计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20150562
52(10):2373-2381,2015

匿名通信系统不可观测性度量方法

谭庆丰[1,2,3] 时金桥[1,2] 方滨兴[1,2] 郭 莉[1,2] 张文涛[1,2] 王学宾[1,2] 卫冰洁[4]

[1](中国科学院信息工程研究所 北京 100093)

[2](信息内容安全技术国家工程实验室(中国科学院信息工程研究所) 北京 100093)

[3](中国科学院大学 北京 100049)

[4](国家计算机网络应急技术处理协调中心 北京 100029)

(tanqingfeng@iie.ac.cn)

Towards Measuring Unobservability in Anonymous Communication Systems

Tan Qingfeng[1,2,3], Shi Jinqiao[1,2], Fang Binxing[1,2], Guo Li[1,2], Zhang Wentao[1,2], Wang Xuebin[1,2], and Wei Bingjie[4]

[1](*Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093)

[2](*National Engineering Laboratory for Information Security Technologies* (*Institute of Information Engineering, Chinese Academy of Sciences*), *Beijing* 100093)

[3](*University of Chinese Academy of Sciences, Beijing* 100049)

[4](*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing* 100029)

Abstract Anonymous communication technique is one of the main privacy-preserving techniques, which has been widely used to protect Internet users' privacy. However, existing anonymous communication systems are particularly vulnerable to traffic analysis, and researchers have been improving unobservability of systems against Internet censorship and surveillance. However, how to quantify the degree of unobservability is a key challenge in anonymous communication systems. We model anonymous communication systems as an alternating turing machine, and analyze adversaries' threat model. Based on this model, this paper proposes a relative entropy approach that allows to quantify the degree of unobservability for anonymous communication systems. The degree of unobservability is based on the probabilities of the observed flow patterns by attackers. We also apply this approach to measure the pluggable transports of TOR, and show how to calculate it for comparing the level of unobservability of these systems. The experimental results show that it is useful to evaluate the level of unobservability of anonymous communication systems. Finally, we present the conclusion and discuss future work on measuring unobservability in anonymous communication systems.

Key words anonymous communications; relative entropy; unobservablity; privacy protection; traffic analysis

摘 要 匿名通信技术作为一种主要的隐私增强技术被广泛应用于互联网的各个方面,然而传统的匿名通信系统很容易被监视、检测.因此,国内外研究者一直致力于改进匿名通信系统的不可观测属性,以防范网络审查和监视.然而,如何量化评估这些协议的不可观测程度则几乎没有相关的研究.针对匿名通信系统提出一种基于相对熵的不可观测性度量方法,该方法从敌手的威胁模型出发,将匿名通信系统的

收稿日期:2015-06-15;**修回日期**:2015-08-13

基金项目:国家科技支撑计划基金项目(2012BAH37B04);中国科学院战略性先导科技专项课题(XDA06030200)

通信作者:时金桥(shijinqiao@iie.ac.cn)

输入、输出状态映射到一个交互式图灵机,并在此基础之上提出一个基于相对熵的不可观测性度量框架,该框架能够有效地度量匿名通信系统的不可观测程度.此外,将它应用于 TOR 匿名通信系统的传输层插件的度量,实验结果表明,该方法能够有效地度量匿名通信系统的不可观测性.

关键词 匿名通信;相对熵;不可观测性;隐私保护;流分析

中图法分类号 TP393.08

随着互联网技术的发展,网络已经深入到人们生活的各个方面,如电子投票、电子支付、电子邮件甚至网页浏览等.匿名通信技术作为一种主要的隐私增强技术被广泛应用于互联网的各个方面,现有的匿名通信系统是利用 Mix 网络和洋葱路由技术隐藏通信过程中通信主体的身份信息以及通信双方的通信关系,从而为在线用户提供隐私保护.然而,随着网络流量分析技术的发展,攻击者不仅可以检测到匿名通信系统的流量指纹特征,还能够更进一步监视匿名网络,从而破解其匿名性、甚至可以阻断互联网用户跟匿名通信系统的连接,如"棱镜"计划显示美国政府部门(位于 Utah 的 NSA 情报中心)已经针对 TOR(the onion router)等匿名网络开展大规模的信息收集工作[1].因此,匿名通信技术虽然能够保证通信主体的不可追踪性和通信客体的私密性,但是,不能掩盖通信主体正在使用该技术这一事实,即对于敌手来说可以检测到该用户正在使用匿名通信系统进行通信.

在这种背景下,国内外研究者开始在匿名通信系统的基础之上研究通信协议的不可观测性,即通过协议伪装、流量模糊化、接入点隐藏和区分发布等方式消除匿名通信系统的通信行为和流量特征,以抵御深度流分析、扫描攻击、女巫攻击等,从而增强系统的匿名性和不可检测性.

已经有很多文献[2-5]提出了匿名支持隐蔽的通信方法,文献[2-4]提出协议伪装和流量模糊化方法,协议伪装主要思想是模仿或者伪装流行的掩体协议,以逃避网络审查,类似于协议层面的"傍大款".文献[5]通过运行真实的目标协议,并将目标协议承载到流行的掩体协议(如 VOIP,P2P,UGC 等)之上,以实现通信行为的不可观测性.据网上公开的文献调研,现有的匿名通信系统的不可观测属性度量和评估目前还没有相关的研究工作.因此,如何度量和评估这些协议的不可观测性对于指导未来系统的设计、评价通信协议的不可观测程度都是一件非常重要的工作.

本文将从敌手的威胁模型出发,将通信系统的输入输出状态映射到一个交互式图灵机,并在此之上提出一个基于相对熵的不可观测性度量框架.我们的方法可以量化评估协议的不可观测程度,如协议的可观测方面在多大程度上可以被攻击者检测出.

1 相关工作

在过去的 30 多年里,自从 Chaum[6] 于 1981 年首先提出 Mix(消息混合)技术和匿名通信的概念以来,Internet 上的匿名通信技术研究越来越吸引人们的注意,无论是学术界还是一般公众.与此同时,研究机构也一直致力于设计、评估、分析和改进匿名通信系统.目前,实用的匿名通信系统主要有 TOR[7],JAP[8]等,并在实际的网络环境中部署、运行.但是,匿名通信系统在设计之初并没有很好地考虑到抗审查的应用需求,因此匿名通信协议本身很容易被检测,从而攻击者可以监视互联网用户正在使用匿名通信工具,甚至可以阻断用户跟匿名网络的连接.

Pfitzmann 等人[9]在匿名与隐私相关术语定义的文献中提出不可观测性,并给出了定义.在现有的各种文献中,通常将互联网上不可观测的通信系统解释为满足匿名和不可检测的通信系统.此后,国内外的研究者提出了各种不可观测的通信系统[2-5].Houmansadr 等人在文献[10]中对协议伪装和模仿给出了详细的评测和分析,认为模仿和伪装是非常困难的,需要正确实现协议的具体规范以及协议内部和协议之间的依赖关系.Geddes 等人在文献[11]中指出,即便在流行的协议基础之上嵌入隐蔽隧道如 Freeware 也会引起各种问题(如架构不匹配、信道不匹配、传输内容不匹配等).但是这些论文都没有对不可观测程度给出可比较的量化评估方法.目前匿名通信系统不可观测性的度量却尚未引起足够的重视,几乎没有相关的研究工作.而匿名通信系统的匿名性度量相关研究工作却不少,下面将阐述匿名通信系统匿名性度量相关的一些研究工作.

Reiter 等人[12]提出用 1-p 表示匿名度,其中 p 是攻击者能够从匿名集中识别出某一特别用户的概

率,它可用于衡量同一匿名集合中不同对象之间的可识别的程度. Berthold 等人在文献[13]中提出了基于匿名集大小的匿名性度量方法,即定义匿名度 $d=\mathrm{lb}\,N$,其中 N 为系统匿名集合中对象的个数;然而在这一定义下,匿名度的度量只考虑了系统的用户数,而没有考虑攻击者通过一段时间的攻击后能够区分出的用户数,因此这种方法无法刻画攻击者对于匿名通信系统攻击后的匿名性变化. Diaz 等人[14]提出基于香农熵的匿名性评估框架. 该方法的优点是不仅仅考虑匿名集大小,还考虑了攻击者对于匿名集合中不同成员之间的可识别状态的概率分布情况,因此能够更好地度量攻击者在获取到相关信息之后,匿名通信系统不同成员的匿名程度.

Hamel 等人[15]认为基于熵的匿名性度量存在诸多缺陷,从而关注于研究敌手的攻击行为如何影响匿名性. 将敌手的攻击能力等价于能够控制匿名网络带宽资源的多少,然后给每一个敌手赋予一定的带宽预算,探讨在一定的带宽资源下对匿名性攻击的影响,从而评估匿名通信系统的匿名度.

2 问题陈述和定义

2.1 问题陈述

越来越多的用户采用匿名通信技术逃避网络审查,匿名通信技术能够保护通信双方的通信内容安全性以及通信主体的身份信息,但是不能够隐藏用户正在使用该技术的事实,也就是说,攻击者可以很容易识别出某一用户正在使用匿名通信工具. 此外,现代的匿名通信技术同样也不能够消除网络数据包的外显行为特征,如数据包的大小分布、网络时延、数据包的内部间隔时间等,这些看似无用的协议指纹特征,实际上对于具有流分析攻击能力的敌手来说可以有效地识别网络用户的身份信息.

针对流分析攻击方法,比较常见的对抗手段是协议伪装,即伪装和模仿某一流行协议的具体实现,用以混淆协议的外显行为特征,如 SkypeMorph[2]将 TOR 的流量伪装成 Skype 视频流量,StegoTorus[3]将 TOR 的流量伪装成 Skype 和 HTTP 的流量,CensorSpoofer[4]模仿基于 SIP 的 VOIP 协议. 实际上 TOR 为了对抗流分析,先后部署了 obfs[16],fte[17],Meek[18]等传输层插件以支持协议混淆和协议伪装. 因此度量这些协议的不可观测程度,对于评估协议的隐蔽性、指导未来协议的设计与实现都具有十分重要的意义.

低时延匿名通信系统的可用性问题本质上依赖于其通信协议的不可观测程度,本文关注匿名通信系统不可观测属性的度量和评估,不可观测性从用户的通信行为来说是一个计算上的谜题,即攻击者在统计上不能够确认或者区分某一用户是否在使用某一特别的协议. 因此,如何去度量匿名通信协议的不可观测程度是一项非常具有挑战性工作. 本文将形式化定义匿名通信系统的模型,研究在该模型下匿名通信协议与所伪装的协议在通信行为上是否具有统计意义上的可区分性,即计算通信过程中数据包的外显行为特征的相对熵值,以度量其可观测程度.

2.2 相关概念定义

在本节,我们将给出匿名通信系统中不可观测性的准确定义,本文采用 Pfitzmann 等人在文献[9]中给出的定义.

定义 1. 匿名性(anonymity). 匿名性是指通信主体在一组匿名集合中不可识别的状态.

如果攻击者能够从获取到的信息中关联到匿名集合中的某一个发送者,如该发送者发送消息的 IP 地址,则认为该用户是可识别的.

定义 2. 不可检测性(undetectability). 从攻击者的角度,攻击者不能够区分感兴趣的通信客体(item of interest)是否存在.

因此,匿名性是研究通信主体的通信关系,保护的是通信主体的身份信息,而不可检测性研究对象为感兴趣的通信客体,保护的是通信行为和通信客体.

定义 3. 完美不可检测性(perfect undetectability). 当感兴趣的通信客体存在与否是完全不可区分的,则认为该通信客体具有完美不可检测性.

定义 4. 不可观测性(unobservability). 不可观测性是指感兴趣的通信客体在任何其他相同类型的通信客体集合中不可区分的状态,不可观测性包括 2 个方面的含义:通信主体的匿名性和通信客体的不可检测性.

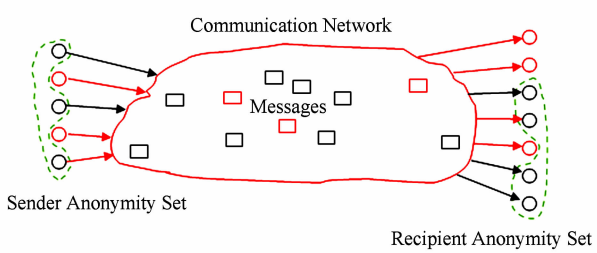


Fig. 1 Network graph for anonymous communications.

图 1 匿名通信网络图

本文关注匿名通信系统不观测属性的度量和评估.在此,我们仅考虑发送者不可观测性,即对于一个特别的消息,攻击者想在发送者匿名集中找出通信过程中该消息是否出现、来自于哪个发送者.在此发送者定义为所有可能的用户集合.

本文定义的不可观测程度是基于某一特定消息出现的概率,即攻击者监视匿名通信系统一段时间后,攻击者给每一个通信主体在通信过程中出现的消息的外显行为特征赋予一个概率值,用以表征该消息可能出现程度.

3 系统建模

匿名通信系统建模:定义匿名通信系统为一个有向图 $G=\langle V,E\rangle$.其中,V 为匿名通信系统的节点,如客户端、服务端、中继节点;E 为路由路径.

定义 E 为匿名通信系统的路由路径的集合,记为 $E=Path(G)$.

V 为 G 上的顶点集合,并定义 V 为一个概率多项式程序,建模为一个五元组 $V=(S,I,O,f,g)$,其中:

S 为客户端、服务器或者中继节点的状态集合,任何一个客户端、服务器或者路由节点都处于 $n=|S|$ 个状态中的一个,记为 $s\in S$;

I 是客户端、服务器或者路由节点可能的输入集合,客户端、服务器或者路由节点的下一个状态通过输入和状态转移函数 f 共同确定;

O 为每一个状态的可能的输出集合,我们可以将这个输出叫作数据包序列,记为 $O=\{\{o_{n,i}\}_{i=1}^{M}\}_{n=1}^{N}$,其中 N 代表包的个数,M 代表每个包上可以观测到的外显行为特征数量,则 $o_{n,i}$ 为第 n 个包上的第 i 维特征值;

f 为状态转移函数:$f:S\times I\rightarrow S$;

g 为输出函数:$g:S\times I\rightarrow O$.

3.1 威胁模型

系统的可观测程度依赖于敌手的攻击能力、拥有的资源以及通信协议本身的外显行为特征的可观测性.本论文将敌手建模为一个被动、局部的敌手:即敌手能够在网络边界监视用户的流量,从而观测通信协议的外显行为特征(如包的大小、包的内部时间间隔等),但是不能够篡改、注入、丢弃数据包.此外,我们也假设敌手只能够在网络边界部署流量分析设备,而不能够控制用户的计算机,该敌手具有有限的计算和存储资源,而不能够在大规模骨干网络环境下完成复杂的计算.本节我们将从攻击能力、可见性、敌手拥有的计算资源这 3 个方面定义敌手的攻击能力:

1) 攻击能力

① 被动攻击.敌手能够在网络边界部署流分析系统,以监视、分析网络流以及通信主体的通信行为.

② 主动攻击.敌手能够操纵网络流(如加时延、丢包、修改包内容、注入攻击等).

③ Proactive 攻击.敌手能够模拟通信系统的客户端发包,以主动探测网络状态、节点行为特征等.

2) 可见性

① 局部可见.敌手仅仅能够监视网络边界(进入和离开某一网络)或者部分网络节点.

② 全局可见.敌手能够监视整个通信网络节点以及网络流量.

3) 资源

① 有限.敌手具有有限的计算和存储资源.

② 充分.敌手具有充分的计算和存储资源,即能够从不同的网络位置同时收集捕获到的大规模网络流量进行存储和处理.

本文假设真实世界的敌手为局部、被动、具有有限的计算和存储资源的敌手.

4 度量方法

4.1 相对熵(relative entropy)

在信息论中,信息熵定义为离散随机事件出现的概率,通常用信息熵来度量信息不确定程度.假设离散随机变量 X,它有 N 个可能的取值,即 $X\in\{x_1,x_2,\cdots,x_N\}$,各个取值发生的概率分别为 $p_i=Pr(X=x_i)$,x_i 为离散随机变量 X 的取值.则离散随机变量 X 的熵定义为

$$H(X)=-\sum_{i=1}^{N}p_i\mathrm{lb}(p_i). \tag{1}$$

相对熵又称为 KL 散度(Kull-back-Leibler divergence, KLD),是 2 个概率分布 P 和 Q 差别的非对称性的度量,典型情况下,P 表示数据的真实分布,Q 表示数据的理论分布、模型分布或 P 的近似分布.假设 $p(x)$ 和 $q(x)$ 是随机变量 X 的 2 种概率密度函数,则两者之间的相对熵定义为

$$D[p(x),q(x)]=\sum_{x\in X}p(x)\ln\frac{p(x)}{q(x)}, \tag{2}$$

当且仅当 $p(x)=q(x)$,$D[p(x),q(x)]=0$.

4.2 不可观测程度度量方法

根据第3节的定义,如果匿名通信协议 π 为目标协议 τ 的一个具体实现,对于给定的集合 $O=\{\{o_{n,i}\}_{i=1}^{M}\}_{n=1}^{N}$,则每一种通信协议可以建模为数据包在外显行为特征上的概率分布.假设 p_π 为匿名通信系统在第 i 维特征向量 $\boldsymbol{O}_i$ 上的概率分布,同样地,p_τ 为目标协议在 $\boldsymbol{O}_i$ 上的概率分布,其中 O 为攻击者攻击后所观测到的数据包的外显行为特征集合,则匿名通信协议 π 和目标协议 τ 在状态 s 下的相对熵定义为

$$D_s[p_\pi,q_\tau]=\sum_{o_{n,i}\in\boldsymbol{O}_i}p_\pi(o_{n,i}\mid s)\ln\frac{p_\pi(o_{n,i}\mid s)}{q_\tau(o_{n,i}\mid s)},\tag{3}$$

其中,s 为匿名通信节点的状态,如匿名通信节点处于握手状态、空闲状态等.

因此,基于相对熵,我们可以度量匿名通信系统的通信行为跟所伪装的协议在统计上的不可观测程度,当攻击者看到匿名通信协议在 O 上的外显行为特征的概率分布和目标协议在 O 上的概率分布完全一致时,他们之间的相对熵最小,即匿名通信协议的不可观测程度最小.

定义5. 假设 S 为匿名通信系统节点的状态集合,定义 $\hat{D}$ 为匿名通信系统的相对熵,记为

$$\hat{D}=\frac{1}{|S|}\sum_{i=1}^{|S|}(D_s[p_\pi,q_\tau]-D_s[p_\tau,q_\tau]).\tag{4}$$

定义6. 匿名通信系统的最大相对熵定义为

$$D_M=\max_{s\in S}(D_s[p_\pi,q_\tau]).\tag{5}$$

定义7. 匿名通信系统不可观测程度定义为

$$d=1-\frac{D_M-\hat{D}}{D_M}.\tag{6}$$

因此,匿名通信系统的不可观测程度可以量化度量匿名通信系统在通信行为上跟其他协议的可区分程度.d 的取值范围为[0,1],如果 d 越小则表明协议的不可观测属性越好,当 $d=0$ 时表明协议完全不可区分.

5 度量TOR传输层插件的不可观测程度

TOR为当今世界上使用最为广泛的低时延匿名通信系统,当前,TOR匿名通信系统大约有6 000多个路由节点、4 000个非公开的网桥节点,每天同时在线用户大约有200多万[19].然而TOR协议本身并不具有很强的不可观测性,因此很难抵御流量分析攻击.东南大学何高峰等人在文献[20]中提出基于TLS指纹的识别方法和基于报文长度分布的TOR匿名通信系统流量识别方法,可以有效识别出TOR的流量.为了抵御流分析攻击,国内外的研究者先后提出了各种协议伪装方法,TOR项目组也先后开发了各种传输层插件以增强协议的不可观测性.本文以TOR的桥接方式以及最近发布的Meek插件为例,给出其不可观测程度.

5.1 网桥机制介绍

为了抵御网络审查和流分析攻击,TOR提出了非公开的网桥(bridge)机制,获取网桥节点只能通过Gmail邮箱向TOR的邮件服务器申请或者到TOR官方网站上去获取(https://bridges.torproject.org).为了抵御枚举攻击,TOR限制每个邮箱地址在一定的时间间隔内只能获得一定数量的网桥节点.对于网页方式,TOR同样限制每个IP地址在一定的时间间隔内只能获得一定数量的网桥节点.此外,为了抵御流分析攻击,TOR的网桥通信协议模仿了Firefox浏览器的TLS握手过程以混淆其流量特征.

5.2 Meek机制介绍

Meek为TOR项目组最近发布的一个传输层插件,其目标是将TOR流量伪装成访问云平台的流量.当TOR的数据流量到达云计算平台之后,会经过TOR的Meek-Server转发到TOR匿名网络,最终转向真正的目标地址.跟以前其他传输层插件不同的是它不是模仿某一协议,而是直接运行Firefox目标协议.即将TOR的流量封装到Firefox的HTTPS载荷中,而HTTPS协议头则为Firefox跟云平台通信的协议头,在到达云平台的服务器后,再解析出HTTPS内容,并经Meek-Server转发给TOR匿名网络.因此Meek传输层插件具有以下2个方面的优点:

1) 隐蔽性强.Meek传输层插件不是伪装而是直接运行目标协议,因此其协议特征跟目标协议特征具有很强的相似性.

2) 不需要像网桥机制那样需要直接发布接入点地址,而是利用云平台的前端域名机制,让用户请求云平台的域名,然后解析得到其IP地址,其过程跟用户访问Google的搜索服务以及访问Amazon,Microsoft云平台的过程完全一致.

5.3 实验方法

为了度量TOR匿名通信系统传输层插件的不可观测程度,我们在TOR官方网站分别下载了Windows平台和Linux平台的最新版本TOR客户

端软件(torbrowser-install-4.5.1_zh-CN),其内置的 Firefox 版本为 Firefox ESR 31.7.0 版本.

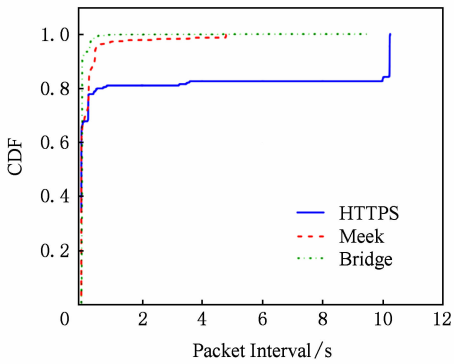
由于 Meek 和网桥方式都是伪装 HTTPS 协议,因此,本实验的方法是分别抓取 TOR 的 Meek-Azure 模式和网桥模式下访问某个 Web 站点的流量,然后用 Firefox 流览器直接访问相同的网址,并且在每一种模式下分别在 Windows 7 和 Ubuntu 下在重复采集 3 次流量,每一种模式采集的流量都包括 HTTPS 协议握手阶段、空闲阶段(等待大约 3 min)、用户输入数据阶段(都访问相同的网址).此外,为了计算 Meek 和网桥的协议与其伪装的 Firefox(版本为 Firefox ESR 31.7.0)的 HTTPS 协议的不可观测程度,本次实验同样采集了用户直接用 Firefox 浏览器访问 Meek-Azure 平台的网址.采集的流量同样包括 HTTPS 协议握手阶段、空闲阶段(等待大约 3 min),用户输入数据阶段(访问相同的 Web 站点).

为了度量 TOR 传输层插件的不可观测程度,我们选择了包的大小分布和数据包的内部到达时间间隔作为网络通信行为可观测性的外显行为特征.为了直观地了解本文选择的特征是否具有可区分性,首先,我们分别计算了它们在不同平台下(Windows,Ubuntu)客户端到服务器端(C2S)和服务器端到客户端(S2C)的累积分布函数(CDF),以及用 Firefox 直接访问 HTTPS 站点(Azure 云平台)的 CDF 值,如图 2 和图 3 所示;然后,我们根据本文 4.2 节提出的方法分别计算在不同平台和状态下的相对熵,如表 1 和表 2 所示;最后根据 4.2 节提出的匿名通信系统不可观测程度计算公式分别给出 TOR 的 Meek 模式、Bridge 模式的不可观测程度的度量值.为了展示效果更加直观,我们对数据作了预处理,过滤了所有 TCP 三次握手的数据包.对于包的内部时间间隔我们精确到 10 ms 级,此外对于 $P(x)=0$ 的情况,我们做了技术上的处理,即将它赋予一个很小的值,在本文将此情况赋予一亿分之一的概率值.

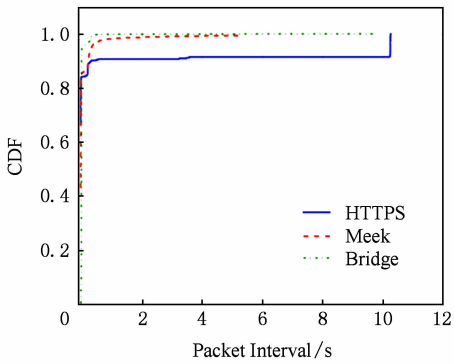
5.4 实验结果与分析

本节我们将讨论实验结果,并对实验结果进行相应的分析.

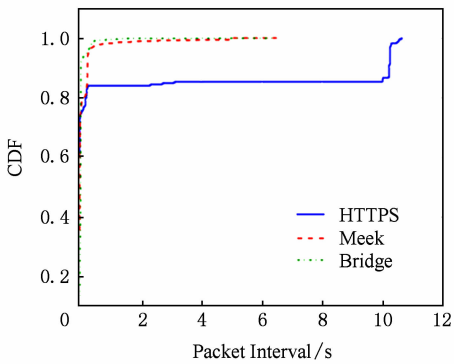
图 2 给出了 Meek,Bridge,HTTPS(Firefox 的 HTTPS)直连的数据包的内部时间间隔分布.从图 2 我们看出 Meek,Bridge,HTTPS 有大约 70%~80%的数据包内部时间间隔分布几乎没有什么差别,但是 Meek 的数据包在 0.5~2 s 附近有一个明显的抖动,而 HTTPS 则有大约 20%的数据包的内部时间间隔在 10 s 附近,经过人工分析 TOR 的源码发现 Meek 的客户端跟云平台之间为了保持长连



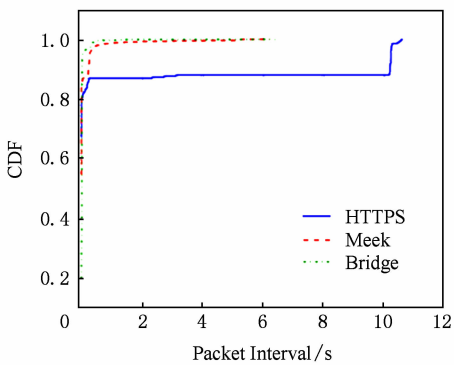
(a) Packet interval distribution (Windows, C2S)



(b) Packet interval distribution (Windows, S2C)



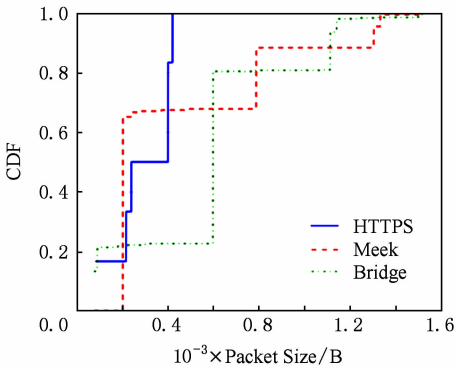
(c) Packet interval distribution (Ubuntu, C2S)



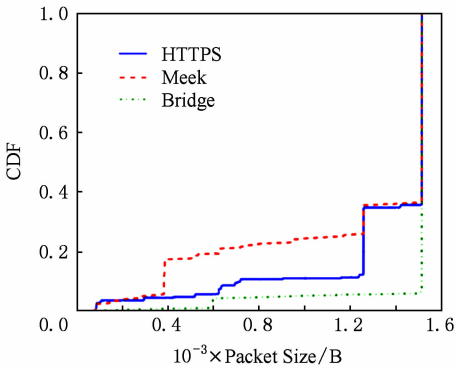
(d) Packet interval distribution (Ubuntu, S2C)

Fig. 2 Packet interval distribution.

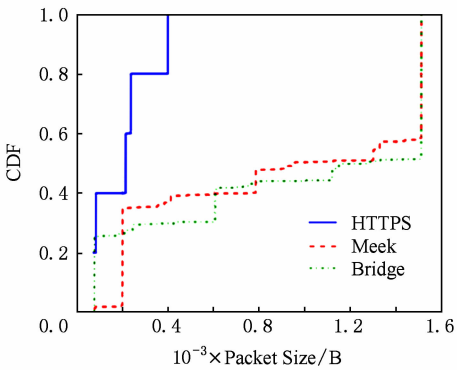
图 2 数据包的内部时间间隔分布



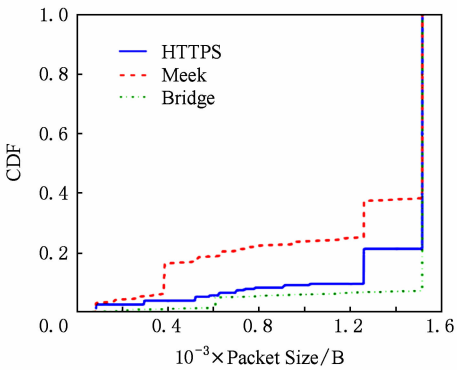
(a) Packet Size distribution(Windows, C2S)



(b) Packet Size distribution(Windows, S2C)



(c) Packet Size distribution(Ubuntu, C2S)



(d) Packet Size distribution(Ubuntu, S2C)

Fig. 3 Packet size distribution.

图 3 数据包大小分布

接,在连接空闲状态时会自动发送一个心跳包,心跳包的时间间隔为 0.1~5 s 之间的一个随机值,而用 Firefox 直接访问 Azure 云平台的时候,如果没有数据,也会发送一个心跳包,心跳包的时间间隔为 10 s.

图 3 为 Meek, Bridge, HTTPS(Firefox 的 HTTPS)的数据包的大小分布图,由图 3 我们可以看到如下 2 个明显的现象:

1) 现象 1. Bridge 模式下客户端到服务器端的数据包大小在 600 B 附近比较集中. 经过分析后发现,TOR 数据包(cell)为一个固定大小值,其在 Windows 平台下为 597 B,Ubuntu 平台下为 609 B.

2) 现象 2. Meek 模式下客户端到服务器端数据包大小为 200 B 左右、服务器到客户端数据包大小为 400 B 左右占比很高,对 TOR 的 Meek 源码分析后发现为 TOR 的 Meek 模式心跳包大小,不同的平台下心跳包的大小稍微有些不同,同一平台则没有区别.

Table 1 Relative Entropy for Packet Interval Distribution

表 1 包内部时间间隔分布的相对熵 nat

Pattern	OS	Handshake		Idle		Data Transmission	
		C2S	S2C	C2S	S2C	C2S	S2C
Meek	Windows	3.56	1.20	18.27	18.14	10.35	5.19
	Ubuntu	2.78	1.63	18.94	18.02	4.29	3.17
Bridge	Windows	18.15	18.47	21.82	21.82	15.64	19.20
	Ubuntu	13.07	13.14	21.85	21.82	14.59	13.76
HTTPS	Windows	1.67	2.05	2.56	4.64	1.96	0.77
	Ubuntu	3.57	2.05	17.03	16.86	1.51	1.54

Table 2 Relative Entropy for Packet Size Distribution

表 2 包大小分布的相对熵 nat

Pattern	OS	Handshake		Idle		Data Transmission	
		C2S	S2C	C2S	S2C	C2S	S2C
Meek	Windows	17.57	4.76	20.91	20.20	19.89	10.86
	Ubuntu	18.14	5.48	10.25	15.01	18.48	5.72
Bridge	Windows	18.00	0.77	21.92	21.92	20.19	4.88
	Ubuntu	18.69	0.73	21.95	21.92	18.44	4.30
HTTPS	Windows	0	0	0	0	0	0.77
	Ubuntu	0	0	0	0	0	1.45

由图 2 和图 3 可以直观地看出,TOR 的传输层插件(Bridge, Meek)在统计上跟所伪装的协议(HTTPS)仍然具有较为明显的差异. 下面将分别计算 TOR 的 Meek 模式、Bridge 模式跟 HTTPS 协议在不同的状态下,包的内部时间间隔分布和包的大小分布的相对熵值以及 HTTPS 跟自己相比较(在同一平台、不同时间下抓取的数据包)的相对熵值,

如计算握手状态下客户端到服务器端包的大小分布和包的内部时间间隔的相对熵.

表 1 和表 2 的结果表明,相对熵能够很好地刻画目标协议与掩体协议之间在统计上的相似程度;包的大小分布相对于内部时间间隔能够更好地表征 TOR 传输层插件的外显行为特征的可区分程度.本文中计算相对熵时,采用的是同一网络环境和操作系统、在同一状态下计算包的大小分布和内部时间间隔分布的相对熵值,因此不同计算结果具有可比性.

因此根据式(6)即可以计算出 Meek 以及 Bridge 传输层插件的不可观测程度.对于 Meek 插件其在 Windows 平台下的不可观测程度 $d=0.54$;在 Ubuntu 平台下的不可观测程度为 $d=0.36$.对于 Bridge 插件,在 Windows 平台下的不可观测程度 $d=0.72$;在 Ubuntu 平台下的不可观测程度为 $d=0.53$.从上面的值可以看出,Meek 传输层插件无论是在 Windows 平台还是在 Ubuntu 平台下,其通信行为上的不可观测程度都好于 Bridge 模式,即在统计上更难区分 Meek 模式的流量;但是其跟 HTTPS 的流量相比较,还是具有较强的流量特征,即很容易从 HTTPS 流量中筛选出 TOR 的 Bridge 流量和 Meek 模式的流量.

6 结论和未来工作

本文形式化地定义了匿名通信系统的通信模型,并在此模型的基础之上提出一种基于相对熵的匿名通信系统不可观测性度量方法.该方法能够有效地度量匿名通信系统的不可观测程度,并将该方法应用于 TOR 的传输层插件的不可观测性度量,本方法同样适用于其他匿名通信系统不可观测性的度量.尽管国内外学者和 TOR 项目组一直致力于改进 TOR 通信行为的安全性,但是本文从理论分析和实验结果都表明,TOR 的传输层插件还没有实现 TOR 官方所声称的不可观测性.目前本论文提出的匿名通信系统的不可观测属性的度量分析只考虑在某一静态时间点通过信息熵的度量方法来评估匿名通信系统的不可观测属性.信息熵对于匿名网络本身的安全评估提供了很好的分析方法,但对任意给定用户,通常需要准确评估其多长时间以多大的概率可以有效识别匿名通信系统的协议,从而进一步破解其匿名性,而这些工作对于度量和评估匿名通信系统都具有十分重要的意义.

参 考 文 献

[1] Bamford J. The NSA is building the country's biggest spy center (watch what you say)[OL]. [2015-05-23]. http://www.wired.com/2012/03/ff_nsadatacenter

[2] Moghaddam H, Li B, Derakhshani M, et al. SkypeMorph: Protocol obfuscation for tor bridges [C] //Proc of the 19th ACM Conf on Computer and Communications Security. New York: ACM, 2012: 97-108

[3] Weinberg Z, Wang J, Yegneswaran V, et al. StegoTorus: A camouflage proxy for the Tor anonymity system [C] //Proc of the 19th ACM Conf on Computer and Communications Security. New York: ACM, 2012: 109-120

[4] Wang Q, Gong X, Nguyen G T K, et al. CensorSpoofer: Asymmetric communication using IP spoofing for censorship-resistant Web browsing [C] //Proc of the 19th ACM Conf on Computer and Communications Security. New York: ACM, 2012: 121-132

[5] Houmansadr A, Riedl T J, Borisov N, et al. I want my voice to be heard: IP over Voice-over-IP for unobservable censorship circumvention [C] //Proc of the 20th Annual Network and Distributed System Security Symp. Reston, VA: The Internet Society, 2013: 1-17

[6] Chaum D L. Untraceable electronic mail, return addresses, and digital pseudonyms [J]. Communications of the ACM, 1981, 24(2): 84-90

[7] Dingledine R, Mathewson N, Syverson P. Tor: The second-generation onion router [C] // Proc of the 13th Conf on USENIX Security Symp. Berkeley, CA: USENIX Association, 2004: 303-320

[8] Department of Business Informatics, University of Regensburg. JAP: The JAP anonymity & privacy homepage [OL]. [2015-05-23]. http://anon.inf.tu-dresden.de

[9] Pfitzmann A, Hansen M. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management [OL]. [2015-05-23]. http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf

[10] Houmansadr A, Brubaker C, Shmatikov V. The parrot is dead: Observing unobservable network communications [C] //Proc of the 34th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2013: 65-79

[11] Geddes J, Schuchard M, Hopper N. Cover your ACKs: Pitfalls of covert channel censorship circumvention [C]//Proc of the 20th ACM SIGSAC Conf on Computer & Communications Security. New York: ACM, 2013: 361-372

[12] Reiter M K, Rubin A D. Crowds: Anonymity for Web transactions [J]. ACM Trans on Information and System Security (TISSEC), 1998, 1(1): 66-92

[13] Berthold O, Pfitzmann A, Standtke R. The disadvantages of free MIX routes and how to overcome them [G] //Designing Privacy Enhancing Technologies. Berlin: Springer, 2001: 30-45

[14] Diaz C, Seys S, Claessens J, et al. Towards measuring anonymity [C] //Proc of the 2nd Int Conf on Privacy Enhancing Technologies. Berlin: Springer, 2003: 54-68

[15] Hamel A, Gregoire J, Goldberg I. The misentropists: New approaches to measures in TOR, CACR 2015-18 [R]. Waterloo, Ontario, Canada: University of Waterloo, 2011

[16] Dingledine R. Obfsproxy: The next step in the censorship arms race [OL]. [2015-05-23]. https://blog.torproject.org/blog/obfsproxy-next-step-censorship-arms-race.

[17] Tor Project. Tor Meek [OL]. [2015-05-23]. https://trac.torproject.org/projects/tor/wiki/doc/meek

[18] Dyer K P, Coull S E, Ristenpart T, et al. Protocol misidentification made easy with format-transforming encryption [C] //Proc of the 20th ACM SIGSAC Conf on Computer & Communications Security. New York: ACM, 2013: 61-72

[19] Tor Project. Tor Metrics [OL]. [2015-05-23]. https://metrics.torproject.org

[20] He Gaofeng, Yang Ming, Luo Junzhou, et al. Online identification of Tor anonymous communication traffic [J]. Jornal of Software, 2013, 24(3): 540-556 (in Chinese)
(何高峰, 杨明, 罗军舟, 等. Tor匿名通信流量在线识别方法[J]. 软件学报, 2013, 24(3): 540-556)

Tan Qingfeng, born in 1981. PhD candidate, assistant professor, Member of China Computer Federation. His main research interests include anonymous communication and covert communications.

Shi Jinqiao, born in 1978. PhD, professor, PhD supervisor. Member of China Computer Federation. His main research interests include network and information security, especially privacy enhancing techniques, and data leakage detection.

Fang Binxing, born in 1960. Professor, and PhD supervisor. Academician of Chinese Academy of Engineering. His current research interests include computer architecture, computer network and information security.

Guo Li, born in 1969. Professor, and PhD supervisor. Member of China Computer Federation. Her research interests include network information security.

Zhang Wentao, born in 1989. Master. Member of China Computer Federation. His main research interests include anonymous communications and information retrieving.

Wang Xuebin, born in 1986. Master. Member of China Computer Federation. His current research interests include visual analytic, computer network and information security.

Wei Bingjie, born in 1987. Recevied her PhD degree in compute software and theory from Institute of Computing Technology, Chinese Academy of Sciences. Her main research interests include microblog retrieval.