

See discussions, stats, and author profiles for this publication at: <http://www.researchgate.net/publication/282694593>

# A bot identification method based on Domain-Flux traffic features

CONFERENCE PAPER · JANUARY 2014

DOI: 10.13140/RG.2.1.2528.3287

2 AUTHORS, INCLUDING:



Tu Dinh Truong

Southeast University (China)

6 PUBLICATIONS 0 CITATIONS

SEE PROFILE

# 基于 Domain-Flux 流量特征的僵尸主机检测方法

Truong Dinh Tu<sup>1,2,3</sup>

程光<sup>1,3</sup>

(<sup>1</sup> 计算机科学与工程学院, 东南大学, 南京 210096, 中国)

(<sup>2</sup> 绥和工业学院, 信息技术部门, 富安 620900, 越南)

(<sup>3</sup> 东南大学计算机网络和信息集成教育部重点实验室, 南京 210096, 中国)

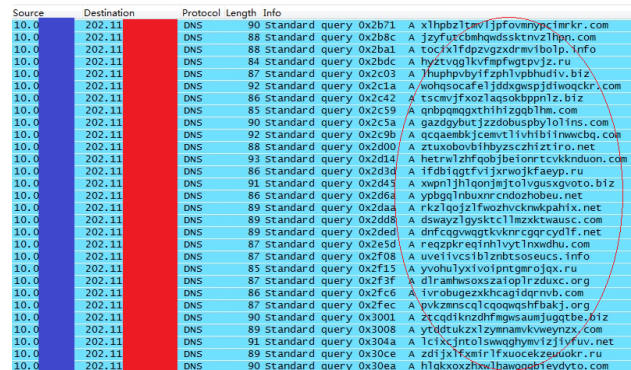
**摘要:** 针对被管网络中如何快速准确地识别僵尸主机问题, 本文提出一种基于 Domain-Flux 流量特征的僵尸主机检测方法。该方法在被管网关处以被动方式收集所有 DNS 流量, 并将其按源地址进行归并, 单独或综合应用 Domain-Flux 所产生伪随机域名流量的长度、字符分布相似性及周期性特征对该归并流进行分析 and 检测, 最后给出结果报告。在实际局域网环境下对该方法进行了实验, 结果表明, 该方法不仅能较为准确地检测出僵尸主机, 还可适用于对时空资源和实时性有不同要求的情形, 表现出较好的灵活性。  
**关键词:** 网络安全; 僵尸网络; Domain-fluxing; DNS 流量

## 1. 引言

僵尸网络 (Botnet) 是攻击者控制大量已感染僵尸程序的主机 (Bot) 而形成的以进行恶意活动 (如分布式拒绝服务、垃圾邮件、网络钓鱼、点击欺诈、窃取敏感信息等) 为目的的覆盖网络<sup>[1]</sup>。目前, 我国已成为世界上 Botnet 侵害较为严重的国家之一。2014 年 3 月底, 国家互联网应急中心发布的《2013 年我国互联网网络安全态势综述》报告显示, 2013 年我国境内感染木马僵尸网络程序的主机多达 1135 万个, 控制服务器为约为 16 万个, 地方政府网站成为黑客攻击的“重灾区”, 利用僵尸网络进行国家级有组织网络的攻击行为显著增多, 不仅影响广大网民的利益, 也危及我国公共互联网的安全运行, 对国家和政府信息安全构成严重威胁<sup>[2]</sup>, 已然成为构建和谐社所面临的重大问题之一。

Botnet 主要活动可分僵尸程序传播、命令与控制 (Command and Control, C&C)、发动攻击三个阶段。正常主机在感染僵尸程序变为 Bot 后, 只有在发现并与 Botnet 的命令与控制服务器 (C&C Server) 取得通信, 才能加入相应的僵尸网络, 被 Botmaster 控制和使用, 从而发挥攻击、窃取信息等作用, 否则, 其威胁性和危险性不大, 将会失去实用价值。因此, Bot 发现或找到 C&C Server 是整个 Botnet 活动的核心部分, 也是决定 Botnet 能够正常运作的关键所在。

目前, Bot 发现 C&C Server 机制的方式有多种如 Domain-Flux<sup>[3-5]</sup>, Fast-Flux<sup>[6]</sup>, 固定 IP 或者域名<sup>[7]</sup> 以及基于 P2P 协议<sup>[8]</sup> 等, 其中, Domain-Flux 是 Bot 使用自身内嵌的域名产生算法 (Domain Generation Algorithm, DGA) 产生大量伪随机域名, 如图 1 所示, C&C Server 域名也位于其中, 然后 Bot 选取部分或全部“伪随机”域名进行尝试连接, 直到从中找到真正的 C&C Server 域名, 能有效逃避相关监控系统的检测, 有效保护 C&C Server 地址信息, 在已发现的 Bot 样本代码中使用较为广泛。



Source	Destination	Protocol	Length	Info
10.0.0.1	202.11.11.1	DNS	90	standard query 0x2b71 A x1hpbz1mw1jpf0vmr9yctmrkr.com
10.0.0.1	202.11.11.1	DNS	88	standard query 0x2b8c A jzyfuctcbhqwdsstkrvz1hpn.com
10.0.0.1	202.11.11.1	DNS	88	standard query 0x2ba1 A tocfx1fdpvgzxdmrv1bo1p.info
10.0.0.1	202.11.11.1	DNS	84	standard query 0x2bdc A hzytvqg1kvmfmpfvgtpv1z.ru
10.0.0.1	202.11.11.1	DNS	87	standard query 0x2c03 A 7huphpvlyfzjh1vphhduiv.biz
10.0.0.1	202.11.11.1	DNS	92	standard query 0x2c1a A wqhqsocfsljddospjdiwockr.com
10.0.0.1	202.11.11.1	DNS	86	standard query 0x2c42 A tscmjfxoz1aqsokbnp1z.biz
10.0.0.1	202.11.11.1	DNS	85	standard query 0x2c59 A qrbpmsqgxtih1hizgblhm.com
10.0.0.1	202.11.11.1	DNS	90	standard query 0x2c5a A gazdgybutjzddobusplylo1rs.com
10.0.0.1	202.11.11.1	DNS	92	standard query 0x2c9b A qqcaemkfcjew11h1h1f1twkbcy.com
10.0.0.1	202.11.11.1	DNS	88	standard query 0x2d00 A ztuxobovb1hbyzsczhit1ro.net
10.0.0.1	202.11.11.1	DNS	93	standard query 0x2d14 A hetrwlzhfobjbeionrtcvkknudon.com
10.0.0.1	202.11.11.1	DNS	86	standard query 0x2d3d A ifdb1ggtfv1jxwjkfaeyr.ru
10.0.0.1	202.11.11.1	DNS	91	standard query 0x2d45 A wxp1j1n1qon1j1to1hgu5xqvoo.biz
10.0.0.1	202.11.11.1	DNS	86	standard query 0x2d64 A ypbq1n1bxnr1cndozhobue.net
10.0.0.1	202.11.11.1	DNS	89	standard query 0x2daa A rkz1qo1z1fwozhvcknwph1fx.net
10.0.0.1	202.11.11.1	DNS	89	standard query 0x2dd8 A dswayz1gysktc11mzxtktausc.com
10.0.0.1	202.11.11.1	DNS	89	standard query 0x2de1 A drfcagvwgk1krk1gpcy1f1.net
10.0.0.1	202.11.11.1	DNS	87	standard query 0x2e5d A reqzpkreq1h1v1t1nxwdu.com
10.0.0.1	202.11.11.1	DNS	87	standard query 0x2f08 A uve11vcs1b1z1nbt5soeucs.1nfo
10.0.0.1	202.11.11.1	DNS	85	standard query 0x2f15 A yvohulyx1v1o1pntg1o1ax.ru
10.0.0.1	202.11.11.1	DNS	87	standard query 0x2f3f A sl1rahsosxsz1o1p1zduxc.org
10.0.0.1	202.11.11.1	DNS	86	standard query 0x2fec A ivrbugezkhcag1dgrnub.com
10.0.0.1	202.11.11.1	DNS	87	standard query 0x2fec A pvkzmsnc1cqoqshfbakj.org
10.0.0.1	202.11.11.1	DNS	90	standard query 0x3001 A ztcdq1knzohfmgwsaum1ugqtbe.biz
10.0.0.1	202.11.11.1	DNS	89	standard query 0x3008 A ytdtdkuxz1z1ymam1vweywx.com
10.0.0.1	202.11.11.1	DNS	91	standard query 0x304a A 1c1x1jnto1swgghym1z1j1yfvu.net
10.0.0.1	202.11.11.1	DNS	89	standard query 0x30ce A zdi1x1fzxm1r1f1xuoekzeuokr.ru
10.0.0.1	202.11.11.1	DNS	90	standard query 0x30ea A h1akooxzhw1h1awonb1evytdto.com

图1 Bot 利用 Domain-Flux 技术产生的 DNS 流量

本文提出一种基于 Domain-Flux 流量特征的识别方法 IBDTF (Identifying Bot based on Domain-flux Traffic Features)。IBDTF 首先收集被管网内 DNS 流量, 然后利用 Bot 程序所产生的伪随机域名长度、字符分布相似性及周期性特征对该 DNS 流量进行分析, 从而

识别出被管网内的僵尸主机，最后本文在可控局域网内对IBDTF进行了实验，验证了该方法的有效性。

## 2. IBDTF僵尸主机检测机制

### 2.1 IBDTF检测框架

本文提出基于 Domain-Flux 流量特征的检测僵尸主机框架，如图 2 所示。首先，DNS 流量与收集模块在被管网网关处以被动方式收集所有 DNS 流量，并将源地址相同的流量归并在一起；然后，将归并后的 DNS 流量送入分析检测引擎。分析检测引擎根据相应的策略对归并后的 DNS 流量分别或综合进行伪随机域名长度、伪随机域名字符分布相似性及伪随机域名周期性分析；最后生成僵尸主机检测结果报告。

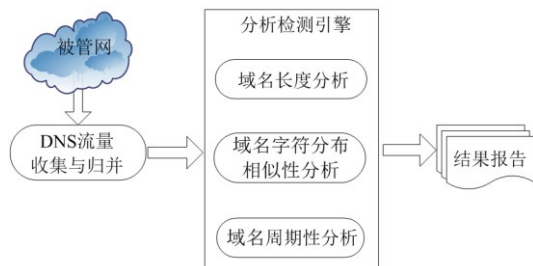


图2 IBDTF检测框架

### 2.2 Domain-Flux流量特征

Domain-Flux 流量是指已经感染僵尸程序的主机使用 Domain-Flux 产生的 DNS 请求流量，如图 1 所示。对 Domain-Flux 流量特征的分析 and 提取是解决僵尸主机检测问题的关键所在。本文在虚拟机环境下运行相关 Bot 程序，并捕获其产生的 Domain-Flux 流量数据，从中可总结出如下特征：

#### Feature1:伪随机域名长度

正常合法的域名 (Domain Name, DN) 是个有意义的字符串，如 sina.com.cn, microsoft.com 等，其背后一般对应着某个站点的一个或多个 IP 地址，它的作用是代替 IP 地址以方便人们记忆。在给定域名情况下，DNS 可通过查询操作来对该域名进行解析，以确定该域名是否已经注册及对应的 IP 地址是否存在。

通过观察发现，Bot 利用 DGA 算法所产生的伪随机域名长度与正常合法域名在长度上相比较，典型例子如图 1 所示。这一方面可能是由于尽量避免 DGA 所生成的域名与正常合法域名发生冲突，另一方面也是通过增大长度来增加伪随机域名的随机性以逃避

追踪和监测。

本文从 Alexa<sup>[9]</sup>网站和 Domain-Flux 的域名中分别随机选取 100 个域名，用 Normal DN、Illegal DN 来标识，并进行长度统计，其分布结果如图 3 所示。

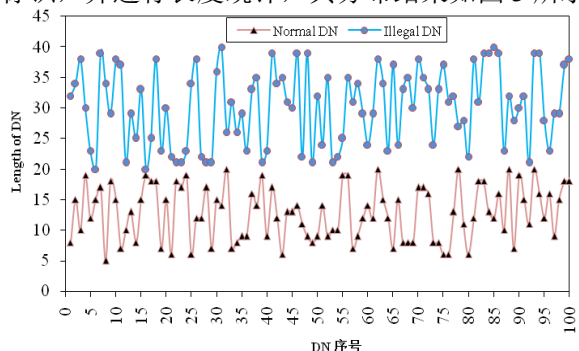


图3 正常 DN 与 Domain-Flux DN 长度对比

从统计结果来看，Domain-Flux 的伪随机域名长度范围为 21~45，而正常合法域名则在 5~19，很显然，总体上 Domain-Flux 的伪随机域名长度还是大于正常合法域名长度，且长度差异较为明显。本文中设置此伪随机域名长度阈值为 20。

#### Feature2:伪随机域名字符分布相似性

由于 Domain-Flux 采用 DGA 算法而产生大量伪随机域名，因此这些伪随机域名不仅在长度上相似，尤其在构成域名的数字和字母分布上也较为相似，可近似看成于等概率事件，如 krb581xdqkrowa47jxj26e5ljuareyc69m59p12.com、pzarntmxaumyn20oug33mqfqi35j36f22oe61.com 等。

而正常合法域名是提供给用户可读性强、易于记忆的具有一定含义的字符串，以取代人们对枯燥、单调的 IP 地址记忆。因此正常合法域名之间的相似度不是很高，其组成的数字和字母分布随机性较强，如 baidu.com, google.com。

因此，可通过比较未知域名与伪随机域名及正常合法域名的相似度来决定该域名归属哪类域名。本文此处采用 Jaccard 系数来衡量未知域名与伪随机域名及正常合法域名的相似度。假设  $X$  为未知域名， $A$  为从 Alexa 网站收集的 2000 个正常合法域名集合， $B$  为从 Domain-Flux 流量中收集的 5000 个伪随机域名集合。 $J_{normal}$ 、 $J_{illegal}$  分别代表  $X$  与集合  $A$ 、 $B$  的相似度，则：

$$J_{normal} = \frac{X \cap A}{X \cup A}, \quad J_{illegal} = \frac{X \cap B}{X \cup B} \quad (1)$$

$$\beta = J_{normal} - J_{illegal} \quad (2)$$

$$X \in \begin{cases} A, & \beta > 0 \\ B, & \beta < 0 \end{cases} \quad (3)$$

若  $\beta > 0$ ，则说明  $X$  与正常合法域名相似度较大，反之，则  $X$  与 Domain-Flux 的伪随机域名更为相似。

### Feature3: 伪随机域名周期性

僵尸程序在成功感染正常主机变为 Bot 后，为了获得命令和控制信息，会定期通过发送大量伪随机域名来尝试寻找和联系 C&C Server，即该过程具有周期性，尤其在 Bot 未联系到任何 C&C Server 情况下，会存在重复不断发送大量伪随机域名过程，使得其周期性更加明显<sup>[10-12]</sup>。

为了描述该周期性，本文此处采用循环自相关方法。该方法可在不了解无 DNS 请求时间等先验知识情况下，确定 DNS 请求的周期性行为，而无需计算具体周期值，其表达式如下：

$$c(r) = \sum_{i=1}^N g(i)g(i+r) \quad (4)$$

其中， $r$  表示自相关函数的时间片偏移量 ( $r=0,1,2,3, \dots, L-1$ )。时间片的长度为  $T$ ， $L$  为时间片总数。 $g(i)$ ， $g(i+r)$  分别代表某主机  $H$  在第  $i$ ， $i+r$  个时间片内访问某个域名的次数； $c(r)$  在  $r$  为周期及周期整数倍时较大，反之则较小，因此可通过计算  $c(r)$  的数值分布情况分析  $g(i)$  是否具有周期性规律。 $g(i)$  具有以下性质：

$$c(r) \leq c(0) \quad (5)$$

$$0 \leq \alpha(r) = \frac{c(r)}{c(0)} \leq 1 \quad (6)$$

根据公式(6)这个特征，一般用  $\alpha(r)$  的值的大小来刻画周期性， $\alpha(r)$  越大说明  $r$  越接近间隔的周期。预先设定的阈值  $\sigma$ ，计算出最大的  $\alpha(r)_{max}$ ，当两者满足公式(7)时，则认为该序列具有周期性。

$$\alpha(r)_{max} \geq \sigma \quad (0 \leq \sigma \leq 1) \quad (7)$$

本文在实验部分给出时间片长度  $T$  和阈值  $\sigma$  的设置。

### 3. 实验结果与分析

为测试IBDTF的有效性，本文实验在一个可控制的局域网内进行，该局域网内主机均能正常访问 Internet。该局域网内共有100台主机，其中50台为正常主机，剩余50台主机安装并能产生Domain-Flux流量的Bot程序。本文在该局域网网关处收集了总共3天

的DNS流量，形成测试数据TestData。

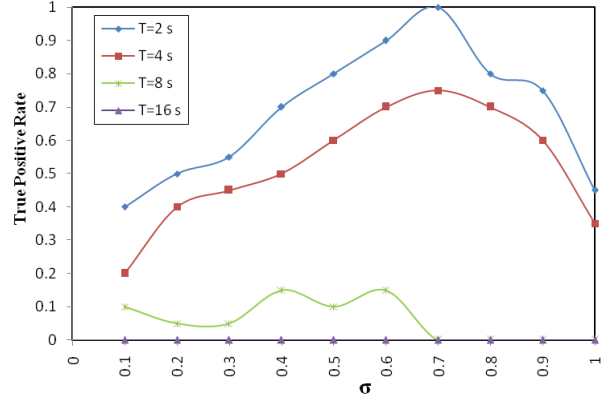


图4 不同 $T$ ， $\sigma$ 时周期性识别结果对比

本文中伪随机域名长度阈值设置为20。对于伪随机域名周期性的  $\sigma$  值设置，本文对数据集 B 进行了周期性测试，其结果如图4所示，可以看出在不同时间分片  $T$  下， $\sigma$  值取不同时对应的识别率差异较大。随着  $T$  值的增大，对伪随机域名周期性识别性剧烈下降，这是因为在  $T$  值很大时，伪随机域名所有出现的时间都完全属于同一个时间分片内，导致对伪随机域名周期性识别失败。当  $\sigma=0.7$  时，对伪随机域名周期性识别效果最好，因此，本文中设置  $T=2s$ ， $\sigma=0.7$ 。

对收集的数据集 TestData，本文首先通过手工方式确认和统计出了僵尸主机的 IP 地址及个数，接着分别单独应用 Feature1、Feature2 及 Feature3 对 TestData 进行检测，最后利用 Feature1、Feature2 及 Feature3 相结合的方法再对 TestData 进行检测，再将这4种方式的检测结果与手工统计结果进行对比，形成 ROC 曲线，如图5所示。其中，Length、Similarity、Periodicity 分别代表单独使用 Feature1、Feature2 及 Feature3 检测结果，而 All 表示综合应用三种特征检测结果。

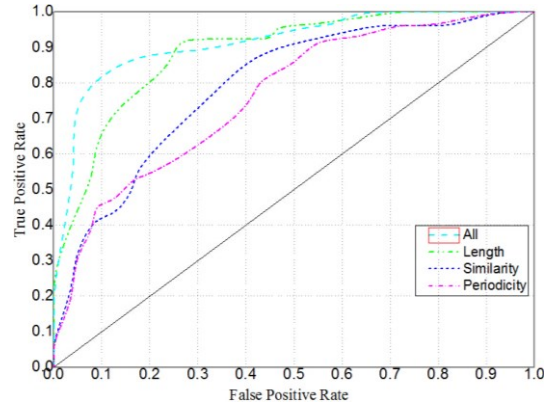


图5 各特征检测效果相应ROC曲线对比

从图 5 中可以看出 4 种检测方式中, 综合应用 3 种特征检测 (All) 时效果最好, 其次是基于伪随机域名长度特征的检测效果最好 (Length), 最后是伪随机域名字符分布相似性 (Similarity) 和伪随机域名周期性 (Periodicity)。其原因主要是 All 方式检测时限制条件较多, 使得错误识别僵尸主机数量减少, 导致假阳性率下降 (FPR), 误判率有所减小, 使得其对应的 AUC(Area Under the Curve of ROC)值较大, 检测效果也相对较好。而 Length、Similarity 及 Periodicity 检测方式限制条件单一, 假阳性率较大, 因而检测效果相对较弱, 但此 3 种方式检测过程简单, 时空复杂度相对较低, 因此可用于时空资源有限或实时性要求较高的场景。

#### 4. 结论

僵尸网络已对当前网络安全构成重大威胁。为从被管网中识别出僵尸主机, 本文给出一种基于 Domain-Flux 流量特征的 bot 检测方法 IBDTF。该方法可单独或综合应用 Domain-Flux 所产生伪随机域名流量的长度、字符分布相似性及周期性等从所收集的被管网 DNS 流量中识别出僵尸主机。通过实验测试, 该方法能较为准确地检测出僵尸主机, 还可适用于对时空资源和实时性有不同要求的情形。下一步将继续挖掘更多 Domain-Flux 流量特征, 并在实际应用环境中对 IBDTF 进行测试, 以进一步提高 IBDTF 准确性和实用性。

#### 参考文献:

- [1] 江健, 诸葛建伟, 段海新, 吴建平. 僵尸网络机理与防御技术[J]. 软件学报, 2012, 23(1):82-96.
- [2] CNCERT. 2013年我国互联网网络安全态势综述[R]. 北京: 国家互联网应急中心, 2014.
- [3] Damballa. Top-5 Most Prevalent DGA-based Crimeware Families[R/OL], [https://www.damballa.com/downloads/r\\_pubs/WP\\_DGAs-in-the-Hands-of-Cyber-Criminals.pdf](https://www.damballa.com/downloads/r_pubs/WP_DGAs-in-the-Hands-of-Cyber-Criminals.pdf).
- [4] Bilge L., Kirda E., Kruegel C. et.al.. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis[C]. In Proceedings of the 2011 Symposium on Network and Distributed System Security (NDSS'2011), 2011.
- [5] Antonakakis M., Perdisci R., Nadji Y. 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.

- [6] Riden J. Know your Enemy: Fast-Flux service networks[EB/OL]. The Honeynet Project, 2008. <http://www.honeynet.org/book/export/html/130>.
- [7] Ken C., Levi L., A case study of the rustock rootkit and spam bot[C]. In Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets, 2007.
- [8] S.Stover, D.Dittrich, J.Hernandez, et.al. Analysis of the Storm and Nugache Trojans:P2P is here, In proceedings of USENIX,2007:18-27.
- [9] Information Company. <http://www.alexa.com/topsites>. 2013.
- [10] Hyunsang C., Heejo L..Identifying botnets by capturing group activities in DNS traffic[J]. Computer Networks. 2012,56(1): 20-33.
- [11]Sousa, R., Rodrigues, N., Salvador, P., et.al..Analyzing the behavior of top spam botnets[C]. In Proceedings of the 2012 IEEE International Conference on Communications (ICC'12), 2012: 6540 - 6544.
- [12] Nelms T., Perdisci R., Ahamad M. ExecScent: Mining for New C&C Domains in Live Networks with Adaptive Control Protocol Templates[C]. In: Proceedings of the 22nd USENIX conference on Security (SEC'13), 2013:589-604.