文章编号:1002-1175(2011)05-0659-09

一种基于 Web 通信行为的抗审查隐蔽通信协议*

谭庆丰†, 时金桥, 郭 莉, 王 啸

(中国科学院计算技术研究所信息安全研究中心,北京市信息内容安全技术国家工程实验室, 北京 100190)

(2010 年 8 月 7 日收稿; 2010 年 11 月 7 日收修改稿)

Tan Q F, Shi J Q, Guo L, et al. A censorship-resistant covert communication protocol based on Web communication behavior[J]. Journal of Graduate University of Chinese Academy of Sciences, 2011,28(5): 659 – 667.

摘 要 以往基于 TCP/IP 协议或 HTTP 协议的隐蔽通信方式通常是利用协议本身各个字段的特点,将信息隐藏在协议的各个字段中.这种方式往往会具有某种结构特征,而基于计时的隐蔽通信又往往具有某种流模式.基于非对称通信理论和马尔科夫模型的 Web 行为预测,提出一种基于 Web 通信行为的抗流量审查隐蔽通信协议.重点描述隐蔽通信协议和原型系统,并分析协议的安全性.测试结果表明,该方法具有很高的效率和安全性.

关键词 隐蔽通信, 抗审查, Web 通信行为

中图分类号 TP391

隐蔽通道的概念最初是在计算机系统领域由 Lampson 于 1973 年提出:即如果一个通道既不是设计用于通信,也不是用于传递信息,则称该通道为隐蔽通道.在信息论领域,隐蔽通道通常被定义为一种寄生于其他信道,没有寄主信道的设计者、持有者与维护者的授权与识别,秘密地进行信息传递的信道.几乎所有的隐蔽通道都占有合法通道的带宽资源.但是,占用的这些带宽资源通常是没有被使用的,或者是未确定的行为,因此,隐蔽通道可以很好地隐藏自身,传统隐蔽通道研究主要集中在操作系统中.随着网络技术的发展,基于网络的隐蔽通道越来越受到研究者广泛的关注[1-3],本文关注的主要是基于网络隐蔽通道的隐蔽通信技术.

网络流量审查是网络隐蔽通信面临的最大挑战之一.审查者可能通过特征分析、协议分析或行为统计分析等方法发现异常通信流量,检测隐蔽通信信道.同时,审查者可能通过流量规格化来消除网络隐蔽通信信道.因此,在设计隐蔽通信协议时需要考虑到隐蔽、安全、可靠等问题.在隐蔽信息传输过程中,如果采用 SSL 加密隧道进行通信,虽然其通信内容不可见,但可能因为加密连接而引起怀疑.基于网络协议的隐蔽通道是隐蔽通信经常采用的传统方法,然而这种方法信道容量比较低,并且容易受到流量规格化的攻击.

本论文关注的是隐蔽通道协议的设计,其核心思想是利用 HTTP 协议中上传隧道和下载隧道的不对称性,将隐蔽通信协议的命令信息隐藏在上传 Web 通信行为中,将信息传输过程中的内容信息利用隐写术隐藏在下行网络流量中,从而抵御流量审查和流量规格化攻击;同时在信息传输过程中,采用加密认证的方式保证信息内容的安全性,利用免费代理和志愿者代理来达到抵御追踪的目的,从而在隐蔽通信客户端和隐蔽通信服务端之间构成一条隐蔽隧道.

隐蔽通信协议设计的最大挑战在于通信行为的私密性,同时又有较好的性能,此外隧道协议还应该

* 国家重点基础研究发展计划(973)项目(2007CB311100)资助

†E-mail:tanqingfeng@ software. ict. ac. cn

能够抵御各种被动攻击和主动攻击方法,如攻击者通过流分析修改包头协议、篡改消息和会话,或者伪装成一个“合法”的 HTTP 请求等. 为了验证协议的可行性,本论文设计了一个原型系统,测试结果表明:该系统为 Web 浏览提供很好的通信带宽,并具有很好的隐蔽性.

1 相关工作

在互联网发展初期,人们通过代理,或者匿名代理绕过审查,如 Anonymizer. com 和 Zero Knowledge[4] 提供匿名的 Web 浏览,通过加密 HTTP 请求来保护用户的隐私. 然而这种匿名代理的方式存在单点失效问题,基于 SSL 的加密连接,内容虽不可见,但是其加密连接的行为会引起怀疑,甚至有些组织会对所有的加密连接进行直接过滤. 后来研究人员提出 Mix – net 和洋葱路由思想,并根据此思想设计了许多匿名通信工具,如 TOR,JAP,I2P 等[5-7]. 然而匿名通信技术本身只能隐藏信息发送者和接收者的身份,或者隐藏谁跟谁在通信,而不能解决通信的隐蔽性问题,而隐蔽通讯的目标是避开上述问题,通过在预先存在的正常通信行为中隐藏额外的信息.

互联网用户往往拥有较高的下行带宽与较低的上行带宽,许多通信技术利用这种不对称性提高通信性能. Adler 与 Maggs[8] 考虑是否可以利用服务器的高带宽去帮助用户发送消息以提高通信性能. 在已知消息分布的情况下,他们证明上述想法是可行的,并且在理论上提出了服务器利用它们所掌握的信息去减少客户端发送请求所需的字节数. Adler 与 Maggs 的工作被后来的许多研究者改善或扩展. Gagie[9] 提出了 4 种协议. 其中 dynamic bit-efficient split(DBES)是对 Adler 等人的 bit-efficient split(BES)协议的改进,该协议不需要服务器事先知道消息的分布. Feamster 等人基于 Adler 与 Maggs 不对称通信理论设计出一个抗流量审查的隐蔽通信系统 Infranet[10]——在 HTTP 中使用隐蔽通道以绕过审查员,即通过 Infranet 将一个隐蔽的 HTTP 请求编码为一系列正常的掩体 HTTP 请求,并利用隐写术将目标内容隐藏在掩体资源文件的图片中,然而测试发现 Infranet 具有迭代次数过多、延时过大等缺点.

本协议设计的目标是抵御网络流量审查,并提供统计上可抵赖性. 此外,协议还保证隐蔽通信服务端的隐蔽性. 并在隐蔽通信客户端和服务器端建立一个安全的隐蔽通道,利用 HTTP 协议中上传隧道和下载隧道的不对称性,把客户端目标 HTTP 请求映射为一系列正常的 Web 请求. 隐蔽通信服务器端解码这个 HTTP 请求序列,获得目标 URL. 相对于其他的抗流量审查的隐蔽通信方法,该方法更加难于检测和阻塞,其通信行为更加难于发现,因此具有更好的隐蔽性.

2 系统架构

这节主要陈述系统的协议设计和系统架构. 首先给出系统的总体架构设计目标,然后详细阐述协议设计的各种考虑,其架构如图 1 所示.

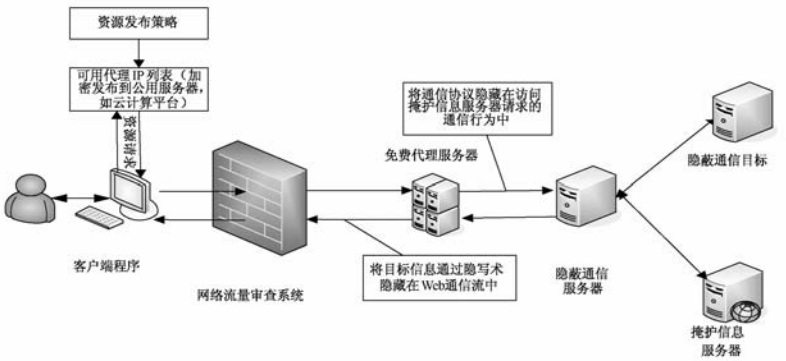

图 1 系统架构图

2.1 概述

如上图所示,系统主要包括以下 3 个部分.

1)隐蔽通信客户端代理　客户端代理接收用户请求,通过隐蔽通信协议将用户请求编码后经代理服务器发送给隐蔽通信服务器;同时接受隐蔽通信服务器返回的结果,提取隐藏信息,通过排序、重组、解密等方法恢复原有信息,并提交给用户. 此外,客户端代理还与服务器进行会话协商、协议初始化和认证等工作.

2)代理资源发布模块　为了更好地隐藏隐蔽服务器,需要设计一个代理资源发布模块. 该模块可以定时发布一批可用的代理资源列表,这些代理资源是通过加密的,其中代理资源发布策略是这个模块设计的重点,在此,通过设计一种资源发布策略使得即便某些恶意节点能够拿到这些代理也不可能拿全所有这些代理资源,以实现隐蔽服务器的私密性即在统计学上的隐蔽性.

3)隐蔽通信服务端　通过客户端的通信行为解码获取用户目标资源信息,则其一方面访问目标服务器获取用户的目标信息;另一方面访问掩体网站,获取掩体信息. 利用信息隐藏方法将用户目标信息隐藏在掩体信息中,然后通过代理服务器转发到客户端. 此外,服务器还与客户端代理进行会话协商、协议初始化和认证等工作.

隐蔽通信服务器端的重点是如何设计一个更好的码书,以实现客户端在统计学意义上的可抵赖性,高通信健壮性以及低的时延. 设计思路是采用自适应映射算法即通过建立马尔科夫模型,并通过增量的学习来预测下一个最有可能的请求页.

2.2　设计目标

因此,我们采用这种隐蔽通信的架构,其目的是设计一个具有如下功能目标的隐蔽通信系统:1)客户端可抵赖性;2)控制服务器的私密性(统计学上的私密性);3)通信的健壮性;4)较好的性能(包括带宽,延时等).

2.3　通信协议

通信协议的设计是本文的关键部分,也是重点和难点,如何设计一个具有安全、可靠、隐蔽匿名的通信协议是本系统的最大挑战. 论文的主要思想是利用 Web 通信行为来实现这一目标,即将隐蔽通信协议的命令信息传送隐藏在上传 Web 通信行为中,将信息传输过程中的目标信息利用隐写术隐藏在下行网络流量中. 隐蔽通信技术相对于加密连接、代理、匿名通信等技术有什么特点,采用的这种隐蔽通信方案有什么优势? 这正是我们要回答并试图解决的问题.

首先,在有的通信系统或者场景中限制加密技术的使用(使用加密技术可能会引起审查者的怀疑,进而对相应的行为进行监控);其次,隐蔽通信的目标是避开上述问题,通过在预先存在的正常通信行为中隐藏额外的信息;最后,由于 Internet 上有大量的数据流和各种不同的协议,因此互联网是一个非常理想的场所,更是一种作为隐蔽通信的工具,相对于基于网络协议的隐蔽通信和基于存储隐蔽通信,基于 Web 的通信行为的隐蔽方式可以更好地抵御网络审查员的审查. 下面重点阐述通信协议构建的 3 个主要部分. 首先给出系统客户端和服务端通信的顺序图,如图 2 所示.

2.3.1　隧道建立

建立私密通信隧道是为了更好地抵御审查,以一种安全、隐蔽、匿名的方式与隐蔽服务器通信. 隧道建立包括 2 个阶段,第 1 阶段是请求可用的代理资源;第 2 阶段是会话初始化阶段,即通过代理与隐蔽通信服务器建立连接,初始化会话,并与服务器交换共享密钥.

2.3.2　命令通道

隐蔽通信的一个关键点在于请求信息的编码和解码,即如何将用户请求的目的 URL 转换成一系列掩体 URL,通过请求这些掩体 URL 来获取目的 URL. 命令通道是一系列的 HTTP 请求,通过 HTTP 请求行为来隐藏目标 URL 地址. 其主要思想是客户端通过产生一系列可见的 HTTP 请求, 并把该 HTTP 请求序列的 URL 映射到客户端秘密的不可见的 URL,即目标 URL. 其方法是通过服务器端动态产生码书,客户端共享此码书, 通过码书把可见的 URL 映射到客户秘密的不可见的目标 URL.

由于 HTTP 协议上行通信的信道容量较低,而下传通道的信道容量较高,因此可以利用 HTTP 协议信道容量的不对称性,把命令信息隐藏在上行通信中,把目标信息隐藏在下行通信流中,通过信息隐藏

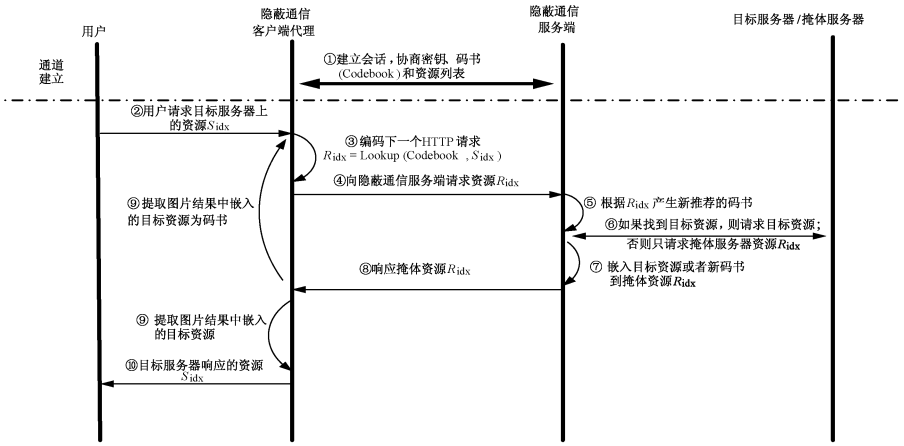

图 2　客户端和服务端通信的顺序图

的方法来秘密传输目标信息. 服务器端通过自适应范围映射算法产生码书,然后把码书隐藏在图片中,返回给客户端,客户端收到这个请求的页面,并提取隐藏在图片中的码书(即客户端和服务器端共享的码书). 在每一次迭代查找的过程中,服务端响应一个有序的 URL 对 (s_i, r_i), 其中 s_i 为目标域名下,当前访问统计上最可能的链接地址或者是被分割的链接地址,即为一个不可见的目标 URL 或者 URL 的一部分;r_i 为下一个 HTTP 请求,这里的 r_i 为掩体服务器的 URL. 当客户端收到码书后,通过 Encode 函数编码,获得下一个 HTTP 请求,其伪码如下所示:

```
Procedure  Encode(CodeBook, URL)
Input 客户端和服务器端共享的码书(Codebook)和当前的 URL,
//码书为一个字典顺序的(s_i, r_i)对,r_i 为通过编码后的下一个请求的 URL.
Output 下一个 HTTP 请求的 URL(即 r_i)
Begin
    For each (s_i, r_i) do
        If  S_i > = url: break
    End for
    把与 s_i 相对应的 r_i 赋值给 R
    Return R
End
```

其中下一个最有可能 URL 的预测问题,可以通过如下概率模型来解决. 如果 W 为某用户 Web 会话序列,其长度为 l,即在此之前已经访问了 l 个 Web 页面,$P(p_i \mid W)$ 为用户访问 W 以后下次访问页面 p_i 的概率,通过式(1)可以计算出下一个 Web 访问页面 p_{l+1} 的概率,其中 P 是某个站点所有页面的集合. 因此,由式(1)可以计算出所有可能被访问的页面 p_i 的概率,然后选择概率高的作为最有可能访问的下一个 Web 页面.

$$p_{l+1} = \arg\max_{p \in P}\{P(p_{l+1} = p \mid W)\} = \arg\max_{p \in P}\{P(p_{l+1} = p \mid p_l, p_{l-1}, \cdots, p_1)\}. \tag{1}$$

为了编码上行通道的 HTTP 请求,客户端每次动态查找来自服务端的码书,然后作为下一个 HTTP 请求. 因此,码书的设计是设计隐蔽通道的一个重点,也是实现隐蔽通信的关键,通过把命令编码转换成流量监管系统不被怀疑的 HTTP 请求信息,并在服务器端可以解码还原出用户目标请求. 本文的思路是联合 Infranet 的范围映射算法和基于马尔科夫模型的 Web 访问预测生成一个动态字典即码书,这个算法称为自适应范围映射(adaptive range map algorithm). 通过该算法来动态产生码书,预测用户下一个最有可能请求的 URL. 下面重点介绍几种产生码书的映射算法,通常映射的方法有如下几种.

1)静态字典　实现起来方便,简单,而且性能很高;但是安全性,私密性很低. 审查者通过对网络流

的监视,可能统计出码书的字典,如图 3 所示.

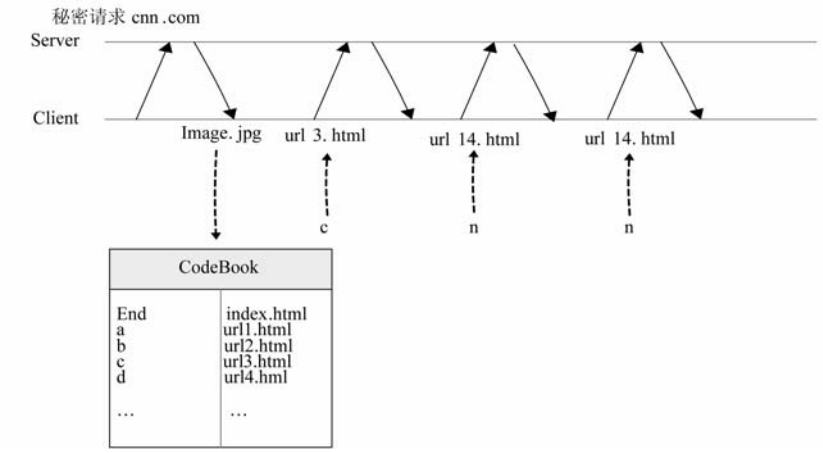

图 3　静态字典示意图

2)范围映射(rang map)　Adler 和 Maggs[4] 提出非对称通信模型,并给出一个相应的非对称通信协议.在这种通信模型中,由于隐蔽通信服务端可以知道所有来自客户端的知识即 HTTP 请求信息,而客户端只知道它自己的请求信息.因此,服务器端可以精确计算出来自所有客户端请求信息的概率分布,这样就可以利用这种非对称通信通道维护掩体服务器上所有存在的 Web 页面 P 的频率分布,即客户可能发送到服务端的 HTTP 请求(URL)和该 HTTP 请求所对应的频率,以此来减少隐蔽通信客户端发送到隐蔽通信服务端的请求次数.在 Infranet 的范围映射中,作者基于这种非对称通信协议,改进了原作者算法的缺陷,提出一种基于范围映射的算法来产生码书.该算法设计相对比较复杂,但安全性比较好,而且每次都会动态地产生码书.图 4 为基于范围映射秘密请求 cnn.com 的示意图,其中左边是不可见的 HTTP 请求(目标 URL),右边是来自掩体服务器的 Web 页面,作为下一个编码后的可见 HTTP 请求,通过范围映射算法可以保证在当前请求下,下一次在统计上最有可能发送到隐蔽通信服务端的可见的 HTTP 请求.

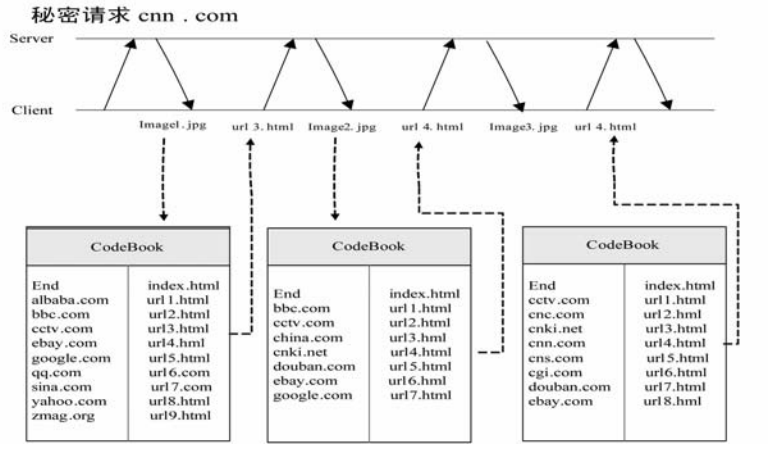

图 4　范围映射算法示意图

3)自适应范围映射　自适应范围映射在范围映射基础之上,联合客户端的用户偏好、使用习惯,去预测下一个 Web 请求行为模式.即对于服务端的预测建立一个马尔科夫模型,去预测下一步用户最有可能会请求的 URL_1, URL_2,…,URL_n,然后生成码书.下面重点介绍自适应范围映射算法.

自适应范围映射包括预处理,马尔科夫预测模型和自适应算法 3 个模块.预处理模块对 Web 日志进行预处理,产生一个 session 序列,并保存为一个 XML 文件.其主要工作是会话识别、用户识别和访问

页识别,该主题不在本文讨论范围之内. 基于马尔科夫模型在 Web 访问行为预测方面的研究,在文献[11-13]中有详细讨论. 本文借鉴 Brian 等人在文献[14]中提出的算法思想,设计一个增量学习的全 k 阶马尔科夫树,通过频率剪枝来提高他们的预测精度,降低时间复杂度. 由于全 k 阶马尔科夫模型具有更好的时间和空间复杂度,而且能够具有更好的预测精度和适应范围. 算法的重点是利用建好的马尔科夫模型预测下一步用户最有可能访问的 Web 页面,然后联合范围映射生成码书,增量建立一个全 k-order 的马尔科夫树. 算法伪码如下所示:

```
//将 HTTP 请求序列,增量插入马尔科夫树中
Procedure BuildMarkovTree( session, root)
Input: session 为 HTTP 请求序列, root 为马尔科夫树的根节点
Output: 马尔科夫树
Begin
    Ptr = root;
    Count = min( |session|, k) //k 为马尔科夫阶数
    For i = 0 to count
        Copy(session.end() - i, session.end(), back_inserter( S ) )
        If | S | = = 0:
            Ptr - > selfCount + +
        Else
            For iter = S.begin() to S.end()
                Ptr - > childCount ++
                If not_exit_first_child( iter, ptr )
                    Ptr - > numChildren ++
                    addFirstChildNode(iter, ptr)
                else if not_exit_next_sibling( iter, ptr )
                    ptr - > numchildren ++
                    addNextSiblingNode( iter, ptr )
                End If
                If iter = = session.end()
                    Ptr - > selfCount ++ ;
                End If
            End For
        End If
    End For
End
```

由于存在 P 个页面的 Web 站点,在全 k 阶马尔科夫模型中总共有 $\Theta(|P|^k)$ 个状态,因此需要对其进行剪枝,通过条件概率来估算那些在统计上最不可能发生的状态,其中最常见的方法就是通过最大似然原理. 如下所示,可以通过计算出$\langle S_j^k, P_i\rangle$序列的访问次数,和 S_j^k 序列的访问次数,然后通过式(2)来计算条件概率 $P(p_i|S_j^k)$.

$$P(p_i \mid S_j^k) = \frac{\text{Frequency}(\langle S_j^k, p_i\rangle)}{\text{Frequency}(S_j^k)}. \tag{2}$$

自适应算法是一个增量学习和剪枝的过程,通过当前的 session 增量建立并维护一个马尔科夫树,并裁减在统计上频率的期望值比较低的枝叶节点. 因此,自适应范围映射算法实质是一种动态字典,它改进了范围映射迭代次数过多、时延太大的缺点,旨在设计一个更加健壮、低时延的隐蔽通信系统. 下面

为联合范围映射和基于马尔科夫模型的 Web 访问行为产生的码书伪代码:

```
//利用马尔科夫模型产生的推荐 URL 序列和 Infranet 的范围映射产生的动态字典生成一个新的码书
Procedure GenCodeBook(SS, dict)
Input: SS 为利用马尔科夫模型产生的推荐 URL 序列,dict 为范围映射产生的字典,其中 dict 为(s_i, r_i)对.
Output: 产生新的码书
Begin
For each s in SS
        //在字典 dict 中查找 s_i,使得 s_i 刚好大于等于 s,并返回相对应的 r_i
        R = Search(s, dict)
            //更新字典 dict,即把(R, s)添加到字典 dict 中
        dict = dict ∪ (s, R)
End for
        return dict
End
```

2.3.3 数据通道

数据传输通道包括 2 个方面,一个是传输码书即传输命令;另一个就是利用隐写术把目标信息隐藏在掩体信息中,然后通过数据传输通道返回到客户端. 因此设计数据通道主要要考虑目标数据的隐藏以及网络性能问题即带宽、时延.

当获得资源即目标信息后,利用 Outguess、F5 等信息隐藏工具将用户请求的资源(网页链接、图片)等嵌入到掩护信息中,然后利用数据通道返回给客户端. 而设计一个较低延时的隐蔽通信系统是系统可用性的一个重要方面. 本文的做法是利用 Web 预取技术和缓存技术来解决. 通过马尔科夫模型预测下一个最有可能的页面,然后预取存到隐蔽服务器端,通过预取,并采用 Cache 技术来降低时延. 如果 Cache 的页面过多可能会造成性能过低,因此必须采用某种算法去置换最近没有请求的资源,在此我们采用 LRU 算法.

3 实验结果及性能分析

这节讨论通信协议的实验结果并分析其性能. 本文设计了一个原型系统来验证我们提出的隐蔽通信协议,该系统服务端运行的硬件平台:CPU 主频 3.0 GHZ,内存 1 G;操作系统为 Ubuntu 发行版 9.04;支撑环境:Java Runtime Environment 6.0、Tomcat 服务器 6.0、Squid 2.6. 客户端运行环境为硬件平台:CPU 主频 3.0 GHz,内存 1 G;操作系统:Linux Ubuntu 发行版 9.04. 主要测试了上行通信和下行通信过程中码书的交互次数,也就是迭代查找目标资源的次数;同时,还测试了每次交互的时延. 测试方案为任意选择一个站点作为目标服务器,然后随机访问其中的 50 个 Web 页面;访问结束 Sleep 一段时间后又继续随机访问这 50 个页面,并比较我们的自适应范围映射算法和范围映射算法在交互次数、时延等方面的区别. 图 5、图 6 为迭代次数对比图,即客户端和服务端交换码书的次数,图中无马尔科夫推荐为原算法的范围映射算法,而有马尔科夫推荐则是自适应范围映射的算法.

从图中可以看到自适应范围映射算法第 1 次访问序列没有明显优化,而在第 1 次访问之后的访问会有明显的优化,即大部分情况只需 2 次就可以找到目标资源;而范围映射其交互次数不管访问多少次其性能都没有明显改进,其在码书大小为 16 的情况下,大部分目标 URL 的迭代查找次数为 6,在码书大小为 8 的情况下为 7.

在时延方面,范围映射算法无论是第 1 次访问某个站点还是后续访问,请求每个页面的平均时延为 4 ~5s 之间,其码书大小对时延没有明显影响. 而我们的算法平均时延为 3.5 ~4.5s 之间.

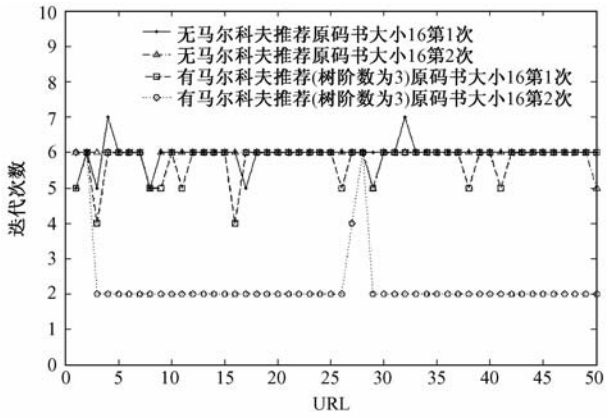

图 5　码书大小为 16 的迭代次数对比图

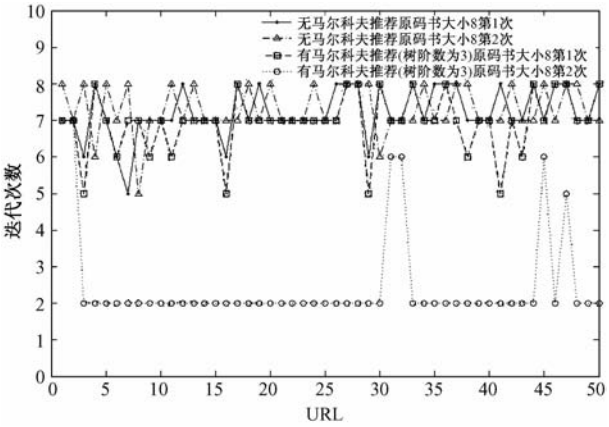

图 6　码书大小为 8 的迭代次数对比图

在通信的信道容量方面,我们的隐蔽通信服务器为一个图片网站的镜像,其平均每页图片信息为 600k 左右,如果按照信息隐藏嵌入量为 20% 估算,则平均信道容量为 120k,即返回结果可嵌入 120k 的隐蔽信息. 另外,根据最新统计,互联网上网页的平均大小为 312k,按 gzip 压缩比 30% 计算,压缩后的内容为 93. 6k,如果 1 页信息隐藏不够,可以采用多页信息隐藏.

4　结束语

本文提出一种基于 Web 通信行为的抗审查隐蔽通信协议,该方法联合非对称通信理论和 Web 行为预测方法,利用 HTTP 协议中的上传隧道和下载隧道的不对称性,将命令信息隐藏在一系列的 HTTP 请求中,将目标资源隐藏在掩体资源中. 其优点是该方法没有增加每次 HTTP 请求的额外负载,每次 HTTP 请求行为都是一次正常的 Web 请求;另一方面,该协议没有采用加密算法,来显式加密 HTTP 连接. HTTP 头字段或者净载荷,而是通过信息隐藏算法把信息隐藏在掩体资源中. 在通信过程中,利用非对称通信理论,在服务器端计算出掩体资源的概率分布,保证每次 HTTP 请求都是在当前请求下概率最大的,从而保证协议通信行为具有统计学意义上的可抵赖性. 最后,利用马尔科夫的 Web 行为预测方法,来预测目标资源的请求. 试验结果显示,我们的协议显著改善了以前方法中性能低下,隐蔽通信客户端和服务端交互次数过多等问题.

参考文献

[1] Handel T, Sandford M. Hiding data in the OSI network model[C]// Anderson R. Information Hiding Workshop (IH 1996). Cambridge, UK: Springer, LNCS, 1996, 1174: 23-38.

[2] Murdoch S, Lewis S. Embedding covert channels into TCP/IP[C]// Proc 7th Information Hiding Workshop. 2005.
[3] Cabuk S, Brodley C, Shields C. IP covert timing channels: Design and detection [C]// Proceedings of the 2004 ACM Conference on Computer and Communications Security. 2004.
[4] Anonymizer[EB/OL]. [2010-08-02]http://www.anonymizer.com.
[5] Dingledine R, Mathewson N, Syverson P. Tor: the second-generation onion router[C]// Proceedings of the 13th USENIX Security Symposium. 2004.
[6] JAP Anonymity & Privacy[EB/OL]. [2010-08-02]http://anon.inf.tu-dresden.de/.
[7] I2P Anonymous Network[EB/OL]. [2010-08-02] http://www.i2p2.de.
[8] Adler M, Maggs B. Protocols for asymmetric communication channels[C]// Proceeding of 39th IEEE Symposium on Foundations of Computer Science(FOCS). Palo Alto, CA, 1998.
[9] Gagie T. Dynamic asymmetric communication[J]. Information Processing Letters, 2008, 108(6):352-355.
[10] Feamster N, Balazinska M, Harfst G, et al. Infranet: circumventing Web censorship and surveillance[C]// Proceedings of the 11th USENIX Security Symposium. 2002:247-262.
[11] Xing D S, Shen J Y. A new Markov model for Web access prediction[J]. Computing in Science and Engineering, 2002, 4(6):34-39.
[12] Brian, Davison D. Learning Web request patterns[C]// Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Springer, 2004:435-460.
[13] Deshpande M, Karypis G. Selective Markov model for prediction Web page access[J]. ACM Transaction on Internet Technology, 2004, 4(2):163-184.

A censorship-resistant covert communication protocol based on Web communication behavior

TAN Qing-Feng, SHI Jin-Qiao, GUO Li, WANG Xiao

(*Information Security Research Center, Institute of Computing Technology, Chinese Academy of Sciences; Chinese National Engineering laboratory for Information Security Technologies, Beijing* 100190, *China*)

Abstract The existing covert communication method based on TCP/IP or HTTP often utilizes characteristics of protocol and hides extra data in specific fields of protocol header. Such a method leaves some obvious signatures. However, timing-based covert communications have some kind of traffic pattern. We propose a censorship-resistant covert communication protocol based on Web communication behavior, which combines dynamic asymmetric communication and Markov model-based Web usage prediction. In this paper we focus on covert communication protocol, prototype system, and security of the protocol. The test results show that our method has high performance and safety.

Key words covert communication, censorship-resistant, Web communication behavior