doi:10.3969/j.issn.1000-436x.2013.05.017

# 基于 DNS 的隐蔽通道流量检测

章思宇[1]，邹福泰[1]，王鲁华[2]，陈铭[3]
（1. 上海交通大学 信息安全工程学院，上海 200240；
2. 国家计算机网络与信息安全管理中心，北京 100017；3. 上海交通大学 密西根学院，上海 200240）

**摘 要**：为提出一种有效检测各类型 DNS 隐蔽通道的方法，研究了 DNS 隐蔽通信流量特性，提取可区分合法查询与隐蔽通信的 12 个数据分组特征，利用机器学习的分类器对其会话统计特性进行判别。实验表明，决策树模型可检测训练中全部 22 种 DNS 隐蔽通道，并可识别未经训练的新型隐蔽通道。提出的检测方法在校园网流量实际部署中成功检出了多个 DNS 隧道的存在。

**关键词**：域名系统；隐蔽通道；入侵检测；机器学习；网络安全
**中图分类号**：TP393.08 **文献标识码**：B **文章编号**：1000-436X(2013)05-0143-09

# Detecting DNS-based covert channel on live traffic

ZHANG Si-yu[1], ZOU Fu-tai[1], WANG Lu-hua[2], CHEN Ming[3]
(1. School of Information Security, Shanghai Jiaotong University, Shanghai 200240, China;
2. National Computer Network and Information Security Administration Center, Beijing 100017, China;
3. UM-SJTU Joint Institute, Shanghai Jiaotong University, Shanghai 200240, China)

**Abstract:** To propose an effective detection method for DNS-based covert channel, traffic characteristics were thoroughly studied. 12 features were extracted from DNS packets to distinguish covert channels from legitimate DNS queries. Statistical characteristics of these features are used as input of the machine learning classifier. Experimental results show that the decision tree model detects all 22 covert channels used in training, and is capable of detecting untrained covert channels. Several DNS tunnels were detected during the evaluation on campus network's live DNS traffic.

**Key words:** domain name system; covert channel; intrusion detection; machine learning; network security

## 1 引言

网络隐蔽通道是攻击者绕过网络安全策略进行数据传输的重要途径，而 DNS（域名系统）则是实现应用层隐蔽通道的常用手段。

DNS 是互联网最为关键的基础设施之一，将域名与 IP 地址相互映射。由于其在网络运行中的重要地位，DNS 协议几乎不会被防火墙策略阻拦，即使在一个内部网络中，也需要架设 DNS 服务器进行主机名解析。DNS 是一个全球分布的数据库，域名递归解析需要本地 DNS 服务器与互联网上其他服务器通信，这也就为基于 DNS 协议构建隐蔽通道创造了条件。通过域名递归解析的 DNS 隐蔽通道客户端只需请求本地 DNS 服务器，而不必与通道的另一方直接通信，大大增加了访问控制策略制定的难度。

过去几年的研究已提出多种 DNS 隐蔽通信方法，且有成熟的软件实现，应用于 IP 或 TCP 隧道、木马和僵尸网络的控制通信。建设无线城市部署的 Wi-Fi 热点，用户认证前通常已开放 DNS 服务，利用 DNS 隧道可绕过认证自由访问互联网，对网络运营商造成损失。而在信息安全

**收稿日期**：2012-01-31；**修回日期**：2012-05-25
**基金项目**：国家自然科学基金资助项目（61071081）；国家 242 信息安全计划基金资助项目（2011A004）；信息网络安全公安部重点实验室开放课题基金资助项目（C11608）
**Foundation Items:** The National Natural Science Foundation of China (61071081); The National 242 Information Security Plan (2011A004); The Open Project of MPS Key Laboratory of Information Network Security (C11608)

要求较高的组织，隐蔽通道则可能造成机密信息泄露等严重的安全事件。文献[1]提出了利用 DNS 传输文件和封装 SSH 隧道的方法。文献[2~4]对现有的 DNS 隧道实现在带宽、延迟和可靠性上进行了比较，实验在低延迟网络中达到了 500kbit/s 的吞吐量。文献[5]利用二进制编码域名进一步提高隧道带宽，文献[6]通过 DNS 数据分组松弛空间数据注入，实现对 DNS 解析器和安全工具透明的被动 DNS 隧道。

相对于层出不穷的 DNS 隐蔽通道传输和隐藏手段，相应的检测技术仍存在诸多不足。目前，对应用层隐蔽通道检测的研究大多针对 HTTP 和 SSH 协议[7, 8]，检测 DNS 隐蔽通道的方法只有如下几类。

1) 请求量、长子域名统计。DNSBL（DNS 黑名单）等频繁请求同一域中大量子域名的情况在实际环境中较为常见，这一方法易产生误报，同时又忽略低频率、低带宽的隐蔽通道。文献[5]提出伪造源地址分散通道的方法，也能绕过这一检测机制，适合构建泄密和远程控制类的非对称通道。

2) 特殊资源记录类型统计。该方法检测 DNS 隐蔽通道常用的 TXT 和 NULL 资源记录，Iodine[9]等使用 A 记录请求即会使此方法失效。

3) Born 等人提出的域名字符频率分析[10, 11]。该检测方法同样存在对 DNSBL 等服务的误判，因为 DNSBL 的子域名一般为 IP 地址或 MD5 散列值，不符合英语单词的字符频率特性。文献[10, 11]实验中合法流量产自 Alexa 排名前百万网站的爬虫，仅代表 Web 服务的域名特征，与实际 DNS 流量特性差异较大。此外，字符频率分析未考虑文献[5]在域名中包含二进制数据的隐蔽通道。

上述 DNS 隐蔽通道检测方法均只适用于基于域名递归解析的通道，而尚未解决隐藏在 UDP/53 流量中客户端与服务器直接通信的通道，如文献[6]的数据注入、Iodine 的 Raw UDP 隧道[9]等。如何设计一种适合各种类型 DNS 隐蔽通道检测要求、不依赖于特定的通道设计模式假定的 DNS 隐蔽通道检测方法，成为亟待解决的问题。

本文提出了一种在 DNS 流量中全面检测各种类型 DNS 隐蔽通道的方法，利用机器学习的分类器对合法请求和隐蔽通信特征进行判别。本文的贡献如下。

1) 对 DNS 隐蔽通道的构建方法进行了总结，全面分析了 DNS 隐蔽通信流量的数据分组特征及会话连接的统计特性。

2) 实验验证了本文的算法能够识别训练涉及的全部 22 种隐蔽通道，并且具有识别未经训练的新型 DNS 隐蔽通道的能力，解决了现有检测算法在可检测的通道类型上的局限性。

3) 本文的算法可应用于实时 DNS 流量分析，并且实现了相应的检测系统，在上海交通大学校园网进行了部署测试，检测到多个实际存在的隐蔽通道并进行了分析。

## 2 DNS 隐蔽通道

DNS 协议的隐蔽通道可分为 2 类：第一类利用 DNS 的递归域名解析，在本文中称为基于域名的隐蔽通道；另一类需要客户端与隐蔽通道服务器直接通信，在本文中称为基于服务器的隐蔽通道。

### 2.1 基于域名的 DNS 隐蔽通道

攻击者注册一个域名，并将其域名服务器（NS）设置为隐蔽通道的服务器。隐蔽通道客户端向任意一台 DNS 递归服务器请求该域下子域名，均可实现与服务器通信。

客户端向服务器发送的数据编码为域名的子域名字符串。DNS 域名标签允许包含英文字母、数字和连字符，且不区分大小写，因此，DNS 隐蔽通道通常将数据进行 Base32 编码。服务器返回的数据则包含在 DNS 回答的资源记录中，其中最常用的资源记录类型是 NULL 和 TXT，前者可包含任意长的二进制数据，后者则要求为可打印字符。A 记录（IPv4 地址）、MX 记录（邮件交换）和 AAAA 记录（IPv6 地址）的请求在互联网 DNS 流量中所占比例最大，尽管带宽利用率低于 NULL/TXT，但它们仍为注重隐蔽性的 DNS 隧道的首选。

基于域名的 DNS 隐蔽通道程序实现，有 IP over DNS 隧道：NSTX[12]、Iodine[9]、DNSCat[13]以及 TCP over DNS 隧道：OzyManDNS[1]、Dns2tcp[14]、TCP-over-DNS[15]和 Heyoka[5]等。

### 2.2 基于服务器的 DNS 隐蔽通道

当网络安全策略允许主机与任意一台 DNS 服务器通信时，可使用基于服务器的 DNS 隐蔽通道。攻击者将基于 UDP 的服务（如 OpenVPN）运行在

53 端口，从客户端直接建立连接。Iodine 发现客户端能与隧道域名 NS 直接通信时，也会自动转入 Raw UDP 模式。在此模式下，整个 UDP 载荷均为隐蔽通道数据，通信效率大幅提升。然而，由于这些报文不是有效的 DNS 消息，流量分析工具解析这些报文时会出现格式异常（malformed），从而引起怀疑。文献[6]在现有 DNS 消息末尾的松弛空间注入数据的方法，不影响 DNS 服务器和流量分析工具对数据分组的解析，可解决上述 Raw UDP 隧道的缺陷。

## 3　检测方法

### 3.1　定义 DNS 数据连接

为了判断各个客户端的域名请求行为是否存在隐蔽通信的可能，本文对 DNS 流量中的“数据连接”进行定义，以确定各个消息的通信双方。

对截获的 UDP/53 数据分组作 DNS 协议解析，如果解析中无任何错误发生，并且解析完毕时指针位于 UDP 载荷的末尾，表明该数据分组是一个无注入的合法 DNS 报文。本文将这类数据分组的 DNS 连接双方定义为<*client_ip*, *pure_domain*>。*client_ip* 为 DNS 客户端地址，即 DNS 查询报文的源，或回答报文的目的 IP 地址。*pure_domain* 为 DNS 请求域名（QNAME）的纯域名部分，即将 QNAME 去除由同一 NS 授权的子域名标签。

如果将数据分组解析为 DNS 时发生错误，或者解析完毕后指针未处于 UDP 载荷末尾，表明该数据分组可能为 Raw UDP 隧道，或存在松弛空间数据注入。这类数据分组的通信双方直接取其 IP 地址<*client_ip*, *server_ip*>，网络边界监控时认为内网主机地址为 *client_ip*。

### 3.2　数据分组特征

通过对 DNS 隐蔽通道构造方法和流量的分析，本文从 DNS 数据分组中提取了一系列可用于区别 DNS 隐蔽通道的特征。

#### 3.2.1　数据分组解析特征

如 3.1 节所述，对采集的数据分组进行 DNS 协议解析时若出现格式异常或有注入数据，则可能为基于服务器的 DNS 隐蔽通道。对于数据注入的情况，本文计算注入数据长度，即协议解析完毕时指针与 UDP 载荷末尾的距离，作为表示松弛空间注入量的特征参数。

DNS 隐蔽通道在数据传输时，充分利用 DNS 协议各字段或 Raw UDP 报文允许的最大长度，以提高带宽利用率、减少数据分组数量。基于域名的通道，其上行和下行数据分别存储在 DNS 消息的问题段和回答段，增加了 DNS 消息本身的长度。基于服务器的隧道，由于无法将其作为 DNS 报文解析，本文将 UDP 载荷长度也作为数据分组解析特征之一。

#### 3.2.2　请求域名特征

请求域名特征针对基于域名的隐蔽通道。DNS 问题段除 QNAME 以外的字段只能容纳 4byte 数据，因此，基于域名的通道上行数据通常只能存储在 QNAME 中，提取 QNAME 的特征：标签数量和子域名长度。DNS 协议限制域名最长为 255byte，每个标签不超过 63byte，因此，DNS 隐蔽通道将上行数据编码为长域名之后，必须分割为多个标签。

文献[5]在域名中使用二进制数据以提高传输效率，也是 QNAME 的特征之一。实验发现大多数 DNS 隐蔽通道使用了 DNS 协议允许的字符集以外的字符。对于使用二进制编码的子域名，请求中所包含的信息量与子域名长度相等；仅使用 DNS 协议允许的字符，则必须 Base32 编码，QNAME 包含的信息量等于子域名长度的 60%。

#### 3.2.3　DNS 消息特征

编码域名中使用前向指针是文献[6]提出的提高数据注入检测难度的手段，前向指针在一般的 DNS 解析器和服务器实现中几乎不会被采用，因而是否存在前向指针是识别隐蔽通道的特征之一。隐蔽性较高的隧道的 A 记录请求，一般采用 CNAME（规范名称）记录来携带较多的下行数据，本文将回答含 CNAME 记录作为第二个 DNS 消息特征。

基于域名的隐蔽通道，下行数据存储在资源记录中，尤其是回答段的资源记录中，因此，本文计算回答段资源记录数据长度（RDLENGTH）之和，以及整个报文包括授权和附加段在内全部资源记录 RDLENGTH 之和，作为表示 DNS 回答所容纳的隧道下行数据量的 2 个参数。

#### 3.2.4　特征总结

通过对数据分组 3 类特征的分析，总共提取了 12 个数据分组特征用于区别合法 DNS 请求与隐蔽通信，总结如表 1 所示。

**表 1 数据分组特征**

| 特征类别 | 编号 | 特征名称 |
|---|---|---|
| 数据分组解析 | $F_1$ | 数据分组解析异常 |
| | $F_2$ | UDP 载荷长度 |
| | $F_3$ | DNS 消息长度 |
| | $F_4$ | 注入数据长度 |
| 请求域名 | $F_5$ | 标签数量 |
| | $F_6$ | 子域名字串长度 |
| | $F_7$ | 域名包含二进制数据 |
| | $F_8$ | 域名存储数据量 |
| DNS 消息 | $F_9$ | 编码域名含前向指针 |
| | $F_{10}$ | 回答含 CNAME 记录 |
| | $F_{11}$ | 回答段资源记录数据尺寸 |
| | $F_{12}$ | 全部资源记录数据尺寸 |

### 3.3 DNS 数据连接特征

仅依据一个数据分组的特征难以判断是否为 DNS 隐蔽通信，因此，算法对属于同一数据连接的数据分组特征进行统计分析，再依据数据连接的统计特性进行分类器判别。

如表 2 所示，DNS 数据连接的统计特征分为 3 类：特征集 $FS_1$ 包含一定时间段内该连接传入和传出的数据分组总数；$FS_2$ 统计了该连接中具有前向指针、CNAME 记录，以及 QNAME 含二进制数据的特殊报文数量;特征集 $FS_3$ 是对数据分组特征 $F_2$、$F_3$、$F_4$、$F_5$、$F_6$、$F_8$、$F_{11}$、$F_{12}$ 的统计，计算该连接所有数据分组上述参数的均值、最大值、最小值及求和。其中，数据分组长度特征（$F_2, F_3, F_4$）对传出和传入 2 个方向分别统计；请求域名特征（$F_5, F_6, F_8$）只对 DNS 请求报文统计，因为回答报文与请求中的 QNAME 保持完全一致；资源记录特征 $F_{11}$ 和 $F_{12}$ 仅在 DNS 回答报文中有意义，因此仅对回答进行统计。

**表 2 DNS 数据连接特征集**

| 特征集 | 特征集说明 | 特征数 |
|---|---|---|
| $FS_1$ | 传入、传出及 DNS 数据分组总数 | 3 |
| $FS_2$ | 含有前向指针、CNAME 记录和二进制域名的数据分组数量 | 3 |
| $FS_3$ | 数据分组特征 $F_2$、$F_3$、$F_4$、$F_5$、$F_6$、$F_8$、$F_{11}$、$F_{12}$ 的统计参数 | 42 |

特征集 $FS_3$ 共含 42 个特征，主要代表了 DNS 请求与应答报文、域名、资源记录的平均长度和变化范围，以及双向的数据传输量。利用合法 DNS 流量样本及DNS隧道流量样本对 $FS_3$ 中特征分布进行分析，其中隧道流量样本含 50%的活跃传输的隧道及 50%空闲、保持连接状态的隧道。如图 1(a)所示，DNS 隐蔽通道对外传输数据时，最长的子域名往往在 25 个字符以上，而合法域名请求中仅 2%的子域名超过 25 个字符；但是，DNS 隧道在空闲状态下的子域名通常小于 10byte，与合法请求相似。图 1(b)显示了 DNS 会话在 5min 内回答数据总字节数分布，98.4%的合法 DNS 通信在 5min 内的回答数据小于 20KB，而 DNS 隐蔽通道活跃时 5min 的下行数据通常在 50KB 以上。

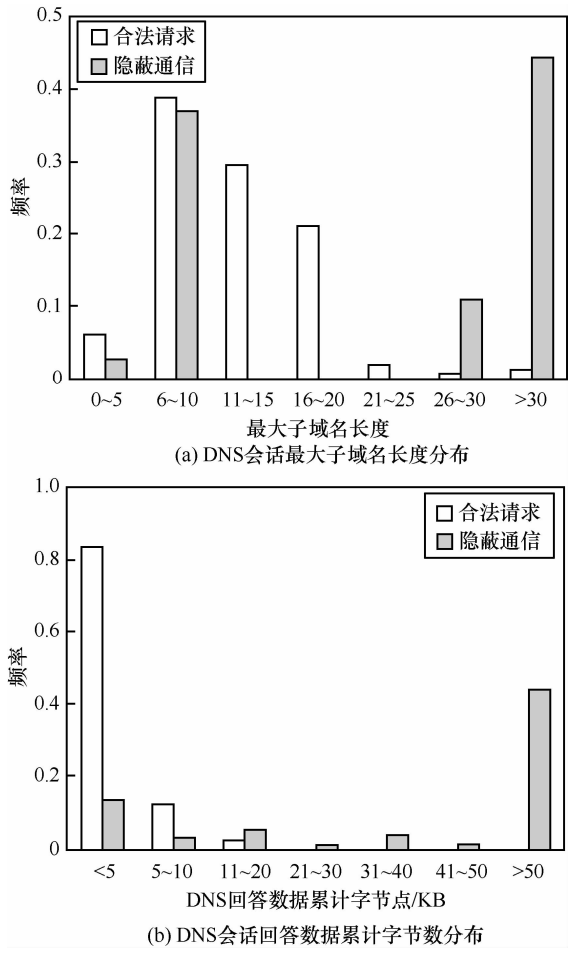



图 1 DNS 数据连接统计特征分布

### 3.4 检测系统流程

本文提出的 DNS 隐蔽通道检测流程如图 2 所示。DNS 流量探针监测所有的 UDP/53 流量。对于 DNS 探针采集到的数据分组，计算其 12 个数据分组特征参数，并进行初步数据分组过滤，去除明显不符合隐蔽通道特征的 DNS 报文。

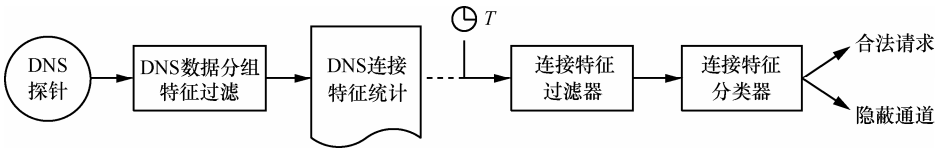



图 2　DNS 隐蔽通道检测流程

研究了 DNS 隐蔽通信的必备要素之后，本文确定了如下的初步过滤规则

$$(F_1 = false) \wedge (F_4 = 0) \wedge (F_6 \leqslant 4) \wedge (F_{11} \leqslant 1)$$

符合该条件，即无格式异常、无数据注入、子域名长度不超过 4 字符且回答资源记录总长不超过 1byte 的，将被数据分组过滤模块丢弃。对余下的数据分组，根据 3.1 节的定义，识别其所属 DNS 数据连接的通信双方，然后将该数据分组特征更新至 DNS 连接特征统计表，以计算 3.3 节提出的 DNS 数据连接统计特征。

一个 DNS 数据连接在 DNS 连接特征统计表中跟踪监测的时间达到统计时限 $T$ 后，进入判别为合法或隐蔽通信的流程。如果其数据分组速率达到 $R_{\mathrm{m}}$，则将 DNS 数据连接特征集 $FS_1$、$FS_2$ 和 $FS_3$ 的全部特征输入一个机器学习的分类器，应用预先训练好的分类模型，判断是否为 DNS 隐蔽通道。对数据分组速率低于 $R_{\mathrm{m}}$ 的连接，认为其请求频率对隐蔽通信而言过小，直接判定为合法的 DNS 请求。

系统参数 $T$ 决定了隐蔽通信开始到系统响应的最大延迟，提高 $T$ 的取值可增强对低速通道的数据分组采集，但将延长响应时间。$R_{\mathrm{m}}$ 取决于安全策略所能容忍的最大隐蔽通道带宽。单个 DNS 请求携带的最大上行数据 $D_{\mathrm{UM}}$ 小于 QNAME 长度，即 $D_{\mathrm{UM}} < 255$。DNS 请求与应答分组成对出现，分析隐蔽信道的泄密风险时，上行带宽 $B_{\mathrm{U}} \leqslant D_{\mathrm{UM}} \times R_{\mathrm{m}} / 2 < 128 \times R_{\mathrm{m}}$。本文实验及系统原型实现采用参数 $T = 5\,\mathrm{min}$ 及 $R_{\mathrm{m}} = 4\,\mathrm{byte/min}$，考虑到 $B_{\mathrm{U}} < 512\,\mathrm{byte/min}$ 对泄密与远程控制等应用均太小，而 5min 的监测足够对低速率通道采集至少 10 次请求用于统计分析。实验将比较几种常用的分类器模型在本算法中应用时的分类效果。

## 4　实验与评估

### 4.1　训练数据采集

对 DNS 数据连接分类器进行训练和评估，需采集一系列合法 DNS 流量样本，以及 DNS 隐蔽通道的流量样本。实验使用的合法流量样本取自上海交通大学校园网的 DNS 流量，在 1h 的流量截取文件中，提取单一客户端对同一纯域名请求次数最多的 6 890 个 DNS 数据连接，并确认了其均不属于 DNS 隐蔽通信。

为了获得 DNS 隐蔽通道样本，实验运行了多个现有的 DNS 隧道软件，截取其产生的流量。测试的 DNS 隧道程序包括 Iodine、Dns2tcp、DNSCat、tcp-over-dns 和 PSUDP。对于支持多种资源记录类型的 Iodine、Dns2tcp 和 DNSCat，分别截取了其在 NULL、TXT、SRV、MX、CNAME、KEY 等资源记录，以及 Raw UDP 模式下的流量。

为使训练产生的模型既能识别数据传输中的大吞吐量隐蔽通道，又可检测低频交互、小流量的通道，对上述 DNS 隧道软件的实验中，分别截取其活动状态（有数据通过隧道传输）和空闲状态（隧道保持连接）时的流量。隧道传输的数据采用 Web 浏览器自动加载网页的方法产生。PSUDP 采用在已有 DNS 流量中注入数据的被动工作方式，自身不产生 DNS 请求，因此，PSUDP 的实验，以 1 次/秒的频率发送 DNS 请求，使 PSUDP 能够进行数据传输。

上述 DNS 隧道程序的实验总共涵盖 22 种隐蔽通信模式。实验对每种模式分别截取 6 个时间段，产生总共 132 个 DNS 隐蔽通道流量样本（如表 3 所示）。

**表 3　　训练样本集**

| 类别 | 样本来源 | 样本数 |
|---|---|---|
| 合法请求 | 校园网 DNS 流量 | 6 890 |
| 隐蔽通道 | Iodine（6 类资源记录） | 78 |
| | Dns2tcp（2 类资源记录） | 18 |
| | DNSCat（2 类资源记录） | 18 |
| | tcp-over-dns | 12 |
| | PSUDP | 6 |

### 4.2　分类器模型比较

本文使用 Weka[16]软件，对 J48 决策树、朴素贝叶斯（Naïve Bayes）和逻辑回归（LR, logistic regression）算法进行比较，分类器模型采用十折交

叉验证进行测试。

分类器模型比较结果如表 4 所示，J48 决策树和 LR 算法的模型正检率（true positive rate）相同，均为 95.6%。朴素贝叶斯算法的准确率明显较低 J48 和 LR 算法，因此不适用于本文研究的场景。LR 算法的误报率（false positive rate）低于决策树算法，而决策树算法 ROC 曲线（如图 3 所示）区域面积（AUC）最大，即其平均性能最优。

J48 决策树算法在实验比较中取得了较为准确的分类结果，并且该算法效率高，输出的模型直观明了。J48 算法采用自顶向下、分而治之的的决策树构造方法，选择信息增益率最大的属性进行分裂，递归直至决策树各节点样本均为相同类别或无属性可分裂。决策树构造过程中对分类信息增益最大的属性进行了选择，模型实际使用的特征数量较少，在实时流量处理时可降低资源开销。因此，在后续的实验及实时流量检测系统的实现中，采用了 J48 决策树作为机器学习的分类器算法。

**表 4　　分类器模型比较**

| 分类器 | ROC 区域（AUC） | 正检率/% | 误报率/% |
|---|---|---|---|
| J48 决策树 | 0.992 | 95.6 | 0.15 |
| 朴素贝叶斯 | 0.793 | 58.4 | 1.18 |
| 逻辑回归 | 0.984 | 95.6 | 0.04 |

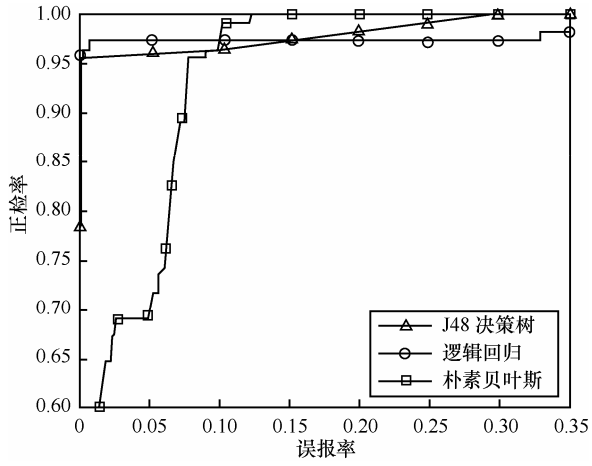



图 3　分类模型 ROC 曲线

### 4.3　特征评估

在不同的特征选取的情况下，分别建立分类器模型进行比较，采用全样本集和十折交叉验证评估，各个模型的误报率均在 0.15%至 0.30%的范围内，因此着重比较其漏检率（false negative rate）。由于 $FS_3$ 包含的特征数量众多，进一步将 $FS_3$ 分为报文解析特征统计（$FS_3PP$）、请求域名特征统计（$FS_3DN$）和资源记录特征统计（$FS_3RR$）3 部分进行了评估，不同特征选取时模型漏检率如图 4 所示。结果表明 $FS_3$ 和全部特征集（$ALL$）均取得了较低的漏检率，而 $FS_1$ 和 $FS_2$ 的漏检率较高。

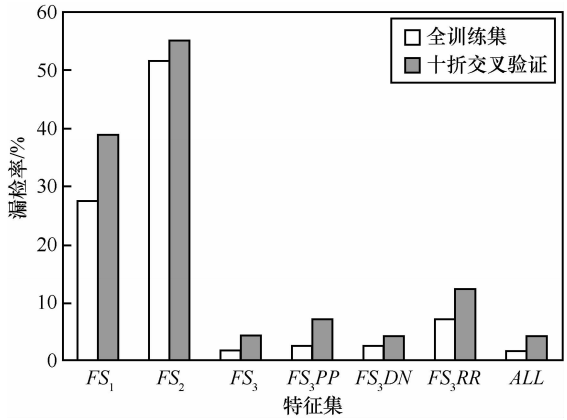



图 4　各类特征集下的漏检率

与本文算法相比，传统的基于高请求频率判断 DNS 隐蔽通道的方法，仅仅依赖于统计同一客户端对同一纯域名的请求量，即仅使用本文算法的特征集 $FS_1$ 中的特征，而缺乏应用层深入分析产生的特征集 $FS_2$ 和 $FS_3$。图 5 对比了合法请求与隐蔽通道样本的报文数量分布，DNS 隧道程序在保持连接和低速率传输时的请求频率与合法应用难以区分，34.5%的隧道样本 5min 报文数小于 150，而有 17.2%的合法样本报文数超过该值。根据此参数设定阈值，为使误报率保持在合理范围内，将不可避免地忽略低带宽的隐蔽通信。因此，特征评估实验利用 $FS_1$ 特征集建立决策树模型的漏检率达 30%左右。

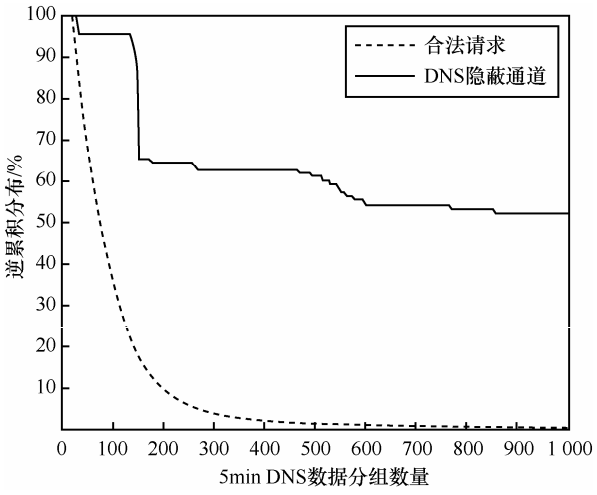



图 5　同域名查询频率分布

特征集 $FS_2$ 分析含前向指针、CNAME 记录和二进制域名的流量异常，符合标准规范实现的 DNS 隐蔽通道避免产生上述特殊报文，因此，仅依据特征集 $FS_2$ 能够检测的隐蔽通道类型较少，在特征评估实验中的漏检率高达 50%以上。

特征集 $FS_3$ 对协议报文尺寸、QNAME 特征和资源记录特征的统计特性分析能够准确区分合法与隐蔽通道。如图 4 所示，根据交叉验证评价，$FS_3PP$、$FS_3DN$ 和 $FS_3RR$ 分别建立模型的漏检率为 7.1%、4.4%和 12.4%。3.3 节提取的全部特征中，对 J48 算法分类信息增益率最大的特征为 $\max(F_8)$，即 QNAME 所携带的最大信息量，因而其所属的 $FS_3DN$ 分类准确率高于 $FS_3PP$ 与 $FS_3RR$。特征集 $FS_1$ 对报文数量的计数 $n_p$ 与 $FS_3PP$ 协议报文尺寸求和 $\Sigma L_i$ 在报文尺寸 $L_i$ 一定的前提下呈正相关性；$FS_2$ 检测的 CNAME 和二进制域名特殊报文，则在 $FS_3RR$ 的回答资源记录尺寸及 $FS_3DN$ 的 QNAME 信息量中对应地进行了计算。从而，$FS_3$ 部分携带了 $FS_1$ 与 $FS_2$ 特征所含的信息，依据 $FS_3$ 建立的决策树模型取得了与全特征集相同的准确率。

### 4.4　未知隐蔽通道模式检测

为验证决策树模型的检测能力，对训练使用的隐蔽通道模式重新截取新的流量，测试结果表明该模型能检出全部 22 种已知的 DNS 隐蔽通道。进一步地，本文实验检验模型对训练未涉及的“未知”DNS 隐蔽通道的检测能力。实验另外测试了 3 个 DNS 隧道程序，分别为 OzyManDNS、Heyoka，以及作者自行设计的 DNS 隐蔽通道 NSChan。NSChan 使用 Base32 域名编码，下行数据采用了与其他隧道均不同的设计，将数据编码为多个 IPv4 地址，通过 A 记录返回。对于上述 3 个未经训练的 DNS 通道程序，检测结果如表 5 所示。

模型成功检测活动和空闲状态的 OzyManDNS 与 Heyoka，以及活动状态下的 NSChan。模型未能检测空闲状态的 NSChan，对其流量分析后发现，两方面因素导致了漏检的产生。首先，NSChan 在空闲、保持连接状态下的请求频率 $f_{\text{idle}}=0.2$，低于训练使用的 Iodine 的 $f_{\text{idle}}=0.33$，以及 DNSCat、Dns2tcp 和 tcp-over-dns 的 1.0、2.0 和 6.0，从而，其 5min 对外报文数 $n_{\text{o}}=Tf_{\text{idle}}=60$ 恰好低于决策树模型在该参数的阈值 68，被判为合法。其次，本文作者实现 NSChan 时，对封装报文头部长度未作优化，空闲时子域名长度与其他隧道活跃时相当，使决策树在请求报文最小长度节点处进入了适用于活跃隧道的分支。尽管如此，NSChan 在数据传输时，不可避免地出现频率提高及域名长度增加的隐蔽通信特征，活动状态依然无法躲避本算法检测。作为改进方案，模型训练时可添加将隧道程序 $f_{\text{idle}}$ 参数减小后采集的样本，截取报文量接近系统参数 $R_{\text{m}}$（如 3.4 节）的流量，以增强模型对请求频率低至下界 $f_{\text{m}}=2/R_{\text{m}}$ 的隧道的检测。

**表 5　检测训练集以外的 DNS 隐蔽通道**

| 通道程序 | 状态 | 检测结果 |
|---|---|---|
| OzyManDNS | 活动 | Yes |
| | 空闲 | Yes |
| Heyoka | 活动 | Yes |
| | 空闲 | Yes |
| NSChan | 活动 | Yes |
| | 空闲 | No |

### 4.5　实际环境评估

本文对 DNS 隐蔽通道检测算法进行了实现，处理实时的 DNS 流量发现其中的隐蔽通信行为。将检测系统在上海交通大学网络中心的 DNS 流量监控服务器上进行了部署，检验算法在实际环境中的效果。系统监测的 DNS 请求来自约 30 万个源 IP 地址，DNS 请求量约 3 000 次/秒。

流量经数据分组过滤后，DNS 数据连接统计表中暂存的记录为 11 万条，内存使用最大 50MB。经过持续 10h 的运行和检测，实际环境测试结果如表 6 所示。系统共产生了 30 万个样本进入分类器，其中 310 个被检出的样本确认属于隐蔽通信，系统总共检测到 7 个独立的隐蔽通道的存在。

**表 6　实际环境测试结果**

| 测试名称 | 测试结果 |
|---|---|
| 样本总数 | 302 528 |
| 检测数 | 310 |
| 独立隐蔽通道 | 7 |
| 样本误报 | 324 (0.107%) |
| 客户端总数 | 9 665 |
| 客户端误报 | 38 (0.393%) |
| 域名总数 | 2 802 |
| 域名误报 | 23 (0.821%) |
| DNS 连接总数 | 64 473 |
| DNS 连接误报 | 45 (0.070%) |


图 6　CipherTrust DNS 隧道流量分析

在误报方面，分类器对样本的误报率为 0.107%，10h 内误报的客户端 IP 地址为 38 个，误报的域名仅 23 个。DNS 流量进入分类器前，经过了初步数据分组过滤、DNS 数据连接过滤 2 个过滤器，前者滤除了 27.5%的合法 DNS 报文，后者则将约 91%的 DNS 会话直接判为合法而无需经过分类器判别。2 个过滤器使得最终进入分类器判决的会话数与系统输入的 DNS 流量相比大幅减少，因此，系统对全部流量的误报率远低于表 6 中分类器模块的误报率。

对检测系统的误报分析发现，系统的误报主要来自于异常的客户端程序实现，如 29%的误报产生自某主机以 5s 为周期对 a.root-servers.net 的重复请求，由于该请求产生的回答含大量授权和附加段资源记录，被认为是隧道下行流量。对于 DNSBL 服务，模型训练中已学习的以 MD5 或 IP 地址为前缀的 DNSBL 在实际环境部署时没有任何误报，但在实际流量中误检了 2 个以 URL Encoding 编码的网址为前缀的 DNSBL 域名：ph.bdaph.com 和 html.ph.bdrbl.com。这 2 个域名由 BitDefender（比特梵德，反病毒软件公司）注册，网址经过 URL 编码，再用减号替代百分号后作为子域名字串（如表 7 所示）。该 DNSBL 子域名长度比 MD5 和 IP 地址大得多，因而被本系统误检。

**表 7　　URL 编码的 DNSBL 示例**

| 编码步骤 | 字符串 |
|---|---|
| 查询网址 | baike.baidu.com/view/671.htm |
| URL 编码的网址 | baike%2ebaidu%2ecom%2fview%2f671%2ehtm |
| DNSBL 请求域名 | baike-2ebaidu-2ecom-2fview-2f671-2ehtm.ph.bdaph.com |

在对本系统检测到的 DNS 隐蔽通道的分析中，也发现了 DNS 隧道在一些合法软件中的应用。McAfee（麦克菲）的软件利用 DNS 隧道技术将用户电子邮件的部分信息传送到服务器，用于其声望评分系统。其 DNS 隧道利用域名 ciphertrust.net，抓包分析显示其 DNS 请求符合典型的 DNS 隐蔽通信模式（如图 6 所示）。

## 5　结束语

本文通过对 DNS 隐蔽通道构建方法的研究和总结，提出了基于机器学习的检测算法，能够全面检测利用域名递归解析和服务器直接通信的多种 DNS 隐蔽通道模型，解决了现有算法在通道类型上的局限性。经过实验比较，本文选择利用 J48 决策树分类器对 DNS 数据连接的特征进行判别。分类器模型可检测训练涉及的全部 22 种隐蔽通道模式，以及多种未经训练的新型隐蔽通道。本文实现了在实时 DNS 流量中检测隐蔽通道的系统，在校园网环境中进行了部署测试，成功检测到 7 个隐蔽通道的存在，并探讨了实际运行中发现的一些特殊的 DNS 隧道的应用。

## 参考文献：

[1] KAMINSKY D. The black OPS of DNS[A]. Proceedings of the Black Hat USA 2004[C]. Las Vegas, 2004.

[2] LEIJENHORST T V, CHIN K-W, LOWE D. On the viability and performance of DNS tunneling[A]. Proceedings of the 5th International Conference on Information Technology and Applications[C]. Cairns, Australia, 2008.

[3] NUSSBAUM L, NEYRON P, RICHARD O. On robust covert channels inside DNS[A]. Proceedings of the 24th IFIP International Security Conference[C]. Pafos, Cyprus, 2009.

[4] MERLO A, PAPALEO G, VENEZIANO S, *et al*. A comparative

performance evaluation of DNS tunneling tools[A]. Proceedings of the 5th International Conference on Complex, Intelligent, and Software Intensive Systems[C]. Seoul, Korea, 2011. 84-91.

[5] REVELLI A, LEIDECKER N. Introducing heyoka: DNS tunneling 2.0[A]. Proceedings of the SOURCE Conference Boston[C]. Boston, 2009.

[6] BORN K. PSUDP: a passive approach to network-wide covert communication[A]. Proceedings of the Black Hat USA 2010[C]. Las Vegas, 2010.

[7] ZANDER S, ARMITAGE G, BRANCH P. A survey of covert channels and countermeasures in computer network protocols[J]. Communications Surveys & Tutorials, IEEE, 2007, 9 (3): 44-57.

[8] DUSI M, CROTTI M, GRINGOLI F, *et al*. Tunnel hunter: detecting application-layer tunnels with statistical fingerprinting[J]. Computer Networks, 2009, 53 (1): 81-97.

[9] ANDERSSON B, EKMAN E. Iodine[EB/OL]. http://code.kryo.se/ iodine/, 2011.

[10] BORN K, GUSTAFSON D. NgViz: detecting DNS tunnels through N-gram visualization and quantitative analysis[A]. Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research[C]. Oak Ridge, Tennessee, 2010. 1-4.

[11] BORN K, GUSTAFSON D. Detecting DNS tunnels using character frequency analysis[A]. Proceedings of the 9th Annual Security Conference[C]. Las Vegas, Nevada, 2010.

[12] GIL T M. NSTX (IP-over-DNS)[EB/OL]. http://thomer.com/ howtos/ nstx.html.

[13] PIETRASZEK T. DNScat[EB/OL]. http://tadek.pietraszek.org/ projects/ DNScat/, 2011.

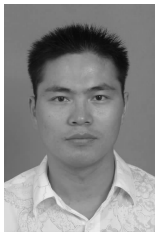
[14] DEMBOUR O. Dns2tcp[EB/OL]. http://www.hsc.fr/ressources/ outils/ dns2tcp/index.html.en, 2011.

[15] VALENZUELA T. Tcp-over-dns[EB/OL]. http://analogbit.com/ software/ tcp-over-dns, 2011.

[16] HALL M, FRANK E, HOLMES G, *et al*. The WEKA data mining software: an update[J]. SIGKDD Explorations, 2009, 11 (1): 10-18.

**作者简介：**

**章思宇**（1989-），男，上海人，上海交通大学硕士生，主要研究方向为网络与信息安全。



**邹福泰**[通信作者]（1973-），男，江西安福人，博士，上海交通大学讲师，主要研究方向为信息安全和分布式计算。E-mail: zoufutai@sjtu.edu.cn。

**王鲁华**（1980-），男，山东曹县人，硕士，国家计算机网络与信息安全管理中心工程师，主要研究方向为网络与信息安全。

**陈铭**（1989-），男，云南景洪人，上海交通大学本科生，主要研究方向为计算机系统和体系结构。