北京大学学报(自然科学版) 第 52 卷 第 3 期 2016 年 5 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 3 (May 2016)
doi: 10.13209/j.0479-8023.2016.064

基于 Passive DNS 的速变域名检测

周昌令[1,2,3,†] 陈恺[1,2] 公绪晓[1] 陈萍[1] 马皓[1]

1. 北京大学计算中心, 北京 100871; 2. 北京大学信息科学技术学院, 北京 100871;
3. 北京大学计算机研究所, 北京 100871; † E-mail: zclfly@pku.edu.cn

摘要 利用 Passive DNS 采集校园网真实运行环境的域名访问记录, 从域名的多样性、时间性、增长性和相关性等方面构建 18 个特征集, 提出基于随机森林算法来识别速变域名的模型。交叉验证实验表明, 所构建的模型对域名分类的准确率超过 90%。在所采集的数据集上, 所构建的模型比 FluxBuster 能更有效地识别速变域名。

关键词 Passive DNS; 速变域名; 随机森林算法; DGA; CDN
中图分类号 TP393

Detection of Fast-Flux Domains Based on Passive DNS Analysis

ZHOU Changling[1,2,3,†], CHEN Kai[1,2], GONG Xuxiao[1], CHEN Ping[1], MA Hao[1]

1. Computer Center of Peking University, Beijing 100871; 2. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871; 3. Institute of Computer Science & Technology of Peking University, Beijing 100871; † E-mail: zclfly@pku.edu.cn

Abstract The authors use Passive DNS to log domain name query history of real campus network environment, and construct eighteen feature sets grouping by diversity, time, growth, and relevance, and then propose a model detect Fast-Flux Domains using random forest algorithm. The result shows that the proposed model can classify domains with accuracy over 90% by cross validation experiments. The model can detect Fast-Flux domains in the datasets used in this study more effectively compared with Fluxbuster.

Key words Passive DNS; fast-flux domain; random forest algorithm; DGA; CDN

目前, 速变服务网络(Fast-Flux Service Network, FFSN)[1]是一些恶意行为(如钓鱼网站、垃圾邮件、僵尸网络等)躲避打击的常见技术手段之一。其基本思想是利用大量被控制的主机提供中转服务, 隐藏背后的控制者。通过域名与 IP 对应关系的快速变化, 避免被 IP 黑名单隔离, 提高其服务的可用性。速变域名(Fast-Flux Domain)的特点是 DNS 应答记录的生命周期(TTL)很短, 应答通常返回的是不断变化的 IP 地址列表, 这些 IP 地址往往属于不同物理位置的不同运营商。由于内容分发网络(Content Delivery Network, CDN)以及循环 DNS (Round-robin DNS)等技术的使用, 使得一些正常的网络服务也有类似的特点。

本文利用 Passive DNS 方法记录域名访问的信息, 从域名的多样性、时间性、增长性、相关性等角度构建 18 个特征集, 基于随机森林算法建立速变域名识别模型, 并在真实的网络环境中进行验证。与 FluxBuster[2]的对比验证表明, 本文的识别模型表现出更好的识别效果。本文提供相关的代码和数据集下载①。

国家 2012 年下一代互联网技术研发、产业化和规模商用专项项目(CNGI-12-03-001)、国家发展改革委员会 2011 年国家信息安全专项和 863 计划(2015AA011403)资助
收稿日期: 2015–06–01; 修回日期: 2015–07–07; 网络出版日期: 2016–05–17
① 相关的代码地址: https://github.com/whodewho/FluxEnder。

1　相关研究

Passive DNS 技术是由 Weimer[3]在 2005 年提出的一种方案，用来解决 DNS 系统 PTR[4]反向查询能力不足的问题。通过将现有 DNS 业务的流量进行镜像或分光处理，解析出查询和响应的数据并存入数据库，然后建立正向和反向的查询索引。在 ISP 或校园网络的递归 DNS 服务器前部署 Passive DNS 系统，可以获得详细的 DNS 查询记录，并且不影响现有 DNS 服务器的运行性能。

Honeynet 项目组 2007 年对 FFSN 进行了系统的介绍[1]，推动了 FFSN 的研究热潮。Holz 等[5]对 FFSN 开展了试验性研究，对比分析 FFSN 和 CDN 域名的差别。Passerini 等[6]提出根据域名注册时间、域名注册商、域名 A 记录和域名 TTL 值等来检测 FFSN。Huang 等[7]提出基于地理位置和地理分布来检测 FFSN。汪洋[8]选取 A 记录数、IP 分散度、TTL 值和域名创建时间 4 个特征构成检测向量，分别采用神经网络和 SVM 进行速变域名(Fast-Flux Domain)检测，是国内较早进行的 FFSN 研究工作。

Antonakakis 等[9]的 Notos 系统通过计算域名的信誉值来判断域名是否为恶意，其模型引入的特征包括域名字符串分析、恶意域名历史记录等信息。Bilge 等[10]提出 EXPOSURE 系统，利用域名的时间特征、IP 分布、TTL 以及域名字符串特征等构建 15 个特征集。Perdisci 等[2]提出 FluxBuster 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。

2　特征选取

为了发现速变域名在 DNS 响应上的特征，本文在取得校园网实际运行的 DNS 数据后，在 FluxBuster 和 EXPOSURE 等系统的基础上，提出如表 1 所描述的 18 个特征，其中 14~18 这 5 个相关性特征是本文首次提出。本文采用 scikit-learn[11]中的随机森林算法[12–13]作为恶意域名的识别算法。此算法具有鲁棒性强、泛化误差会收敛、不存在过度拟合等优点。以下是各特征的具体描述。

2.1　时间性特征

min_ttl, max_ttl 和 diff_ttl 用来记录域名响应中 TTL 的最小值、最大值和差值。

表 1　特征列表

Table 1　Feature list

分类	编号	特征	描述
时间性	1	min_ttl	最小 TTL 值
	2	max ttl	最大 TTL 值
	3	diff_ttl	max_ttl-min_ttl, 域名活动期
多样性	4	ip_count	域名的 IP 数目
	5	p16_entropy	按/16 前缀熵值
	6	subd_count	子域名个数
	7	subd_len_entropy	子域名长度熵
增长性	8	p16_growth_1	1 个周期内，新增 IP/16 前缀的比
	9	p16_growth_4	4 个周期内，新增 IP/16 前缀的比
	10	p16_growth_8	8 个周期内，新增 IP/16 前缀的比
	11	subd_growth_1	1 个周期内，新增子域名的比例
	12	subd_growth_4	4 个周期内，新增子域名的比例
	13	subd_growth_8	8 个周期内，新增子域名的比例
相关性	14	relevant_domain_count	与该域名相关的域名集合大小
	15	unique_2ld_ratio	相关域名中，不同 2ld 的比例
	16	unique_ip_ratio	域名指向的非公共 IP 的比例
	17	noshare_domain_ratio	具有非公共 IP 前缀的相关域名比例
	18	max_dga_ratio	IP 指向的域名 DGA 比例的最大值

2.2 多样性特征

ip_count 用来记录域名解析出来的 IP 地址集的大小。

p16_entropy 是衡量 ip 地址分散程度的特征。为了表示 IP 地址集的分散程度，本文选取/16 前缀来分析。假设 IP 地址集的集合为 P, $p(x)$是 IP 的/16 前缀 x 在 P 中所占的比例，$p(x)=\text{count}(x)/|P|$，则

$$\text{p16\_entropy}=\frac{-\sum_x p(x)\cdot\log_2 p(x)}{\log_2|P|}。$$

subd_count 和 subd_length_entropy 分别记录子域名的数量和子域名的长度熵。设某域所有子域名的集合为 S, s 是子域名的长度，$p(x)=\text{count}(x)/|S|$，则有

$$\text{subd\_len\_entropy}=\frac{-\sum_s p(s)\cdot\log_2 p(s)}{\log_2|S|}。$$

2.3 增长性特征

p16_growth_n 和 subd_growth_n 分别记录相对于前 1, 4, 8 个数据周期，新增的 IP/16 前缀比例和新增的子域名比例。

设 P_m 为数据周期内的 IP/16 前缀的集合，定义公式如下:

$$\text{p16\_growth\_n}=\frac{|P_m-\bigcup_{j=1}^{k}P_{m-j}|}{|p_m|}。$$

类似地，子域名增长性公式如下:

$$\text{subd\_growth\_n}=\frac{|S_m-\bigcup_{j=1}^{k}S^r_{m-j}|}{|S_m|}。$$

其中，S_m 为子域名的集合。

2.4 相关性特征

本文认为，同一个域名解析出来的多个 IP 之间是相关的；类似地，多个域名解析到同一个 IP 时，这几个域名之间是相关的。特征 relevant_domain_count 是相关域名的大小。如图 1 所示，d1~d9 是相关的，故相关域名的数量是 9。通常，CDN 网站(如 taobao.com)的一组 IP 地址会有多个域名指向它们，则此特征的值比较大；而速变域名的相关域名一般较少。

unique_2ld_ratio 指所有相关域名中不同 2LD (域名的第二部分，是域名的主要内容，如 www.baidu.com 的 2LD 为 baidu.com)的比例。

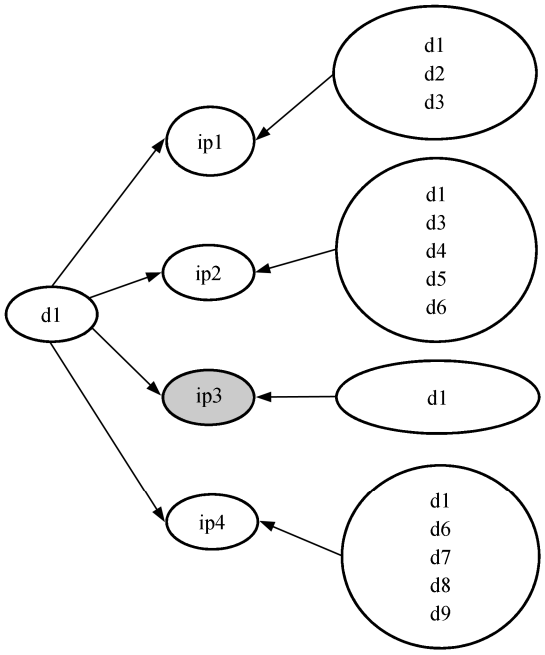

图 1 相关域名

Fig. 1 Relevant domain

unique_ip_ratio 用来描述相关域名之间的非公共 IP 比例，如图 1 中 ip3 是非公共 IP，其比例为 1/4= 25%。

由域名关联到域名后，如果考虑不同域名之间的非共享 IP/16 前缀，noshare_domain_ratio 定义为具有非共享前缀的相关域名比例。试验数据表明，一些速变域名特征取值较高，而正常域名(包括使用 CDN 和 RRDNS 的)特征大多为零。如图 2 所示，假设除 ip3 外，其他 ip 都具有相同的 IP/16 前缀，则 d2 为非共享域名，noshare_domain_ratio= 0.25。

max_dga_ratio 指通过域名生成算法(DGA)自动生成的域名所占的比例。通过 DGA 生成的一组域名往往会指向相同的 IP。本文选取一个开源的识别 DGA 域名的工具[14]来计算所指向 IP 的域名中 DGA 域名的比例。

3 试验数据

本文的数据来源于北京大学校园网的真实运行数据。目前的部署情况是将校园网主要的三台递归 DNS 查询服务器的流量，用端口镜像的方式转发到采集服务器。在采集服务器上运行 PassiveDNS 工具[15]，分析得到的 DNS 查询记录数据定时保存到文件中，文件内容如图 3 所示。为了保存长期的记录，按照日期对日志进行分割，保存在不同的

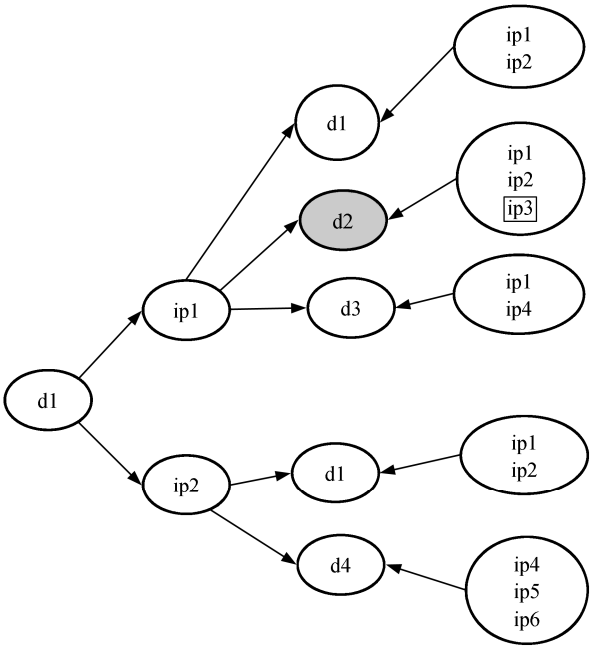

图 2　非共享域名
Fig. 2　Noshare domain

文件中。

为了便于后期分析和查询, Passive DNS 日志数据经过处理后，保存在 TokuMX 数据库[16]中。TokuMX 是 MongoDB[17]的一个分支，采用分形索引算法以及高效的压缩方案，与传统的 MongoDB 相比，具有更高的性能，同时节约近 90%的存储空间[16]。实际上，我们也发现它的数据文件大小只有原始版本 MongoDB 的 10%~20%，并且写入和查询的性能更好。为了提高后期特征提取以及分析的效率，我们将 ip→domain 和 domain→ip 的映射关系以及一些统计数据预先计算好，并分别保存在不同的 collection[17]中。

数据处理流程如图 4 所示。

3.1　数据预处理

由于本文关注速变域名，只有满足如下 3 个条件[2]的域名记录才会被分析计算，并进入到机器学习步骤中:

```
#timestamp||dns-client ||dns-server||RR class||Query||Query Type||Answer||TTL||Count
1449763231.446715||162.105.129.27||205.251.199.14||IN||bsveri.com.||A||54.199.145.144||60||38
1449763231.446715||162.105.129.27||205.251.199.14||IN||bsveri.com.||A||54.65.54.163||60||38
1449763231.471277||222.29.32.135||162.105.129.105||IN||vip.clickzzs.nl.||A||87.233.228.115||600||1
1449763231.502727||162.105.129.58||140.205.81.26||IN||3rf78.duomeng.net.||CNAME||r.duomeng.net.||600||1
1449763231.633706||162.105.129.58||211.100.44.216||IN||www.logistank.com.||A||101.200.72.72||3600||1
1449763231.666274||162.105.129.85||193.223.77.3||IN||www.dikant.de.||A||185.21.102.223||3600||1
1449763231.671776||162.105.129.85||193.223.77.3||IN||www.dikant.de.||AAAA||2a00:1158:0:300:6d0b::1||3600||1
1449763231.673524||162.105.129.31||88.221.81.192||IN||e4306.g.akamaiedge.net.||A||23.7.130.165||20||1
1449763231.727131||162.105.129.58||140.205.81.16||IN||EtiZ8.duomeng.cn.||CNAME||r.duomeng.cn.||600||1
1449763231.927777||162.105.129.27||110.75.38.29||IN||5ibada.taobao.com.||CNAME||shop.taobao.com.||1800||1
1449763231.966965||162.105.183.235||162.105.129.135||IN||www.bustynikkibenz.com.||CNAME||bustynikkibenz.com.||600||1
1449763231.966965||162.105.183.235||162.105.129.135||IN||bustynikkibenz.com.||A||173.225.176.136||600||1
```

图 3　日志格式
Fig. 3　Log format

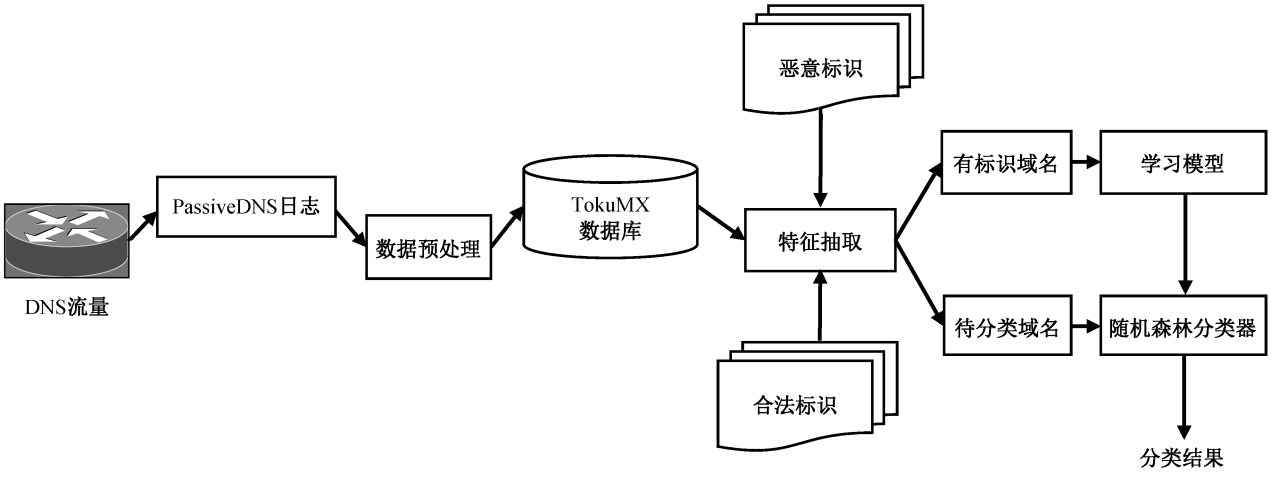

图 4　流程图
Fig. 4　Flow diagram

$$\begin{cases} \min\_\text{ttl} \leqslant \theta_{\text{ttl}}, \\ |R| \geqslant \theta_R, \\ \text{div}(R) \geqslant \theta_{\text{div}}, \end{cases}$$

其中, min_ttl 是最小 ttl 值, R 是域名 d 指向的 IP 地址集合, $|R|$是地址集合的大小, div(R)描述 IP/16 前缀的散度。通过试验确定, 本文最终使用的各阈值如下: θ_{ttl} = 20000 s, θ_R = 1, θ_{div}= 0.1。

本文将观测到的部分特殊情况列入免于处理的列表中: 域名污染(由于 DNS 劫持, 查询如 facebook、twitter 等域名返回的是伪造的同一组地址)和待售域名(运营商或域名注册经营者保留的域名, 可读性高)。

3.2 训练数据

本文的合法域名标识来自 Alexa[18], 它提供过去 3 个月内全球范围内站点的流量排名。排名靠前的域名一般是正常域名(也包括使用 CDN 和 RRDNS 的域名)。FFSN 相关域名标识取自表 2。

实际情况中速变域名所占比例较少, 本文借鉴褚燕琴等[19]的主动查询的办法, 在校园网内部署主动查询节点, 定期向表 2 中部分域名发起查询请求。由于采集的是校园网的递归 DNS 流量, 在校园网内部对这些恶意域名的主动查询数据就会被记录下来。表 2 中的域名数据实际上还包括一些非速变域名的数据, 不过, 经过预处理后, 被留下的基本上可以认为是速变域名。

本文用于训练的数据是北京大学校园网 2014–03–01 到 2014–03–17 递归 DNS 服务器的数据, 采集到的不同域名数量如图 5 所示。

表 2 域名信息来源

Table 2 Source of domain information

类别	网址
垃圾邮件	http://untroubled.org/spam
钓鱼网站	http://www.phishtank.com
恶意域名	http://www.abuse.ch
	https://zeustracker.abuse.ch
	https://spyeyetracker.abuse.ch
	https://palevotracker.abuse.ch
	http://www.malwaredomains.com
	http://atlas.arbor.net/summary/fastflux
	http://www.malwaredomainlist.com
	http://hosts-file.net
	http://cybercrime-tracker.net

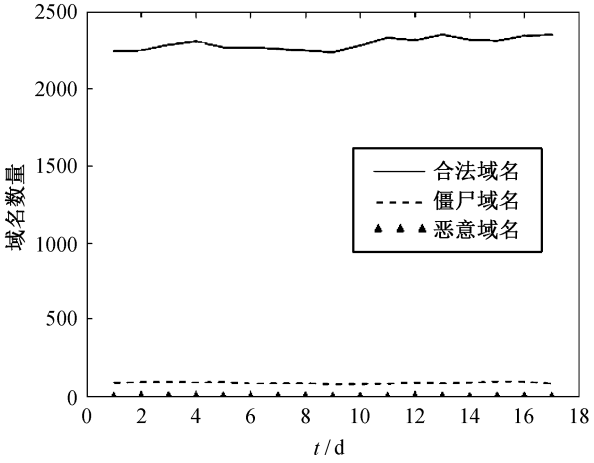

图 5 不同域名数量

Fig. 5 Number of different domains

3.3 交叉验证

本文采用交叉验证(cross validation)对模型进行检验。图 6 是对 2014–04–14 的数据进行 10 次交叉检验得到的准确率结果, 其中合法域名 2109 个, 恶意域名 338 个。可以看到每次分类的准确率都超过 90%。具体到其中一次(选取 20%的标识数据作为验证集, 其中合法域名有 437 个, 恶意域名有 53 个), 正确识别的 TP (恶意域名被识别成恶意域名)和 TN (合法域名被识别成合法域名)分别为 90.6%和 94.5%, 错误识别的结果 FN (恶意域名被识别为合法域名)为 9.4%, FP (合法域名被当成了恶意域名)为 5.5%。

可以将多次交叉验证的结果取平均值作为准确率。图 7 是 2014–04–02 到 2014–04–25 期间的数据

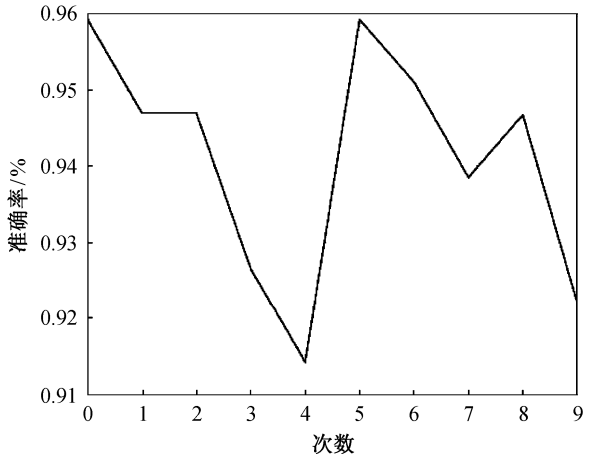

图 6 交叉验证

Fig. 6 Cross validation

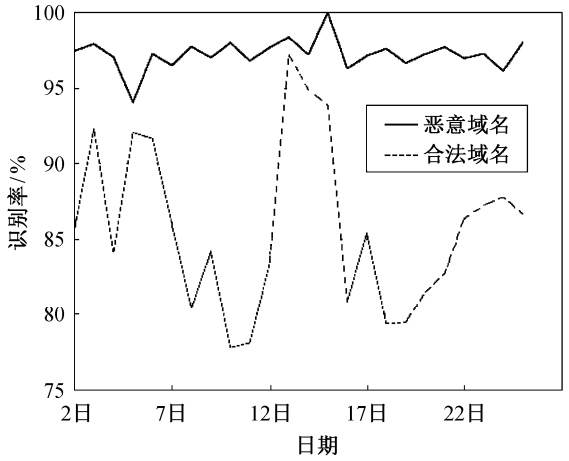

图 7　正确识别的比率
Fig. 7　Correct classification rate

集进行 10 次交叉验证取平均的结果，可以看到恶意域名被正确识别的比例基本上都超过 95%，合法域名的正确识别率也基本上都在 80%以上。

3.4　对比试验数据

本文选择 FluxBuster 作为对比，它提供源代码下载[20]。将本文使用的数据集转换成 FluxBuster 可以识别的格式，并进行重新训练和聚类检测，FluxBuster 所发现的 2014–04–02 至 2014–04–25 期间的恶意域名数量如图 8 所示。具体地，这段时间 FluxBuster 标记为速变域名簇的 2LD 只有两个：weminemnc.com 和 gccdn.net。经过人工验证，这两个域名簇里的域名实际都不是速变域名(前者是 bitcoin 的 dns-seed[21]，后者是 CDN 域名)。

分析原因，除在特征个数、模型参数以及算法选择等方面的差异外，FluxBuster 与本文的最大区别在于对相关性特征的使用上。FluxBuster 没有直接包括相关性特征，而是把不同域名的 IP 地址集的重叠程度作为聚类依据，对聚类后的域名簇判断是否为速变网络。近期的研究发现速变域名已经出现一些新的特点[22]，使得很多原来的检测特征不再有效，但速变域名之间却有了更多的相关性。本文提出的 14~18 这 5 个相关性特征有助于从相关性去识别速变域名，故本文方法取得更好的效果。

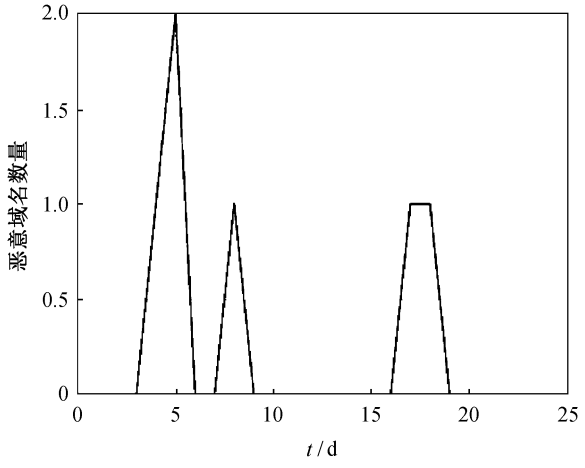

图 8　FluxBuster 识别出的恶意域名
Fig. 8　Malware domain identified by FluxBuster

4　结论和展望

本文利用 Passive DNS 方法采集域名信息，构建识别速变网络的 18 个特征集，用随机森林算法建立了相应的识别模型。在真实的网络运行数据集上验证了模型的有效性，并与开源工具进行对比。试验表明，本文提出的速变域名识别方法识别正确率高，取得比 FluxBuster 更好的识别效果。

本文的局限性在于采集到的速变域名数据集较小，因此我们只能选取 alexa 正常域名的一小部分来做正向标签，这就使得整个训练集规模较小，对模型精度有影响。另一方面，模型的部分特征计算量较大，有待进一步改进。

参考文献

[1] Riden J. Know your enemy: fast-flux service networks [EB/OL]. (2008–08–16)[2015–05–01]. http://www.honeynet.org/papers/ff

[2] 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

[3] Weimer F. Passive DNS replication // FIRST Conference on Computer Security Incident. Singapore, 2005: 1–13

[4] Mockapetris P V. Domain names, concepts and facilities [EB/OL]. (1987)[2015–03–01]. http://tools.ietf.org/html/rfc1034

[5] Holz T, Gorecki C, Rieck K, et al. Measuring and detecting fast-flux service networks // NDSS. San Diego, 2008: 487–492

[6] Passerini E, Paleari R, Martignoni L, et al. Fluxor: detecting and monitoring fast-flux service networks // Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin: Springer, 2008: 186–206

[7] Huang S Y, Mao C H, Lee H M. Fast-flux service network detection based on spatial snapshot mechanism for delay-free detection // Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security. Beijing, 2010: 101–111

[8] 汪洋. Fast-flux 服务网络检测方法研究[D]. 武汉: 华中科技大学, 2009

[9] Antonakakis M, Perdisci R, Dagon D, et al. Building a dynamic reputation system for DNS // USENIX Security Symposium. Washington DC, 2010: 273–290

[10] Bilge L, Kirda E, Kruegel C, et al. EXPOSURE: finding malicious domains using passive DNS analysis // NDSS. San Diego, 2011: 1–5

[11] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. The Journal of Machine Learning Research, 2011, 12: 2825–2830

[12] Ho T K. Random decision forests // Proceedings of the Third International Conference on Document Analysis and Recognition. Montreal, 1995: 278–282

[13] Quinlan J R. C4.5: programs for machine learning. San Francisco: Morgan Kaufmann Publishers, 2014

[14] Sconzo M. DGA detection [CP/OL]. (2014–01–21) [2015–05–11]. https://github.com/ClickSecurity/data_hacking/tree/master/dga_detection

[15] Edward B. A network sniffer that logs all DNS server replies for use in a passive DNS [CP/OL]. (2011–04–29) [2015–05–11]. https://github.com/game linux/pas sivedns

[16] Percona. High-performance MongoDB distribution [EB/OL]. (2006–01–01) [2015–05–11]. https://www.percona.com/software/mongo-database/percona-tokumx

[17] MongoDB. MongoDB for GIANt ideas [EB/OL]. (2009–08–20) [2015–12–11]. https://www.mongodb.org/

[18] Alexa Internet. Actionable analytics for the web [EB/OL]. (1996–04–01) [2015–12–11]. http://www.alexa.com/

[19] 褚燕琴, 应凌云, 冯登国, 等. 速变服务网络行为特征分析. 计算机系统应用, 2013(8): 1–8

[20] Perdisci R, Corona I, Giacinto G. Early detection of malicious flux networks via large-scale passive DNS traffic analysis [EB/OL]. (2012–10–20) [2015–12–11]. https://code.google.com/p/fluxbuster/

[21] Bitcoin Community. Satoshi client node discovery [EB/OL]. (2014–03–13) [2015–12–11]. https://en.bitcoin.it/wiki/Satoshi_Clie nt_Node_Discovery

[22] Xu W, Wang X, Xie H. New trends in FastFlux networks [EB/OL]. (2013–12–04)[2016–04–07]. https://media.blackhat.com/us-13/US-13-Xu-New-Trends-in-FastFlux-Networks-WP.pdf

[23] 陈恺. 基于 Passive DNS 的恶意域名识别研究[D]. 北京: 北京大学, 2014