**基于HTTPS流量的目标网页分类与识别**

1. **课题来源、选题依据、课题研究目的、工程应用价值**

**1、课题来源**

本课题来源于网络运营商的实际需求。网络运营商需要利用自己手头收集到的用户访问各个网页所产生的数据，分析自己用户当前关心的热点内容**。**

1. **选题依据**

流量加密是保护隐私的一个重要手段，访问流量加密的站点，能够保护用户的数据不被窥探，也能够阻止犯罪分子窃取用户信用卡信息、应用的使用习惯或密码。随着HTTPS协议的普及应用，Web访问数据越来越多地被加密传输，据统计截至到2018年，超过80%的访问流量被HTTPS协议加密。虽然这对于重视隐私的用户来说是一个福音，但HTTPS 在给人们提供了性能良好、安全隐私性更高的服务的同时，也给不法分子散步不良信息带来了更大的便利和隐蔽性。隐藏在HTTPS下的互联网信息难以在网络流量层面进行监管，虚假的社会消息、负面的言论得以在加密的流量中快速传播，对社会稳定产生重大危害。

为此，网络服务提供商希望能够在不侵犯具体用户隐私的前提下，实时地对用户的流量进行监控，第一时间发现民众关心的热点，进而采用更进一步的手段判断这些热点内容是否为敏感问题、不良内容等，而其中最重要的就是针对用户浏览网页时产生的HTTPS流量进行分析，发现其中隐藏的URL访问目标。因此，对于访问特定Web页面所产生的HTTPS流量的识别、检测以及分类工作，显得重要而有意义。

1. **课题研究目的**

本课题针对网络服务提供商对流量进行监控的需求，解决以下两个方面问题：1）判断网络流量中是否包含针对特定网页的HTTPS访问，并统计给定时间内访问的次数，为网络舆情监控提供基础；2）识别HTTPS流量中用户分别访问了哪些网页，统计出特定时间内，用户访问最多的网页是什么，进而识别出热点网页，为网络内容监管提供支持。

由于用户与服务器之间的交互是经过HTTPS加密的，因此需要先收集并分析用户与服务器访问所产生的流量，根据TLS协议的实现细节，提取此次访问流量的特征，分析用户访问的是哪个网页。在对用户的访问数据进行辨别前，需要手工构建训练数据，选择适当的机器学习方法，训练出一个可以分类流量的分类器，确定某一类特征相似的流量属于某个特定的网页。可以使用不同的主机和浏览器，在多种网络条件下多次访问所关注的网页，并使用wireshark等抓包工具抓取这些流量，提取访问流量的特征并为其打上标签，完成分类器的训练。随后用户访问未知网页时所产生的流量就可以通过这个分类器，第一时间判断当前用户访问了哪些网页，以此来推测特定网页的访问热度。

**4、工程应用价值**

流量分类识别，是网络服务质量保证、网络规划建设、健康网络环境的重要环节，也是网络舆情分析的基础，随着互联网的发展，各种加密协议应运而生，这对流量识别带来了一定的挑战，本论文希望通过对网页建立指纹来对加密的流量进行分类识别，以此来识别经过HTTP加密的流量。

1. **国内外研究现状、发展动态**
   1. **TLS(Transport Layer Security Protocol)协议**

HTTP协议传输的数据都是未加密的，也就是明文的，因此使用HTTP协议传输隐私信息非常不安全。为了保证这些隐私数据能加密传输，需要使用另一种协议HTTPS。HTTPS在HTTP的基础上加入了TLS协议，该协议依靠证书来验证服务器的身份，并为浏览器和服务器之间的通信加密。HTTPS的流量安全依赖浏览器的正确实现以及服务器软件、实际加密算法的支持。

TLS的设计目标是构建一个安全传输层（Transport Layer Security），在基于连接的传输层（如TCP）之上提供密码学安全：

1. 保密性，message privacy (通过加密实现，所有信息都加密传输，第三方无法窃听 )
2. 完整性，message integrity（通过MAC校验机制，一旦被篡改，通信双方会立刻发现 ）
3. 可认证性，mutual authentication （双方都可以配备证书，防止身份被冒充 ）

TLS协议由握手协议和记录协议两部分组成，握手协议负责通信前的相关信息协商以及必要的认证，记录协议负责实际数据传输和相关控制命令传输，在握手阶段：

1. 客户端请求建立SSL链接，并向服务端发送一个随机数–Client random和客户端支持的加密方法，比如RSA公钥加密，此时是明文传输。
2. 服务端回复一种客户端支持的加密方法、一个随机数–Server random、授信的服务器证书和非对称加密的公钥。
3. 客户端收到服务端的回复后，加上新的随机数–Premaster secret 通过服务端下发的公钥及加密方法进行加密，发送给服务器。
4. 服务端收到客户端的回复，利用已知的加解密方式进行解密，同时利用Client random、Server random和Premaster secret通过一定的算法生成HTTP数据传输的对称加密密钥 – Session key。

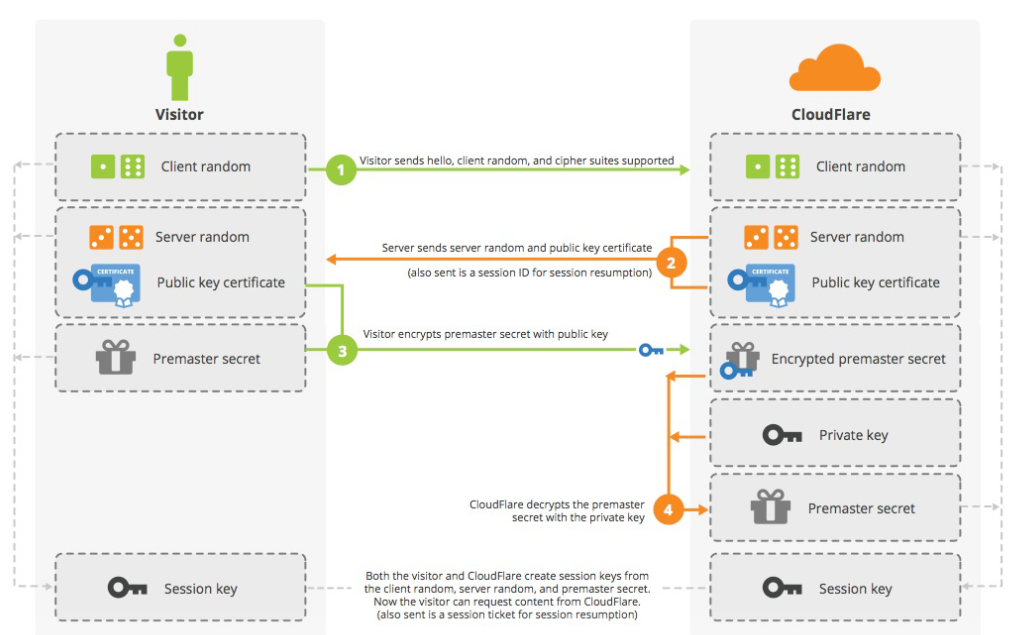


图1：TLS握手协议

此后的HTTP数据传输通过对称加密方式进行加密传输。TLS协议的报文结构如下图所示

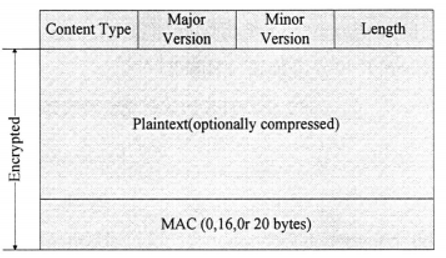


图2

内容类型(Content Type)是用以区分当前TLS记录协议承载的是什么内容，值为APPLICATION DATA(23)时，意味着此时承载的是应用数据，Length指图中整个加密数据的大小，当不启用压缩算法时，Length的大小不超过。当有压缩算法时，Length的大小不超过+1024。

* 1. **针对TLS协议的攻击**

针对TLS协议的攻击，可以分为主动攻击和被动攻击。

主动攻击又称为证书中间人攻击，该攻击过程需两个步骤：DNS欺骗过程和证书欺骗过程。当攻击者可以截取网关设备流量或者用户的DNS服务器被污染的情况下，可以无须DNS欺骗，直接进行证书伪造欺骗攻击，如果此证书是自行制作的，浏览器会给予警告通知，但当浏览器中的根证书颁发机构认定攻击者提供的证书安全时，该机构可能会认定攻击者的站点为可信任的，那么浏览器会接收该证书并以正常方式显示页面内容，无法区分攻击行为。

被动攻击又称为流量分析攻击，攻击者对网络信道中传输的数据进行窃听、采集，然后按照已有的策略进行处理，推断数据流的各种属性。流量分析的特点是对数据流量进行统计分析，进行数学建模，生成能明显区别于其它个体的特征，这些特征被称为指纹。

由于HTTPS隐藏了通信的实际内容，攻击者只能直接检测到参与通信的源IP和目的IP，这些信息量远远不能构成有效攻击。而HTTPS架构是HTTP Over TLS，攻击者从HTTP的相关特性入手，通过数据包搜集、时序记录等操作，有可能提取到相关HTTPS流量模式的特征。

针对HTTPS指纹攻击的第一步是指纹搜集，攻击者会寻找若干感兴趣的站点，通过访问该站点，搜集并捕获数据流，从中提取出对应的流量，利用数据挖掘、概率论等知识进行分析，寻找出有效的特征，我们称这些特征为该网站的指纹。

第二步随机采集一段加密数据流，按第一步的方法提取流量特征。

第三步，利用相关算法将指纹归类，分析这个当前流量是否为特定的网站

* 1. **现有加密流量识别手段**

研究人员经过大量的尝试后，发现对加密的网站进行攻击多为提前为已知流量建立分类器，根据已经建立的分类器，分析流量是否为特定网站，也有一些研究向真实网络环境的流量进行了挑战，即不提前预设流量是特定的网页，而是判断该流量是否来源于互联网的某站点。

Sun等人在文献[1]中指出，加密的通信仍然揭示了一些有关内容的信息，作者从浏览过程中抓取了https网页内嵌对象的数量和长度来识别目标网站，它们从不同的网站收集了大约10万个网页，然后对它们进行分类以获得一个标签数据库。此后，使用特征度量每两个网站之间的相似性。最后，应用Jaccard系数相似性度量来识别页面对象的长度，并通过改变相似度阈值来重复实验，以使假阳性率降到最低。

另一项工作是由Bisias等人完成的。他们在文献[2]中通过抓取Firefox的流量数据，得到经过OpenSSH隧道加密的数据流量信息，收集了一年左右的数据后，他们总结出每个HTTP流量的两个特征，即数据包的大小和到达时间，使用互相关度量来识别不同时间间隔的两个数据流量的相似性。尽管成功率较低，但取得了合理的效果。

Liberatore和Levine在文献[3]引入了一种用于识别网站指纹的新技术，使用带有密度估计的朴素贝叶斯分类器(NB)。这是通过识别加密数据包的特性、方向和大小来实现的。类似地，Herrmann等人在文献[4]中做了相同的实验(具有相同的特征)，但使用多项朴素贝叶斯(MNB)分类器。Herrmann方法的主要创新是统计了流量特征的出现概率。

Panchenko等人在文献[6]的研究中改进了svm（支持向量机）分类器在流量识别中的工作。分类器同时在封闭和真实网络环境的数据上验证通过，使用的是流量大小、到达时间和方向等特征。该项工作是首次成功地攻击真实网络环境网站，允许对ISP没有列出的网站进行指纹识别。

Shi等人通过对网络流量的分析，在文献[5]中提出了一种新的网站指纹识别技术，将输入和输出的数据包按照流量变换方向为区间进行分割，然后将其转换成向量，利用余弦相似度公式计算了观察到的向量与已知指纹的相似性，并对实验结果进行了理论和实践评价，证明了该技术在降低用户匿名性方面是有效的。

Cai等人在文献[7]中提出了攻击网站的新指纹方法。针对使用了tor加密的网页在Firefox浏览器中访问所产生的流量，作者通过使用tSharp 捕获数据包，为100个网站收集数据，所提出的分类器利用了数据包的大小、顺序、方向等特征以及其他有用信息，使用damerau-levenshtein 距离(DLD)算法来识别网页。此外，他们还使用隐式马尔可夫模型(Hmms)来识别哪一组网页与同一网站相关。

在文献[7]的基础上，Wang和Goldberg开发了一种新的指纹攻击[8]，称为最优串对距离(Osad)，这是damerau-levenshtein 距离(DLD)的一个改进版本，可以识别两个流量实例之间的相似之处。作者将该方法应用在tor这样的多跳匿名系统上，在准确度方面取得了令人满意的结果 。Panchenko 等人在文献[6]中进行了针对真实网络环境的实验，得到了更高的准确度。

我们根据相关文献中提到的分析加密流量的方法，对比其研究对象特征和准确率，制作了表1：

表1:加密流量识别的研究

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **相关文献** | **目标协议** | **特征** | **分类标准** | **流量范围** | **准确率** |
| Statistical identification of encrypted web browsing traffic[1] | Http流量 | 数据包数量，大小 | Jaccard 系数 | 特定网站 | 75% |
| Privacy Vulnerabilities in Encrypted HTTP Streams[2] | SSL流量 | 数据包大小，包传输间隔 | 滑动卷积 | 特定网站 | 40% |
| Inferring the source of encrypted http connections[3] | OpenSSH | 数据包大小，方向 | 朴素贝叶斯 | 特定网站 | 90% |
| Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier[4] | Tor网络 | 数据包大小，方向 | 多重朴素贝叶斯 | 特定网站 | 2.96% |
| Fingerprinting attack on the tor anonymity system[5] | Tor网络 | 数据包大小，方向 | 余弦相似度 | 特定网站 | 50% |
| Website finger printing in onion routing based anonymization networks[6] | SSH，TLS，IPSec | 数据包的尺寸，传入数据包百分比，数据包总数 | SVM | 真实网络环境 | 82% |
| Touching from a distance: Website fingerprinting attacks and defenses[7] | Tor网络 | 数据包大小，方向，顺序 | 使用SVM技术DL距离 | 特定网站 | 87.3% |
| Improved website fingerprinting on tor[8] | Tor网络 | 数据包大小，方向，顺序 | OSAD | 真实网络环境 | 91% |

* 1. **总结**

如表2所示，在讨论加密流量识别的过程中，往往需要对以下几个特征加以明确：

|  |  |  |
| --- | --- | --- |
| **特征** | **描述** | **可能的选择** |
| 范围 | 确定该技术应用的位置。 | 网站，无线网等 |
| 目标 | 实现该技术的协议或网络环境。 | HTTP,SSL,TLS,SSH,Tor等 |
| 特征 | 为了对网站或应用程序进行指纹识别，应该有一些特定的特征或模式，通过这些特征或模式，一种技术是否可以识别数据包与需要识别的网站或应用程序有关。 | 数据包大小、计数、方向、顺序、包间时间等. |
| 分类器 | 改进新指纹技术的主要部分是选择适合于训练和测试给定数据包集的机器学习分类器。 | SVM，HMM，朴素贝叶斯等 |
| 数据收集 | 大多数指纹技术收集大量数据并过滤，然后以离线的方式处理它们，另外一些技术会在线进行流量分析。 | 封闭的流量数据或真实网络环境 |
| 准确度 | 使用分类器对所获得的数据进行识别后，使用精确度，召回率，真阳性，假阳性等参数验证其准确度。 | 百分率 |

表2：流量分析领域关注的重点

各种识别方式，根据其应用的领域不同，其采用的特征和准确率都各不相同。

TLS协议本身会针对这些攻击做出相应的更新，以确保自身的安全，如：将所有流量填充到同样的长度，但是为抵抗这些攻击而提出的防御措施在现实部署中是不可行的，因为它们在增加延迟和带宽开销方面代价高昂。

代理服务器同样会对本课题产生很大的影响，代理的实现机制以及网络性能使这些可见的特征更加复杂化。

代理服务器转发客户端的请求到实际Web服务器，在接收到目标Web服务器的应答后，修改应答相关属性值后再转发到客户端，对于HTML页面，部分JavaScript，代理必须修改其URL，使其从原始的指向目标Web服务器的地址改成指向代理服务器的地址，在URL变换流程中，Proxy必须缓存被修改页面当前的全部信息，如整个HTML文本页面、JavaScript文本，并以正则表达式等规则进行替换，如果此页面过大，超出代理服务器设置的缓冲区，代理会返回页面过大的失败信息给客户端；对于实际图片、音频等数据流不需要翻译，代理会直接转发或缓存到数据接收完成再转发。这种代理够正常下载文件，观看视频，但每次待转发数据量是不固定的，从而转发后的数据包会存在大小不确定的非MTU，且服务器应答的对象越大，非MTU出现的概率越大，对攻击者而言是一种严重的噪声，破坏了站点固有指纹。

HTTP代理服务器一般会允许Cookies，这会改变HTTP请求流量，由于客户端浏览器实际在和代理服务器的Web站点通信，随着不断访问不同的站点，积累在客户端的Cookies也会成线性增长，从客户端的角度看，我们抓取到的数据是逐渐增大的，在使用数据包大小作为特征之一时，需要考虑到这个问题，我们在讨论时一般禁止使用Cookies。

上述论文中，仅有少数两篇论文提到了“针对HTTPS流量识别某个网页”，而且使用的方法准确率不高，因此我们在以上讨论的基础上讨论“HTTPS加密流量的识别准确率是否可以进一步提升”，以及贝叶斯算法能否对HTTPS的加密流量进行识别。

1. **研究目标与研究内容**
   1. **研究目标**

本文研究的目的是抓取用户访问HTTPS网页时产生的流量，通过分析流量的特征形成网页指纹，再使用朴素贝叶斯算法对流量进行分类，形成一个实时的网页分类器。通过使用朴素贝叶斯算法，提高流量分类的准确度。

* 1. **研究内容**
  2. 流量的抓取

在我们进行客户端识别和检测可疑甚至恶意活动之前，我们必须了解网络流量。因此，我们必须观察网络流量，以深入了解典型模式。具体而言，在这种情况下，我们必须检索包含尽可能多的不同模式的加密网络流量的记录。在此前提下，我们可以获得更有趣的结果，这些结果对网络管理员，安全从业者和科学界有用。

* 1. 流量特征的识别

出于本文的目的，我们必须确定建立SSL / TLS通信的选项有哪些，以及在实际流量中使用哪些选项。然后，我们必须找到哪些选项变化最大，以及这些选项的可变性是否会影响网页的特征。

为此我们需要深入了解网络流量，将不同浏览器与不同操作系统对流量特征产生的影响降至最低。

* 1. 使用朴素贝叶斯分类器分析流量数据

朴素贝叶斯分类器对流量的部分变化不敏感，因此在流量发生部分变化时有较好的鲁棒性，我们使用朴素贝叶斯分类器首先对实验数据进行分类，给特定的网页确定分类器，之后将收集到的数据通过分类器进行判断，得到最终的结论。

1. **实施方案及可行性分析**
2. **实施方案**

整个研究方案围绕如图3所示的流程进行



图3：研究方案

* 1. 数据采集

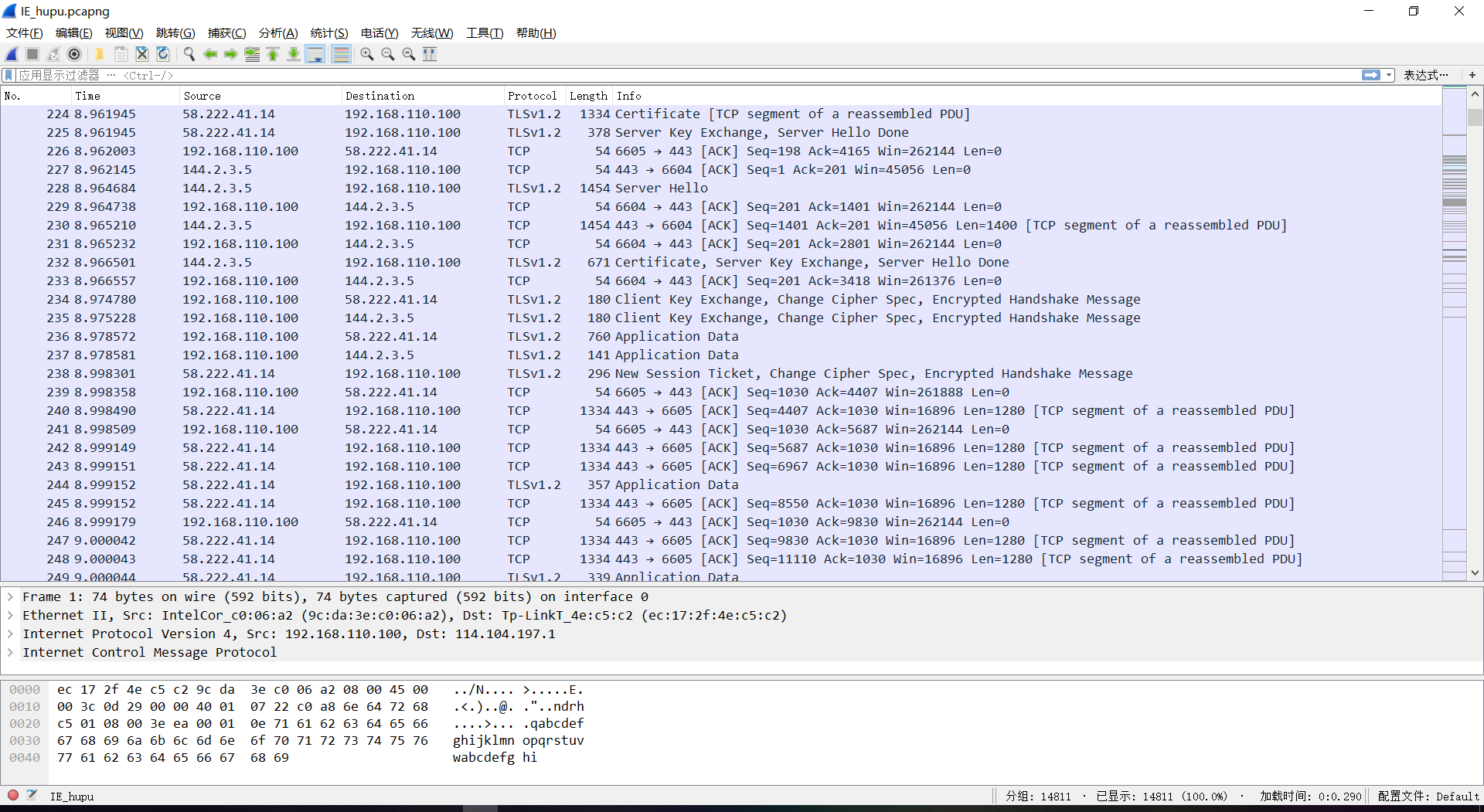
通过使用不同的主机和Web客户端浏览不同的服务器，生成不同的流量数据，所有数据包都是加密的应用层流量，从客户端应用程序的角度来看，Web页面的内容均来自目标服务器。通过流量捕获软件wireshark监控浏览器产生的SSL/TLS流量，客户端在不同机器上访问目标网页。

图4：wireshark抓取的流量

访问的网页来自Alexa.cn站点统计的热门网站top10，每个站点随机选取10个不同的页面，由于页面中包含新闻提要、状态更新等信息，因此每次访问同一页面时的流量可能是不同的，我们利用wireshark或tcpdump等工具对流量进行采集，数据收集的时间跨度为10天，每页平均每天采集40个样本。连续的数据收集至少间隔20分钟，以确保网页有足够的差异。作为对照组，我们还会收集几个静态网页的数据。对照组的网页随时间变化不大。来自实验组和对照组的跟踪是在相同的条件下收集的。

* 1. 过滤流量数据

通过观察采集到的流量数据，我们发现，长度为54的流量为TCP的ACK报文，这些ACK报文数量非常多，我们在统计时应该考虑删除掉这些报文，以免发生大量ACK报文掩盖掉报文特征的情况。

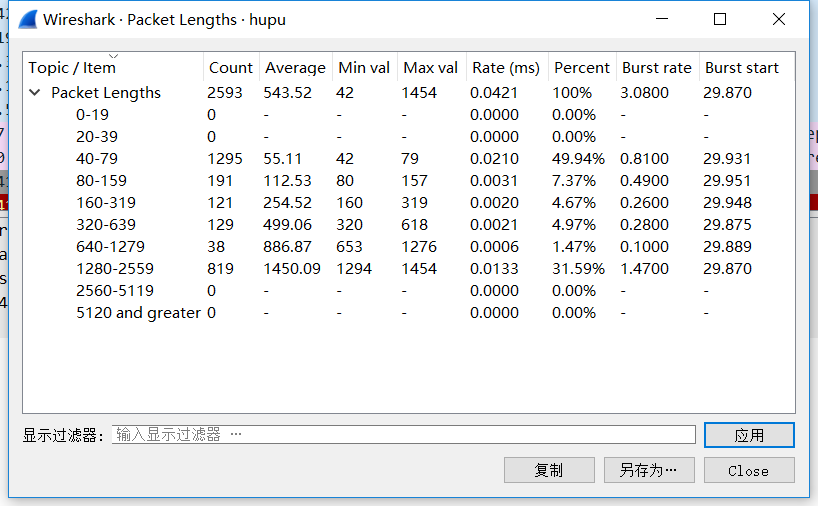


图5：wireshark报文大小统计

在现实的网络环境中，有些站点页面的页面结构比较复杂，所以在网络传输的过程中会采用分块编码传输的策略进行传输，代理程序需要缓存并重组各个应答报文块，并且去掉应答报文头的部分，将应答报文长度整理后记录。

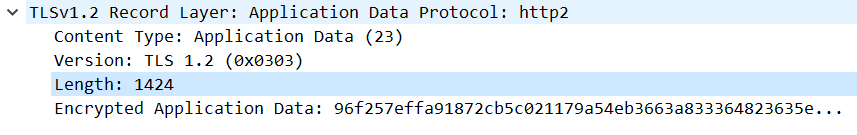


图6：TLS流量细节

最终，我们将访问某个页面的一次网络流量，按目标页面的URL作为索引，存储为json的形式



图7：经过整理后的流量格式

* 1. 分析网页指纹

网页指纹是由访问给定页面过程中产生的 HTTP 应答报文所组成的序列，它以从之前过滤出的流量数据中分析得出。

页面流量序列可以作为 HTTPS 页面的标识信息，但它不能直接作为访问该页面过程中，产生的 HTTPS 流量的流量标识。究其原因，是我们无法从 HTTPS 加密流量中，直接提取 HTTP 应答报文对象的大小。在这种情况下，就需要我们设法从 HTTPS 流量中，提取出适当的流量特征。提取出的特征，应当能够与已有的页面指纹按照一定的策略进行匹配计算，计算出的匹配度可用以衡量流量特征所标识的 HTTPS 流量与页面指纹所标识的 HTTPS 页面之间的相关度，从而，完成“识别特定 HTTPS 流量是在访问哪个 HTTPS 页面过程中产生的” 这一流量识别工作。

一方面，由于 HTTP 簇协议的基本通讯形式，是以请求应答对的形式呈现的，所以，即使经过了加密封装，通过报文传输方向，我们依然可以根据数据块传输方向分析出应用层协议传输过程中的请求/应答序列。另一方面，[13]中提到了一种 SSL／TLS 协议传输过程中协议信息的泄漏，根据这一漏洞，可以推算出 SSL／TLS 协议之上传输的应用层报文的大小区间，从而，对于 HTTPS 的应答序列中的每一个传输的密文数据块，我们能够推测出其明文报文大小区间。

SSL／TLS协议由握手协议和记录协议两部分组成，握手协议负责通信前的相关信息协商以及必要的认证，记录协议负责实际数据传输和相关控制命令传输。在客户端与服务器握手阶段，客户端会发送出自己所支持的密码族和压缩算法，服务器端应答会确定使用其中的一种密码族以及压缩算法。目前常用的压缩算法如表3，以加密方式的不同可将密码族分为流加密(Stream)和分组加密(Block)。:

|  |  |  |  |
| --- | --- | --- | --- |
| 密码族名称 | 密码族类型 | 哈希大小 | 分组大小 |
| TLS\_RSA\_AES\_1128\_CBC\_SHA | BLOCK | 20 | 16 |
| TLS\_RSA\_AES\_256\_CBC\_SHA | BLOCK | 20 | 16 |
| TLS\_RSA\_RC4\_128\_SHA | STREAM | 20 | N/A |
| TLS\_RSA\_3DES\_EDE\_CBC\_SHA | BLOCK | 20 | 8 |
| TLS\_ECDHE\_RSA\_AES\_128\_CBC\_SHA | BLOCK | 20 | 16 |
| TLS\_EDHE\_RSE\_AES\_256\_CBC\_SHA | BLOCK | 20 | 16 |
| TLS\_DHE\_DSS\_AES\_128\_CBC\_SHA | BLOCK | 20 | 16 |
| TLS\_DHE\_DSS\_AES\_256\_CBC\_SHA | BLOCK | 20 | 16 |
| TLS\_DHE\_DSS\_3DES\_EDE\_CBC\_SHA | BLOCK | 20 | 8 |
| TLS\_RSA\_RC4\_128\_MD5 | STREAM | 16 | N/A |
| TLS\_NULL\_NULL\_NULL | NULL | 0 | 0 |

表3：TLS加密流量算法

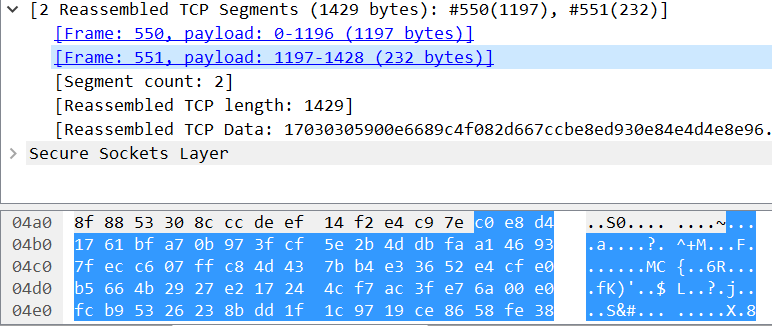
其中，块加密算法的原理是将固定长度的未加密数据块，利用事先约定的密钥，转换成长度相同的代码块。而流加密的原理是利用密钥产生随机码，利用该随机码与数据流作异或运算产生机密数据流。对于应用层 HTTP 报文，SSL/TLS 数据的加密示意图如图8。



图8：TLS数据加密示意图

基于以上两点事实，可以给出一个HTTPS 流量特征的概念。对于之前加密的 HTTPS 流量，可以通过 HTTPS 密文应答数据块反推出 HTTPS 明文应答报文的尺寸。

有时，一个经过TLS加密过后的TCP segment的流量过大，无法放入一个TCP报文中传输，TCP协议会对已经加密的流量再次切割成固定大小的流量包，要获取TLS流量的大小，必须对收集到的数据进行处理。



* 1. 流量识别

朴素贝叶斯是经典的机器学习算法之一，也是为数不多的基于概率论的分类算法。朴素贝叶斯原理简单，也很容易实现，多用于文本分类，比如垃圾邮件过滤等。

逻辑回归通过拟合曲线（或者学习超平面）实现分类，决策树通过寻找最佳划分特征进而学习样本路径实现分类，支持向量机通过寻找分类超平面进而最大化类别间隔实现分类。相比之下，朴素贝叶斯独辟蹊径，通过考虑特征概率来预测分类。

c表示的是随机事件发生的一种情况。x表示的就是证据，泛指与随机事件相关的因素。

* ：在x的条件下，随机事件出现c情况的概率。（后验概率）
* ：（不考虑相关因素）随机事件出现c情况的概率。（先验概率）
* ：在已知事件出现c情况的条件下，条件x出现的概率。（后验概率）
* ：x出现的概率。（先验概率）

现在我们有了数据集D，想要获取计算D中c的各个情况出现的频率即可。比如计算，直接用情况出现次数在所有情况中所占的比例值即可。

而获得就略显困难，因为x往往包含多个相关因素（是一个多种因素构成的向量），即它可能有多个需要考虑的属性值：=(,,,...,)

任意一个 都代表了所有相关因素中的其中一个。因此当x是一个向量时，我们若要计算，实际上就是要计算。这个理论上也是可以利用我们的数据集D来进行估计的，但是现实情况是，n的值往往非常大（属性非常多），而我们的数据集往往不能保证我们的样本包含了属性值的所有可能组合，这些样本很可能仅仅是我们的数据集中没包含到，即“未被观测到”，但不代表它们现实中“出现概率为0”。于是这就给我们计算出真实合理的目标值造成了障碍。

这里引入了朴素贝叶斯的概念，这个假设认为每个属性取它的各个值的可能性是独立的，与其它属性的取值不相关。在这个基础上，公式可以转化为：

属性条件独立性假设实际上是忽略掉了某些属性之间可能存在的关联，假设属性的取值可能性都是独立的，以简化的计算为

按照这种方式计算，我们前面谈到的“数据集无法覆盖所有可能的属性组合的样本”问题也就基本解决了，因为根本不需要有那么多样本，保证每个属性的各个取值都能有样本取到过就行了。

* 1. 分析数据与分类

经过训练后，进行测试用例来衡量所使用的分类器是否能够识别跟踪身份。例如，如果访问“www.google.com”40次，那么训练过程的次数将为36次，其余的4次将用于测试。通常，训练和测试是通过几次随机选择进行的，每次试验结果的精度是独立计算的。

分类准则用公式表达如下：

代表一个由Naive Bayesian (nb)算法训练出来的hypothesis（假设），它的值就是贝叶斯分类器对于给定x的因素下，最可能出现的情况c。y是c的取值集合。

等式右边的式子就是argmax加上基于属性条件独立性假设的贝叶斯定理的分子部分（分母P(x)前面提到过，可以在计算中略去），即我们的输出值就是这样的一个c，这个c使的计算结果值最大。

最后，最终的成功率结果是通过平均所有的实验结果精度来计算的。最终给出关于TLS加密流量对应网页的分析与评估结果，并实现热门流量的第一时间反馈。

1. **参考资料**

[1]Q. Sun, D. R. Simon, Y.-M. Wang, W. Russell, V. N. Padmanabhan, and L. Qiu, “Statistical identification of encrypted web browsing traffic,” in Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on. IEEE, 2002, pp. 19–30.

[2] G. D. Bissias, M. Liberatore, D. Jensen, and B. N. Levine, “Privacy vulnerabilities in encrypted http streams,” in Privacy Enhancing Technologies. Springer, 2006, pp. 1–11.

[3] M. Liberatore and B. N. Levine, “Inferring the source of encrypted http connections,” in Proceedings of the 13th ACM conference on Computer and communications security. ACM, 2006, pp. 255–263.

[4] D. Herrmann, R. Wendolsky, and H. Federrath, “Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naive-bayes classifier,” in Proceedings of the 2009 ACM workshop on

Cloud computing security. ACM, 2009, pp. 31–42.

[5] Y. Shi and K. Matsuura, “Fingerprinting attack on the tor anonymity system,” in Information and Communications Security. Springer, 2009,pp. 425–438.

[6] A. Panchenko, L. Niessen, A. Zinnen, and T. Engel, “Website finger printing in onion routing based anonymization networks,” in Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. ACM, 2011, pp. 103–114.

[7] X. Cai, X. C. Zhang, B. Joshi, and R. Johnson, “Touching from a distance: Website fingerprinting attacks and defenses,” in Proceedings of the 2012 ACM conference on Computer and communications security. ACM, 2012, pp. 605–616.

[8] T. Wang and I. Goldberg, “Improved website fingerprinting on tor,” in Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society. ACM, 2013, pp. 201–212.

[9] C. V. Wright, L. Ballard, S. E. Coull, F. Monrose, and G. M. Masson, “Spot me if you can: Uncovering spoken phrases in encrypted voip conversations,” in Security and Privacy, 2008. SP 2008. IEEE Symposium on. IEEE, 2008, pp. 35–49.

[10] M. Tavallaee, W. Lu, and A. A. Ghorbani, “Online classification of network flows,” in Communication Networks and Services Research Conference, CNSR’09. Seventh Annual. IEEE, 2009, pp. 78–85.

[11] F. Zhang, W. He, X. Liu, and P. G. Bridges, “Inferring users’ online activities through traffic analysis,” in Proceedings of the fourth ACM conference on Wireless network security. ACM, 2011, pp. 59–70.

[12] J. Atkinson, O. Adetoye, M. Rio, J. Mitchell, and G. Matich, “Your wifi is leaking: Inferring user behaviour, encryption irrelevant,” in Wireless Communications and Networking Conference. IEEE, 2013, pp. 1097–1102.

[13] 吴家顺. Website指纹识别攻击与防护技术研究[D]. 南京理工大学, 2013.