数据库。同时，系统连续采集一周的域名数据，建立域名数据库作为判别新增域名的数据基础。系统输入的 DNS 镜像数据流量平均为 10 万条 /s，按照 Alexa 提供的知名网站进行过滤后 DNS 请求流量约为 2000 条 /s。

实验中，通过调整 δ、α 和 β 输出域名，当条件设定为 $\delta \leqslant 5$，$\alpha \geqslant 0.9$，$\beta \leqslant 0.1$ 可达到较好的检测效果。其中，漏报率小于 0.3%，误报率较高，约为 12%。通过对误报结果进行跟踪分析，发现误报域名主要是由未关联到的新增域名引起。通过手动收集误报域名的相关信息，包括解析 IP、web 站点、端口开发情况等，误报域名主要包括以下一些用途 游戏网站登录、聊天工具的下载以及网盘类网站的用户访问。其中，大部分域名采用了泛域名解析技术，利用这项特征对结果进行过滤后，系统误报率降到 5%以下。

对于系统观察到的新增域名，如可在已知数据中关联到相关域名分类，域名检测结果的输出时间小于半小时。当无法在已知数据中关联到结果后，新增域名缓存到内存中，并定期进行域名的分组，以天为单位，计算域名的活跃情况，判别域名分类。

表 2 为按照时间列出某恶意软件使用的部分域名。

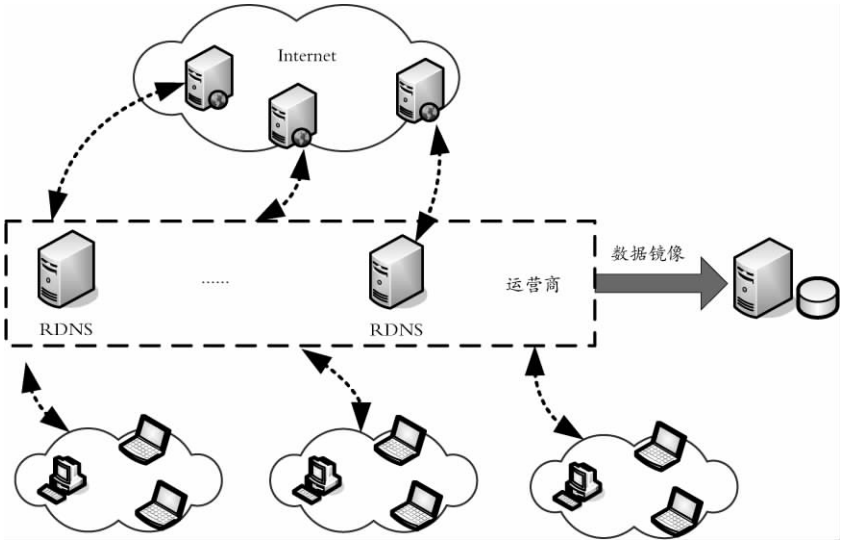

图 4 系统部署示意图

表 2 某恶意软件使用的部分域名

域名
19.vfrtg.com
49.okjyu.com
17.vfrtg.com
6.vfrtg.com
49.ujhvg.com
66.ujhvg.com

5 结束语

本文利用已确认的恶意域

名数据，基于新增域名与已分类恶意域名之间的关联关系，提出一种算法生成域名的实时检测方法，对新增域名进行快速检测。通过在某省运营商机房部署本系统，实验并验证了本检测方法。与已有方法相比，本方法能够快速筛选用于恶意网络行为的算法生成域名。但本方法在系统运行时需要消耗大量的计算资源，且内存资源占用随着系统运行而不断增长，这需要在后续的工作中研究解决。 MSTT

参考文献

[1]P Porras, H Saidi, V Yegneswaran. A Foray into Conficker's Logic and Rendezvous Points[R]. SRI INTERNATIONAL,2009:10-11.
[2] Leder F, Werner T. Know Your Enemy: Containing Conficker [R], The Honeynet Project & Research Alliance, University of Bonn, Germany, 2009.
[3] 江健，诸葛建伟，段海新等. 僵尸网络机理与防御技术[J]. 软件学报, 2012, 23 (1) : 82-96.
[4] Stone-Gross, B Cova M, Vigna G. Your Botnet is My Botnet: Analysis of ABotnetTakeover [C]. in ACM Conference on Computerand Communications Security (CCS) , 2009, 635-647.
[5] Yadav S, Reddy, Ranjan S. Detecting Algorithmically Generated Malicious Domain Names [C]. in 10th Annual ACM Conference on Internet Measurement, New York, USA, 2010, 48-61.
[6] Z Zhu, V Yegneswaran, Y Chen. Using failure information analysis to detect enterprise zombies[J], in SecureComm'09, 2009.
[7]Jiang N, Zhang Z. Identifying Suspicious Activities through DNS Failure Graph Analysis [R]. in Network Protocols (ICNP) , the 18th IEEE International Conference, 2010, 144-153.
[8] Y Jin, E Sharafuddin, Z-L Zhang. Unveiling core networkwide communication patterns through application traffic activity graph decomposition [C], in SIGMETRICS '09, 2009.
[9] Manos Antonakakis, Roberto Perdisci, YacinNadji, et al. From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware [C], in the 21th USENIX Security Symposium, Bellevue, WA, August 8-10, 2012.

(上接第 2 页)

式终端的测试内容在 3GPP TS 51.010 第 70.11 节中做了详细规定；WCDMA 制式终端的测试内容在 3GPP TS 34.171 中做了详细规定；CDMA 制式终端的内容在 3GPP C.S0036 中做了详细规定。A-GPS 控制平面下的射频一致性测试项，通常包括接收灵敏度测试、定位精准性测试、多径条件下的定位精准性测试、动态范围测试、移动环境下的定位精准性测试等。

3.2 A–GPS 用户平面测试

用户平面的安全用户平面测试 SUPL (Secure User PLane)，由 OMA 组织制定。目前主要是基于测试规范 OMA- ETS- SUPL- V2_0。这类测试主要在高层的业务平面进行测试，与底层的无线承载方式无关，因此对各类的终端都适用。SUPL 测试，主要验证具备 SUPL 能力的移动终端 (SET，SUPL Enabled Terminal) 和 SU-PL 定位平台 (SLP，SUPL Location Platform) 的信令交互符合标准要求。SUPL 定位平台主要提供定位信息服务，可以分为两部分，SUPL 位置中心 (SUPL Location Center) 和 SUPL 定位中心 (SUPL Positioning Center)。SUPL 在应用中有两种模式，代理模式和非代理模式。在代理模式中，SET 与 SLC 进行通信，当需要进行位置计算时，SLC 作为代理连接 SET 和 SPC。在非代理模式中，SET 可以和 SPC 直接进行通信。在 SUPL 测试规范 OMA-ETS-SUPL 中，分别对两种模式的测试进行了定义。

4 总结

目前，绝大多数运营商和终端设备制造商都支持 A-GPS 服务，并且随着互联网技术的发展，人们对位置服务的需求量也大大提高。A-GPS 目前仍然是全球运用最广的定位技术之一，而 A-GPS 终端一致性测试保证了移动终端的 A-GPS 功能和性能，在 A-GPS 技术的发展和应用过程中起到极为重要的作用。 MSTT