IEEE/ACM Transactions on Networking

SPABox：在MiddleBox进行深度数据包检查期间保护隐私

范靖远，关朝文，任奎，*IEEE研究员*，崔勇，*乔春明，IEEE研究员*

摘要-Widely用于在互联网上的流量进行加密 · HTTPS 提供安全 和 私有 数据 通信 客户端和之间 的服务器。

但是 · 为了应对快速变化和复杂的安全攻击 · 网络运营商经常使用从简单的关键字匹配到更高级的机器学习 和数据挖掘分析等技术 · 部署中间盒 来执行深度包检测 （DPI） · 以 检测攻击和潜在的安全漏洞。但是 · 这带来一个问题：如何能中间件 · 其中雇用DPI · 工作在HTTPS使用加密的交通连接 · 同时保持隐私？

在本文中 · 我们介绍了SPABox · 这是一个基于中间盒的系统 · 该系统在加密流量上支持基于关键字和基于数据 分析的DPI功能。SPABox通过使用新颖的协议以有限的连接设置开销来保护隐私。我们在标准服务器上实现SPABox · 并 表明SPABox对于长期连接和短期连接都是实用的。与国家的最先进的Blindbox系统相比 · SPABox是大小更快的多于5个数量和需要较少数量级带宽的7个订单连接设置而 SPABox可以达到更高的安全 级别。

*索引*词-DPI · 隐私保护 · 中间盒。

I. 我导论

HTTPS是一种流行的因特网协议，它使用反式-端口层安全（TLS） 或者 它的 前身， 安全套接字 层 （SSL） [27] 到 加密 通信之间客户端和服务器以确保数据完整性和隐私性。

传统上，提供深许多中间件包检测（DPI）功能由网络部署OP- erators通过搜索特定检测攻击关键字 在非加密的流量[23]或签名。由于恶意软件使用各种隐藏技术，如混淆，和聚Morphic环或变质的策略来逃避检测[15]，[16]，工业界和学术界都认为加入更先进的机器学习 和 数据 挖掘分析 的DPI [50]，[8]，[12]。例如，赛门铁克和Proofpoint都声称通过机器学习，他们能够 到检测除系统更攻击和威胁而不机器学习 技术 [10]， [9]。 然而，与 在成长

2002年1月20日收到手稿；修改后的1月30日，2002年这一工作的一部分是由中国国家自然科学基金的支持下，格兰特61422206和部分由下格兰特CSR-1409809的美国国家科学基金会。i的研究得到了美国国家科学基金会的部分资助，资助号为CNS-1262277。

J.风扇，C.关，K.仁和C.巧与该 部 的 国家计算机科学与工程，布法罗大学，大学新的纽约，布法罗，NY 14260 USA（电子邮件：jfan5 @ buffalo.edu; chaoweng@buffalo.edu; kuiren@buffalo.edu; qiao@computer.org）。

Y.崔是与该系的计算机科学和Tech-清华大学nology，北京，100084（电子邮件：cuiy- ong@tsinghua.edu.cn）。

数字对象标识符：10.1109 / TNET.2017.2753044

由于采用HTTPS，现有方法无法执行关键字或签名匹配，更不用说用于加密流量的恶意软件检测的高级机器学习分析了。

Naylor等。[39]试图通过提出mcTLS解决此问题，该协议允许与中间盒服务共享加密密钥。然而，这种做法不能保护用户的私人 信息 从 一个 服务 提供商 谁部署中间件。实际上，服务提供商一直 在使用其部署的中间盒来捕获Internet流量以进行监视和营销[6]，这一做法已受到广泛批评。雪利酒等。[43]采取了第一步，通过引入Blindbox来弥合差距， 支持对加密流量进行关键字匹配和正则表达式评估。他们通过开发一种新颖的可搜索加密方案实现了这一目标，并通过实验证明了其可行性。然而，其功能范围由限定的底层加密方法。具体而言，Blindbox仍需要在中间盒解密流量，以执行正则表达式评估，从而可以揭示最终用户认为是私有的信息。此外，它无法支持用于恶意软件检测的复杂 机器 学习 分析。毛皮- 此外，连接设置的开销使Blindbox不切实际。在后续工作[35]试图解决的 高设置 的开销通过 采用 一个 修改的 架构。

在本文中，我们介绍了SPABox，这是第一个 基于中间盒的系统，该系统通过有限的开销在加密流量上支持基于签名和数据分析的DPI功能。与[35]相比，SPABox在更一般的条件下实现了更大的设置开销减少，而[35]仅适用于企业环境。使用SPABox，当两个端点创建通信链接时，中间盒无需参与，与其他现有中间盒部署一样，中间盒也保持 了透明性。SPABox通过引入新颖而简单的方法来实现这些目标， 高效 加密 策略 建造 上 顶部的 所述*离散 对数问题*[36]和用于一种新颖的协议的安全两方之间的通信。虽然我们的协议具有加密关键字（和部分的映射关键字），以他们的纯文本表示，SPABox既不能解密流量，也没有任何推断私人信息，这是比强大的隐私模式中的以前的工作[43]。

具体来说，SPABox支持三种类型的操作：关键字匹配（第IV-B节），正则表达式评估（第IV-C节）和通过机器学习进行恶意软件检测（第IV-D节）。这三个操作背后的设计主要思想是：1）代替 直接匹配关键字，我们对关键字进行标记化，以便我们可以构建一个

类似于Trie的结构，可加快DPI中间盒的搜索和匹配速度； 2）在接收方处理正则表达式，乱码和遗忘传输用于保护流量和规则集的隐私，以及3）加法和标量乘法的同态性质。

为了验证我们的设计，我们分析了设计的安全性，在标准服务器上实现了SPABox，并使用多个实际数据集和流量评估了其性能。 我们考虑性能指标，如吞吐量，频带-宽度和发件人，中间设备的内存开销和 接收器两侧，表明SPABox是实际两种长寿和短命 的连接。

其余 的 纸 被 安排 为 如下。 第二 部分描述了安全模型和协议要求。第三节介绍所提出的系统架构中，接着 通过详细的协议在第IV节的设计。第五部分讨论 了我们系统的实现。我们评估perfor-第六节曼斯显示效果和效率的 SPABox。第七节调查的相关工作和第VIII讨论可能的扩展，替代办法，并我们的方案和Blindbox之间的差异，然后通过结论 和 今后 的工作 中 科 九。

II. 小号ECURITY中号ODELS AND P ROTOCOL ř EQUIREMENTS

这项工作的目的是保护来自执行DPI的中间盒 的用户流量的隐私，同时利用高级DPI功能检测恶意流量。在换句话说，中间设备可以检测到过加密的流量与规则集和训练有素的机器学习攻击（ML）模型，这是由（第三方）提供规则发电机。在本节中，我们首先描述安全模型，然后概述协议要求。

A. 安全 模型

在本工作中考虑的场景中，我们假设 1）规则生成器是诚实的，并且2）中间盒 是*诚实但好奇的*，即，它将实现规则并诚实地遵循协议，但可能会尝试推断私有 执行期间从加密交通信息的DPI。

因此，我们提出的系统的目标是实现DPI过程，同时保护端点的数据隐私。更正式，安全 目标，这SPABox被设计 来实现被归纳为如下：

1) 为了保证无与伦比的保密通信， 即，不与已知的流量攻击在规则设置的关键字将保持从秘密的中间件。
2) 为了确保没有私人信息可以通过中间件来推断，所有的分析结果，只能看到由 客户端。

此第一个目标表明，允许中间盒仅学习 与中间盒规则集中的已知可疑关键字完全相同的数据，第二个目标旨在使中间盒无法应用数据分析 技术来推断有关的流量，例如如泄漏滥用攻击 [24]。

B. 协议 要求

基于以上安全模型，我们确定了协议应满足的一些 要求。由于我们的目标是 在中间设备上提供用户数据的隐私HTTPS连接，首先我们需要SPABox维护和扩展的 特性 提供 通过 现有的 TLS / SSL 为 如下。

私人连接。应该使用在会话开始时协商的秘密密钥对所有流量进行加密，并且协商应该既安全又可靠。请注意，只有端点可以读出未加密流量，而所有的暴露于中间件业务应保持加密状态，在所有的 时间。

身份认证。端点的身份可以相互验证，甚至可以与DPI设备进行身份验证。这可以使用公钥密码术来实现，并且可以选择以减少开销。

可靠的连接。会话的一方发送的每条消息的完整性应能够由该会话中的其他各方（包括受信任的中间盒）进行验证，以检测未经授权的修改并防止在传输过程中出现未发现的丢失。

此外，我们还要求SPABox满足以下新要求：

中间盒 透明度。在通信期间，客户端通常不直接与现有部署中的DPI中间 盒通信。为了符合当前的部署，我们尝试在协议中保留DPI设备的透明性。

隐私保护。在我们的案例中，中间盒的攻击者永远都不能读取两个端点之间交换的明文，也永远不能从数据中提取私人信息使用分析技术。TLS / SSL会话密钥在任何情况下都不应在中间盒中暴露给攻击者。

端点验证。为了安全起见，除了能够认证会话中另一通信方的身份之外，一个人还应该能够验证另一方是否遵循该协议以防止其行为
恶意地。

功能种类。该协议应该足够通用，以支持基于签名或关键字以及基于数据分析的DPI功能。

最小开销。最后，我们的协议应该在没有大量开销（包括带宽，延迟，计算等）的情况下运行。具体地说，建立连接的延迟和带宽开销应该很小，以便 支持短而独立的流和大规则集。端点处的计算开销应受到限制。我们还打算使计算独立于任何特定的 硬件 平台， 以便 所有 用户 都 可以从我们的协议中受益。

III. 小号YSTEM一个体系结构的设计

图1显示了系统架构，高亮显示的框表示SPABox添加的组件。如在预vious 工作 [43]， 存在 有 4 各方： 发送者 （S）， 接收器（R），中间盒（MB）和规则生成器（RG）。为了定义一个

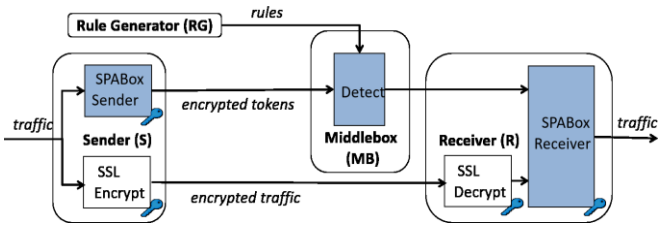

图1：SPABox系统架构

攻击时，诸如Symantec [13]和McAfee [4]之类的RG通常提供一系列攻击规则。每个攻击规则都包含一组关键字以及可能的其他信息，例如每个关键字的偏移量，两个模式匹配之间的距离以及一个正则表达式。此外，RG还提供了一种经过训练的模型，用于对加密流量进行分类，以确定流量是否包含恶意程序。MB通常由AT&T [46]等网络运营商部署。在安全的交通威胁可以通过两种方式确定：一）的MB的加密流量与规则集进行比较提供由RG观察到流量与规则集中的攻击规则之一之间的匹配；*b*）该MB进行分类 与由所提供的训练模型的加密流量的RG和来自分类结果由R来决定流量是否包含 恶意软件。

当S和R要在MB监视的 网络中建立HTTPS连接时，将执行以下步骤 ：

连接设置。当S和R尝试通信时，它们 首先运行SPABox握手以交换 SSL 会话密钥k_{ssl}，就像SSL握手一样。此外，在SPABox握手过程，S和R也需要协商7个 参数（SPAPara），*克*，*Ñ*，*- [R*，*s* ^，*Ñ* ²，*盐*和散列函数*H* ^（·），所有的这些都用于对SPABox协议中的加密，解密和检测。请注意，只有S和R涉及SPABox握手，该握手会从 现有SSL握手中引导出来，并保留MB透明 性属性。

发送流量。 在S，两个逻辑连接被设置，以 作为SSL连接和被称为 SPA分别连接。在SSL连接上，S使用未经修改的SSL加密流量 。在SPA连接上，S进行复制流量，将其令牌化，然后使用基于离散对数问题（第IV-A节）的建议方法对令牌进行加密。为了支持恶意软件检测，令牌需要在加密之前进行预处理（第IV-D节）。

MB检测。一旦MB通过SPA连接从S接收到加密的令牌， MB 将执行两个任务。 一个）的MB的加密流量比较与所述规则集提供由所述RG （包括关键字和定期表达）; *b*）MB使用RG提供的训练模型对加密流量进行分类；如果有匹配的交通和规则集之间并没有正规表达式 中 的 相应的 攻击 规则， 该 MB 可以选择丢弃数据包，并通知管理员或发出警告，就像什么正规MB会做通过未加密的流量。如果要进一步评估一个正则表达式的需求，它会被转发到ř用于进一步处理（第IV-C节）。对于使用ML检测恶意软件时，MB不应看到分类结果。而是应将它们发送到R ，并且 只能在R处看到（IV-D 节）。这样做的目的是防止MB上的攻击者使用 脚本来分析数据（第II-B节中的协议要求隐私保护）。

接收流量。在R接收到SSL流量后，R将使用常规SSL解密并验证流量。然后R将标记化和加密明文回收其将被 比较 用 的 流量从 该SPA连接。的
我们之所以采用这种方式，而不是解密和解密 来自SPA连接的流量以进行比较，是因为在我们的协议中，加密比解密要快得多。通过比较从流量所产生的密文经对所述加密的令牌的SSL连接接收 来自SPA连接，R可确定是否符合S在此会话遵循SPABox正确协议，包括两个关键字匹配和恶意软件检测使用ML（协议要求端点验证在II-B部分）。如果是任何差异，R可以认为S是攻击者和 立即断开连接。否则，R可以处理由含有分类结果的MB转发信息和正则表达式来决定从S以上的流量是否 SSL 连接 是 恶意的 或 不。

发件人	g，n，$盐_0$，d，r，s，N^2，H（·）
中间盒	$g^{盐_0}$，g^{d}，n，g^{s}，N^2，H（·）
接收者	g，n，$盐_0$，d，s，N^2，H（·），λ，μ

表I：协议参数列表

IV. Protocol Designs

在本节中，我们给出的详细说明我们提出的协议。首先，我们描述一个困难的问题，以此为基础 构建协议。然后，我们将介绍有关协议如何分别通过机器学习处理关键字匹配，正则表达式和恶意软件检测的加密过程。表I总结了每个实体使用的参数，其中*d*用于计算伪随机盐，（λ，μ）是用于解密恶意软件检测分类结果的私钥对（请参阅详细信息）在第IV-B，IV-C和IV-D节中介绍），其余内容已在前面进行了介绍（第 III节）。

A. 直觉

Blindbox需要之间互动初始化发送者 和中间盒（协议要求中间件反式parency在第II-B），其结果在速度较慢的连接设置。为了避免这样的相互作用，公开密钥cryp- tographic原语是优选的，我们的协议。因此，该设计能够满足我们的需求的协议关键字既效率和安全性是选择在其上建立协议的合适的计算难题。正如在协议要求所提到的，我们的目标是设计一种协议与所述 以下 3个 属性：一个）它不 依赖 于 任何特定的硬件支持，但可以有效地对所有各方共同努力来处理数据; *b*）MB能够执行必要的操作

对于不同类型的DPI功能上的操作恩加密后 的流量; Ç）没有信息交换的需要之间 建立连接时客户端和MB。请注意，第三 财产不能，如果我们使用任何可搜索加密方案[18]，满足[44]因为这些方案要求MB到 来自客户端，这可能会成为获取搜索键主要性能瓶颈。考虑所有这三个属性为考虑，我们构造基于我们的加密策略上的难解 离散对数问题定义为如下。

定义1.*离散对数问题。给定一个有限 群* G ^ ， *一个 元件 克* ∈ ģ *和 一个 元件b 中 的子组 所产生克，即，（克），找到一个整数x使得克*^{x} = *湾*

这样的整数*x*是*b*与底数*g*的离散对数。 离散对数问题是现代非对称密码术中最重要的单向函数之一。 许多公共密钥算法是基于它建立，如在广泛使用Diffie-Hellman密钥交换协议[28]和在ElGamal加密方案[29]。

当另外两个重要和有用的性质encrypt- ING 消息*米*作为一个元素*克*^{米}在 ġ 是 所述 同态
功能 上 除了 和 标量 乘法。 在 一个 概括地说，

这两个属性使我们的协议可以直接在加密流量上工作，而无需在MB上解密有效负载。

假设$c_1 = g^{m_1}$和$c_2 = g^{m_2}$，其中m_1和m_2是明文。该2个特性可以被归纳为如下：

- 同态 加成： 所述 产品 的$Ç_1$和$c \wedge_2$给出$g^{m_1 + m_2}$对消息进行加密（$m_1 + m_2$）。
- 同态标量乘法：提高$Ç_1$到所述p ò WER的$米_3$产生$克^{米_1 \cdot 米_3}$，其加密信息$米_1 \cdot 米_3$，其中， $米_3$是在另一个明文相同的消息空间$米_1$， $米_2$。

B. 关键字匹配

在本小节中，我们介绍了用于 在S，R和MB处执行关键字匹配的协议设计。注意，仅需对MB，S和R进行少量更改即可支持在IV-C和IV-D节中介绍的其他两个功能。攻击规则可能包含多个关键字以及这些关键字的位置信息。仅当一个攻击规则中的所有关键字都具有正确的偏移量匹配时， 我们才 认为 此 攻击规则 已 匹配。

*1) 发送侧：*图2示出了体系结构上 的SPABox 发送机侧。第一步是使用固定长度的滑动窗口标记要通过SPA连接 发送的 流量。例如，假设要发送的流量为“ NETWORKING”，并且每个令牌使用5个字节，则生成的令牌为“ NETWO”，“ ETWOR”，“ TWORK”， “ WORKI”， “ ORKIN”和“ RKING”。如果我们记号化中的关键字的 规则集在该MB在一个类似的方式，我们可以搜索 的 关键字，其长度都等于或大于5举例来说更大，说一个的 的 关键词 是 “网络”。 如果 该 MB 标记化此使用5个字节的窗口关键字长两个关键字令牌 “网络及通讯系统”和“TWORK”被生成。那么MB能比较各地收到的令牌SPA连接与这2个关键字标记，并看看是否有是2接收

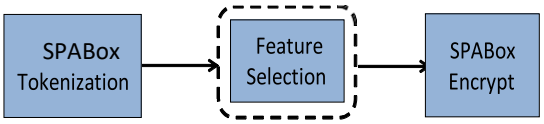

图2：SPABox发送器架构

标记，由2个标记分隔，分别等于这两个关键字 标记 。我们选择以这种方式标记流量 的 原因将在IV-D节中讨论。在我们的实现中，我们使用5个字节作为滑动窗口的长度。我们选择使用5个字节的原因是：（i）消息空间足够大，可以确保我们的加密安全；（ii）滑动窗口越长，加密速度越慢。注意，令牌 在加密之前需要转换为适当格式的整数（即mpz t ）。

令牌化后，S可以根据*离散对数问题*（在IV-D节中使用功能选择模块），使用我们的 加密方案对每个令牌进行加密。假设明文用于从所述标记化是获得一个令牌*吨*，所述密文*Ç*被给出为：

$$c = \text{克}^{\text{š}^{\text{一升}}_{\text{吨}}} \bmod n \quad (1)$$

其中*Ñ*是一个素数，*克*是一个元件在乘法Z *组和*盐*是添加组Z $_{n\text{中}}$的元素，所有这些都在连接建立阶段定义。其中一个可能找到的是我们的加密机制是相似到了众所周知Paillier密码的方法[40]，但有多大的不同，我们将在第讨论他们八。具体地，*盐*是用于随机密文，即是，确保没有两个连续的密文进行加密的相同的消息是相同的（如随机加密方法类似），并且因此使频率分析抗性加密的令牌。然而，如果一个新的随机*盐*被 选择用于每个令牌，它使所得到的协议难以 以支持高效的关键字匹配操作在该MB，

并且不再可能预先计算$g^{salt} \bmod n$ 。为了解决这个问题，在我们的协议，初始*盐* $_0$，$d \in$ ž $_Ñ$ 被定义。甲计数表也用于存储定义$ķ_{米}$，所述

每个明文次数*米*已经被加密，从而在远SPABox标记化模块。因此，为了加密一个 令牌*吨*，š第一长相向上的数目$ķ_{吨}$（即，如何许多

次令牌*吨*迄今）在加密计数器表和然后加密*吨*为$C \wedge = 克^{\text{š}^{\text{一升}}_{\text{吨}_ķ \cdot \text{吨}}}$米直径：d *Ñ*其中$盐_ķ = 盐_0 + ķ_{吨} d$ MOD ž $_Ñ$。注意的是$ķ$被最初设置为1。

出于安全目的，所述位长度 的*Ñ*应是在至少 1024（2048通过NIST推荐）。不幸的是，这将导致巨大的通信开销。为了维持在安全水平，同时减少开销，我们哈希每个令牌*吨*使用一个密码散列函数$ħ$（·）后到ķ enization和集*Ñ*到160个比特，从而其每个令牌可以是 代表

加密后的20个字节。在我们的实现中，我们使用哈希函数SHA-1，其输出大小为160位。这是一个有效的步骤，因为令牌的消息空间（40位）是大大 小于 该 160个 比特。 通过 此 修改， S加密

通过计算令牌

$$Ç = 克^{s_{一升吨_k}} \cdot ħ（吨） 米 ø d \tilde{N}（2）$$

其中$盐_k = 盐_0 + kd \bmod Z_n$。请注意，这可确保同一消息的至少两次连续加密不会产生相同的密文（类似于随机加密方法），因此使加密令牌具有抗频率分析的能力。不幸的是，当第一个加密令牌被标识为恶意关键字（即规则集中的关键字令牌之一）时，上述设计不够 安全（第IV-E节）。为了使我们的协议安全反对　　在这种情况下，要求S选择一个随机数r的良性令牌，并将此令牌序列设置为SPA连接上令牌化流量之前的前缀。 在从该S选用良性令牌可以是源可信第三方等证书颁发机构（CA）或任何系统文件。我们将此过程称为*良性前缀填充*（BPP）。

后 的 令牌 被 加密的， 它们 被 发送 出去一起 在SPA与辅助信息的连接（AUX-信息）$克^{小号一升吨_0}$，$克^d$，$-[R$ ，$\tilde{N}$和所述散列函数$H \wedge (\cdot)$，其是 由所使用的MB用于关键字匹配。一个能想到的在AUXInfo作为公共密钥对　　一届 特别 会议。对于

符号的简单性，我们表示等式。2作为$enc(t, k)$，其中 k 是令牌t到目前为止已出现在流量中的时间。 为了 防止 计数表变得过大，S复位$盐_0$每一个*中号*不同的令牌中发送，其中*中号*可设定为一个大的数字，说10M，所以作为以保持的开销为以最小的复位水平。

*2) 在中间盒侧：*为了支持关键字匹配，MB需要执行以下操作。规则 准备。 在 关键字 中 的 规则 集 需要 以成为

处理。 在 接收 的 规则 集 从 该 RG， 该MB

首先，使用一个长度相同的窗口来标记规则集中的所有关键字，R使用相同长度的窗口来标记通过SPA连接发送 的流量。对于长度的关键字$升$，$\lceil 升 \rceil$关键字令牌被生成。那么MB将改变所有 切分 的关键字 到 大 编号 为 提到 前面

（即mpz t 格式）。

当S尝试向R发送流量时，MB将从S接收到一组AUXInfo，通过这些信息，MB可以哈希然后将所有标记化的关键字w 加密为$enc(w, 1)$。这里$k = 1$的原因是MB尚未在流量中看到关键字w。值得注意的是，在我们的协议，没有需要S和MB之间的信息交换过程中 的规则准备步骤，而这是在[43]的主要性能瓶颈。我们将在第六节中比较开销。

关键字匹配。使用加密的关键字令牌，MB 可以按以下方式执行关键字匹配。首先考虑一个简单的示例，其中有一个接收到的加密令牌令牌$enc(t, 1)$和一个加密关键字 令牌$enc(w, 1)$。为了检查是否*吨* 等于*瓦特*，所述MB仅需要检查是否$ENC(T, 1)$ 是等于 以$ENC(瓦特, 1)$。 一旦 匹配， 该 MB 只是 计算

$enc(w, 2)$为$enc(w, 1) \cdot g^{d \circ w} \bmod n$，并替换$enc(w, 1)$

这一新 的未来，因为值 时间S发送 令牌*牛逼了*

在SPA连接中，相应的密文为$enc(t, 2)$。接下来，我们用多个加密的关键字标记扩展一个关键字匹配示例，并收到一个

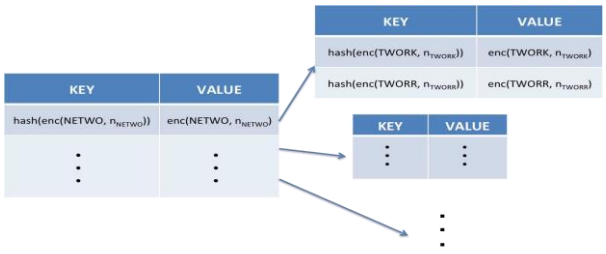

图3：分层哈希表

加密令牌。为了支持搜索多个关键字标记，MB需要维护一个标记表，该表 记录每个关键字标记出现在流量中的次数。一旦存在匹配项，MB可以像一个关键字令牌的情况一样重新加密匹配的关键字令牌，并用新的加密关键字令牌替换令牌 表中的匹配条目。但是，实际的规则 集可能有几百加密关键字标记的[11]，*如何能 在 MB 执行 一个 有效的 搜索*?

第一个想法是建立一个搜索树， 每个 元素都是一个加密的关键字标记。当通过SPA连接从S接收到加密的令牌时，MB可以查找树以查找是否存在匹配项，这使得搜索时间与关键字令牌的数量成对数。然而， 即使对数搜索时间，这将仍然导致开销太大：一）拥有超过万个规则集合这是典型的关键字（不同的关键字令牌的数量可能会更多），它可能需要更多比15比较以找到匹配项。*b*）由于遇到一个恶意关键字是相对罕见的，计算能力可以 被上寻找一个加密的令牌不存在浪费在搜索树。*c*）由于关键字是标记化的，因此匹配规则可能需要在树上进行多次搜索。在我们的《Snort新兴威胁》 [11]中的规则集中，一条规则最多有81个关键字标记。相反，我们的解决方案　　既简单又有效。我们没有建立搜索树，而是

使用令牌哈希表（THT）。 的 键 为 每个 条目是的每个加密关键字令牌中的散列值*瓦特*，并且每个条目的值是一个元组（$ENC(瓦特, \tilde{N}_{瓦特})$， *瓦特*）。我们需要存储w 本身的值 的原因是为了在发生哈希表冲突时打破联系。而且，THT可以基于MB保持计数的令牌表来构建的 数目 的 外观 的 每个 关键字令牌。 一旦 一个

建立匹配后，MB可以从THT中删除相应的密钥和值对，重新加密密文并将其作为新条目插入。

截至目前，我们已经讨论了如何高效地搜索 规则集中与数百名的加密令牌关键字标记 在 了 MB。 该 问题 留下 来 的答案 是 *如何将多个匹配的高效组合关键字标记一个关键词，因为关键词被也被标记化 MB？如何将多个匹配的关键字映射到特定规则， 因为 一个 规则 可能 包含 多个 关键字?*

我们试图用一块石头杀死两只鸟。洞察这里 是*一旦MB找到一个匹配的关键字 标记，它可以减少 了 搜索 范围 为 在 未来 的 道理， 这 是 无论是*

下一个的所述第一令牌的关键字 在相同的规则或所述 下一关键字与所匹配的令牌一起关键字令牌 构成 一个 部分 的 一个 关键字。 对于 例如， 如果 有 是只有2个关键字中的规则集与第5个字节为“网络及通讯系统”，即”网络”和‘NETWORR’，并有一个匹配的令牌‘网络及通讯系统’在交通中，MB只需要找出来 如果“NETWO”之后的第三个加密令牌是“TWORK” 或“TWORR”。因此，我们可以构建如图3所示的层次 哈希表（HHT），以保留有关 如何对每个规则进行标记的信息。HHT的第一级只有一个哈希表，其每个条目对应于每个规则中first关键字的前五个字节。一级哈希表的每个条目还指向一个新的哈希表，其中包含所有可能的后续关键字标记，依此类推。每个关键字标记的位置信息也可以轻松地嵌入到HHT中。由于S使用以下方式标记流量 的滑动窗口，该偏移的每个接收到的加密令牌上 的SPA连接可被容易地推导出来。

HHT指示每条规则如何用关键字标记构造，一旦 从RG接收到规则集，就可以在MB中预先构建甚至硬编码，而THT需要 专门为每个连接构建。请注意，可以通过将哈希表与相同条目组合在一起来进一步压缩HHT ，我们将其保留为将来的工作。的存储器开销和时间设置THT将评估 在第 VI。

*3) 在接收器侧：*为了防止期从被恶意和发送非法代币，R在接收流量 从 所述 SSL 连接 和 流量 转发 由 所述 MB上的SPA 连接可以先解密SSL加密通信，然后标记化并对明文形式的SSL流量进行加密。然后，R可以将SSL流量生成的密文与SPA连接中的加密令牌进行比较；R还需要检查是否该AUXInfo 通过将 S发送的消息与其自己的副本进行比较， 是有效 的。仅当两者都匹配时，通过SSL连接接收的流量才被视为安全。可以使用SSL中指定的标准方法对SSL加密的流量进行解密。我们继续进行下去的理由在这种方式，而不是解密，然后detokenizing的从交通SPA比较连接是tok- enizing和加密的明文等式。2一起是多比从解密通信，速度更快SPA 连接。此外，R可以开始该比较过程，只要它得到SSL流量和所述第一加密令牌在SPA连接。具体来说，R可以首先解密，标记化然后加密SSL通信。只要R通过SPA连接接收到加密令牌，它就可以立即进行比较。然而，这个过程可能会导致在R侧有一些延迟，但 我们 认为 这 是 一个 可以接受的 折衷。

C. 正则表达式 评估

评估正则表达式的原因有两个。 首先，除了关键字匹配之外，某些规则集还需要对正则表达式求值。通过启用此类操作，可以解决许多公共和工业规则集的所有 攻击规则[43]。举例来说，在我们从规则集的Snort新兴的威胁[11]，一些中的规则被允许到是使用与Perl兼容的正则表达式编写。以新兴信息。规则中的规则74为例：

警报http $ HOME NET any → $ EXTERNAL NET any flow：
已建立，到服务器；
内容：“ 8866.org”；
PCRE ：“/主机\ X3A [\ - [R \ Ñ] * \ x2E8866.org/Hi”

在 此 示例中，如果可以在流量中找到关键字“ 8866.org”的匹配项，则将触发由动作“ pcre”表示的 正则 表达式 评估。其次，正则表达式可以帮助MB检测关键字的是少于5个字节长。注意S的标记化交通使用5字节长的滑动窗口，以便MB可搜索的关键字，其长度都等于到或更大的比
5.然而， 如果一个规则包含的关键字较短 那5个字节，方法部分中提及IV-B将不工作。在我们的规则集，该关键字的25％以上都超过5个字节长的短。为了能够对长度小于5个字节的关键字进行匹配，MB可以轻松地根据每个规则从长度小于5个字节的关键字中构建一个正则表达式，从而使关键字搜索问题成为正则表达式评估问题。因此，在我们的情况下，匹配正则表达式可启用完整的关键字匹配功能。

接下来，我们尝试回答*如何通过加密流量有效评估 正则表达式？* 一个 稻草人 溶液是嵌入SSL密钥k_{SSL} 在加密令牌，并且一旦有一个加密的令牌中发送之间的匹配比 的 SPA连接和一个关键字令牌在MB中，MB 可以 提取k_{SSL}通过重新加密的匹配的令牌，例如
$ENC(吨, \tilde{N}_{吨})$与$\tilde{N}_{吨}1$ ，并与异或它$ENC(吨, \tilde{N}_{吨}+1) \oplus k_{SSL}$ 被作为一对连同发送$ENC(吨, \tilde{N}_{吨})$在SPA被S连接。然而，三个问题可能会出现在此
解决方案：*a*）流量可以增加一倍；*b*）匹配的一个 部分关键词令牌将将揭示SSL密钥到 的MB; *c*）MB被授予某些用户过多的权力 。特别是，第二个突破了安全性要求。举例来说，如果仅在MB关键词是“网络”和交通含有“NETWORR”，这不 匹配 的 关键字， 在 SSL 密钥 将 仍然 被曝光 以来令牌“网络及通讯系统”匹配的MB。因此， 我们需要能够执行规则的协议表达评价 和 保存 用户 数据 隐私 在 该 相同 的时间。

通常，正则表达式可被转换到一个去 terministic有限自动机（DFA）与接受状态$^{\circ}F$指示输入字符串中包含的恶意因素如果这个DFA状态结束$^{\circ}F$。甲DFA 将作为输入在所述加密的流量SPA连接一个字节在时间[30]，[19]。以隐私保护方式执行正则表达式的一种方法是使用Eq加密每个单个字符。2.但是，这将使生成的协议容易受到暴力攻击。而不是透露SSL MB的密钥，以便MB可以解密流量以运行正则表达式，我们建议将正则表达式 发送给R，然后让R解密SSL加密流量并运行 正则表达式匹配算法，这只会导致 较小的 开销 （第 VI节））。 如果 在 相应的DFA CAN

达到接受状态时，R会知道S是恶意的（攻击规则的命中）并且可以丢弃数据包。如果RG出于任何实际原因不愿公开其规则集，则可以使用*Yao的乱码*[49]和*2分之一的遗忘转移*（OT）[26]来向最终 用户隐藏正则表达式。

为了进一步改进我们的协议的性能，而不是使用姚的 ，其被设计乱码电路技术对于一般的电路，我们应用高效错乱技术即是定制的DFA从[37]。注意的错乱的DFA的 可以在连接建立之前完成下线以来的错乱过程中并不需要从信息的其他方，但在MB本身。因此，它不会影响MB的运行时性能 。

值得指出的是，即使先前已匹配相同的常规表达式，MB仍必须为每次匹配发送“新的”乱码DFA。这样可以确保R不会得知相同的规则重复被匹配。请注意，这并不意味着DFA每次比赛都需要即时乱码。相反，MB可以预先计算并存储它们以备将来使用。

在我们的协议的一个实际的假设是，RG将提供MB各种的DFA和MB将发送的 乱码DFA至R而不让其中R知道DFA中被执行。有了这个前提，在我们的协议，R将不会有足够的信息来发现的结构的任何 特定DFA，因为R是不能够学到什么乱码 DFA 它已经计算以及是否出现乱码的DFA它已经计算出 如此 远 的 相关 性 或 不。

回想一下，当正在传输的流量具有匹配项时，会将乱码的DFA发送给R，该匹配项可以触发正则 表达式评估。一个可能被关注的是一个攻击者 可以导致MB送出去很多乱码的DFA至R只是通过发送匹配的令牌流量触发一个正规表达式求值，这将成为拒绝服务上的r网络链接。请注意，要实现这种攻击，攻击者需要了解哪些关键字可以触发正则表达式评估。 然而，在此文章中，我们假设关键字的信息保持从用户的秘密，它是唯一已知的（值得信赖）规则发生器。在此期间，因为底层加密 的关键字可以提供语义安全性，攻击可以学习任何关于关键字本身。因此，这类的攻击是从场景正在考虑淘汰的纸。

另一个关注的问题可能是由于R运行乱码 DFA以稍稍明文净荷（即， 该 有效载荷是未加密但具有随机字符串输入到替换的乱码DFA）， R将至少学习有效载荷，显然匹配一个正则表达式基于规则。可以通过在DFA上添加虚拟操作来避免这种情况。结果，在使乱码的DFA与某种纯文本有效负载一起运行时，它将使R感到困惑。

D. 通过机器 学习检测恶意软件

在本节中，我们将展示我们的协议如何使用ML分析启用对加密流量的恶意软件检测。回想起那个为关键字匹配，S需要来标记流量用 的滑动窗口，其也被称为*n-gram中*或*带状疱疹*。在过去的几十年中，研究人员提出使用*n-gram*表示使用ML方法进行恶意软件检测的功能[45]，[34]。在我们的协议中，我们将使用*支持向量 机*（SVM） [41]。

*1) 在规则生成器侧：*的主要工作的 RG是培养一个ML模型（SVM模型）。假设该 RG有两套文件，其中分别由恶意软件和良性程序的集合，的。 在特别地，恶意软件程序包括各种形式的的敌对或侵入软件，如蠕虫，特洛伊木马 马和病毒[16]。为了建立代表所有文件中的每一组中，RG应提取正克（功能）的所有文件中的每一组将作为该组特征。为了减少 对在我们协议的特征维度中，我们要求RG计算每个集合中每个n-gram的频率，并使用前 *K个*最常见的n-gram作为代表该集合的特征。然后，RG可以训练一个SVM模型和获得必要的参数 瓦特 和*b*为所述以下决定功能

$$°F(X) = 瓦特^{\top} \cdot X + b \quad (3)$$

其中，X是所述特征向量和瓦特是一*米*维向量$(w_1, w_2, \cdots, w_K)$。需要注意的是一对（w∧，*B*）可以定期更新， 通过 对 RG中 为 他们 收到 新 的恶意软件报告每天[15]。我们还将在VIII-B节中讨论SPABox如何支持其他ML模型。

2) 在发送方： 为了在 MB上执行ML，需要首先提取对象的特征。但是， MB提取一个输入对象的特征并不是一件容易的事，因为很有可能并非所有重要特征都是恶意关键字，而MB只能通过我们的关键字匹配协议保证知道这些关键字的模式。因此，在我们的协议，选择的作业的输入对象转移至S谁可以做到的功能这 通过在几乎没有成本的计数表，其存储的频率 的 所有 令牌 是 已经 出现了 这样 远。 具体来说，S首先使用图2所示的特征选择模块来选择特征，例如最频繁的前400个n-gram， 即$K = 400$。 然后 小号 加密 这些 正克， 那 就是， 对于$我 = 1, \cdots, ķ$，ś加密$X_{我}$为$X' = 克^{小号.X_{我}}$中号öd $\tilde{N}^2$，其中$\tilde{N} = PQ$与P，Q为相等的比特长度的两个素数和*小号*是一个随机元素在 ž $\tilde{N}$。 之后 是， 小号 发 的SPA连接上的加密功能与g^s和N^2一起使用，以便MB可以执行分类。为了确保令人满意的安全性，我们选择长度至少为1024位的p，q。但是，实际上，使用1024位长的p和q会导致开销很小。注意，只要RG可以提供相应的训练模型，就可以实现自己的特征选择模块。

*3) 在所述中间边：*最显着的特征的SVM是核函数，其作为的映射数据 ，以改善它的相似的线性可分集。在我们的工作中，我们使用简单的线性核[32]，我们将在第VIII-B节中讨论其他可行的核函数。召回该每个 加密 的n-gram 是 的 该 形式 如 等式2。

从S接收到加密的n-gram（特征）后，MB使用训练有素的SVM执行以下步骤

模型：（i）用于*我*个输入到所述决定功能（与*我* = 1，⋯，400），产生加密的n-gram x 的力量值w_i，即x' $\bar{w}_i = g^{s \cdot x_i \cdot w_i}$；（ii）乘在一起所有所述 中间 结果 产生 在 所述 前 步骤；（iii）相乘 的 结果 产生 在 步骤 （ii） 由*克*^{SB}和然后通过*小号*$^{\tilde{N}^2}$ g^2（此乘法使R能够学习分类符号）以加密形式给出分类结果，*小号*（$\sum$ *瓦特*·$x + b + \tilde{N}^2$）即g $^{我=1\,ii}$ ^{2}。最后，MB可以发送在步骤结果返回（iii）至R.所述的正确性烯加密后的分类结果被保证 在部分所述同态特性IV-A。

*4) 在接收器侧：*为了得到的分类结果，所有的R需要做的是 ，以 解密 该 分类 使用由MB结果SENT（λ, μ）（私钥在Paillier密码系统[40]），然后检查是否它是大于或小于 比 0。 不幸的是， 在 后面的 步骤 不能 被 做直接因为在我们的加密方法中，令牌（采用mpz t格式） 是非负的。但是 我们 观察到 的是 该 分类由等式 输出 的结果 。 3 是 内 的 间隔 $[-\tilde{N}^2, \tilde{N}^2]$。

因此， 在 为了 要 恢复 的 迹象 中 的 分类结果，R可以解密分类结果，然后进行比较它与N^2。如果恢复值大于N^2，则R

$^{2\uparrow 2}$加密令牌是关键字，在对手能知道那知道 是 它 是 一个 积极的 结果; 否则， 它 是 一个负结果。 在 正确 的 这一 操作 是简单的，因为如果 °F（X）=瓦特$^{\top}$ · X + B> 0，则 °F（X）+ $\tilde{N}^2 > \tilde{N}^2$。

E. 安全 保证

关键字匹配。回想一下，在关键字匹配操作中，S需要标记流量，执行BPP，然后使用Eq加密标记。2.在MB端，它检查规则集中是否有任何加密的关键字令牌。
等于通过SPA连接接收到的加密令牌。
在第IV-B节中，我们提到了BPP旨在避免前几个加密令牌是恶意关键字令牌的情况。如果没有BPP步骤，则生成的协议将变得不安全。为了说明这一点，假设生成的令牌是“ NETWO”，“ ETWOR”和“ TWORK”（分别表示为m_1，m_2和m_3）以及第一个令牌

米$_1$是一个恶意关键字令牌而在其他2是良性。因此，MB可以学习什么是第一 加密 令牌。作为 该 MB 也 知道*盐* $_0$， d 和 所述计数器 *k*对于每个不同的令牌开始从 1， 该 MB 可以学习

米$_2$由拾取消息*米* = *ETWO*$_2$*，加密它和比较生成的密文与接收到的 加密

到*k*烯*克*$^{s\,一升吨_1 \cdot 米_2}$，其中，*是一个字节和罐*h*一个v *e*向上到256种 可能性。*米* $_2$不能是一个恶意关键字令牌;
否则，MB将找到与规则集匹配的内容。显然，通过反复这样做，MB绝对可以收回*米* $_2$内恒定的时间（即256个比较中的最坏的情况），这会使协议不安全。同样，MB可以学习以下所有标记。但是，如果S适用令牌，因此不能发现的模式加密
如果S在没有BPP步骤的情况下进行传输，则按原样分配令牌。

为了找到合适的*r*值，我们再次考虑

第一个令牌是恶意关键字令牌的 情况
应用所述规则set.After BPP，第一*r*令牌是良性和 所述 （- *[R* + 1）个 令牌 就是 一个 恶意 令牌。 注意 的是该目标 的 应用 BPP 是 给 扰乱 了 原有 模式的加密 令牌。 在 其他 的话， BPP 是 预计 到 带来 的良性 令牌 这 会 也 出现 在 了 交通。我们表示这个事件由*C*表示，*C*发生的概率由Pr [*C*]表示。假设由BPP处理拍摄每个标记将出现在通过加密令牌序列原的SPA的概率连接*p*。然后PR [*C* ∧] = 1 − （1 − *p*）$^{-[R}$。注意，MB未知值*r* 。因此，它是 不可能的（至少是不可行的）的MB推测出什么*[R*是。从而， 一个 平均 情况*[R*将足够用于实际应用程序。
我们提出的协议的安全目标是为不匹配的流量提供不可区分性。在高层次上

$^{2\uparrow 2}$给定加密令牌那是不等于以任何关键字在到目前为止， 根据规则集，没有多项式时间的对手可以推断出 这些标记的任何信息。但是，当给定它们位于关键字集中，而没有学习有关基础关键字的其他信息。

$^{2\,2}$要评估我们的加密方案的安全性，首先我们给出了*决策Diffie-Hellman问题（DDH）的定义。*

定义2. *决策Diffie-Hellman问题（DDH）。如果没有针对*G的*DDH算法*A *，则族*G*满足DDH假设，使得对于某些*$\alpha > 0$*且*n*足够大：*

$$|\Pr[A(G, g, g^a, g^b, g^{ab}) = true] - \Pr[A(G, g, g^a, g^b, g^c) = true]| > \frac{1}{\tilde{n}^{\alpha}}$$

然后，我们可以通过以下安全性博弈来定义所提议协议的不可区分性。

定义3. *安全游戏。考虑带有算法（Setup ，Enc ）和相关消息空间*M*的协议。*
*设*A*为ppt对手。安全游戏*Exp$_A$（1^{λ}）*是定义如下：*

1）盐$_0$，$d \leftarrow$设定（1^{λ}）
2）$T^0 = (t^0_1, \cdots, t^0_n)$，$T^1 = (t^1_1, \cdots, t^1_n) \leftarrow A(1^{\lambda})$
3）$b \xleftarrow{\$} \{0, 1\}$
4）$c_1, \cdots, c_n \leftarrow$ Enc（*盐*$_0$，d，$t^b_1, \cdots, t^b_n$）
5) $b' \leftarrow A(c_1, \cdots, c_n,)$
6) 如果$b' = b$*，则输出1。*

我们说，对于所有ppt对手来说，该协议都是安全的一个S*，并为所有足够大的*λ*：*

$$\Pr[\text{实验}_A(1^{\lambda}) = 1] \leq \frac{1}{2} + \text{negl}(\lambda)$$

在 通过SPA连接发送BPP到令牌化流量之前，可以避免这种不希望的情况，因为*r*是MB无法学习 的随机数 。由于MB无法学到这些的加密形式*[R*良性令牌， 该 MB 是 不 能够 来 上班 了什么 的良性

基于以上定义，提出的协议可以实现
上述引理1中所述的安全目标。

引理1.*假设DDH问题难以解决，那么我们在IV-B节中提出的关键字匹配协议对于那些不在规则集中的令牌是安全的。*

证明。如果有攻击者一个在谁能够在MB学习的私有值任何进入加密的权标有不可忽略的概率，那么我们就可以用这个攻击者一个作为一个黑箱，以构建能解决的算法DDH
有效率的。

更具体地，给定的DDH的一个实例（*克*，*克*^{一个}，*克*^{b}，$\check{Z}$），我们想建立一个算法Ç使用甲的解决能力为这样b。为此，我们可以构造一个算法C ，该算法将在A玩的安全游戏中充当挑战者，其过程如下：

设置阶段。C如第IV-B节中所述定义*盐*$_0$，d，并计算$h = g^{盐_0} = g^{a} \bmod n^2$和$k = g^{d} = g^{b} \bmod n$。然后，$h$，$k$将作为协议的公共参数发送到A。

挑战阶段。A提交$T^0 = (t^0, \cdots, t^0)$和$T^1 = (\dot{t}^{吨^1}, \cdot\tilde{n}\cdot\cdot, 吨^1)$。Ç翻转一个硬币$b$从{ 0， 1 }和然后加密$\check{T}^{b}$

通过计算$c_i = (g)^{i} \cdot (g)^{i的一}$组令牌，然后
转发该组{ C_1， $\cdots$， $C_{\tilde{N}}$ }到甲。

输出相位。对手A输出 b'。如果$b' = b$，则A获胜。

我们可以看到，算法C 被构造为A的质询者。现在，我们可以从下面检查$ç_{我}$是的正确 形式为密文加密明文*牛逼*^{b} 在我们的建议协议。

$$(g)^{在^{b}}_{我} \cdot (g)^{AB吨^{b}}_{我} = (盐_0)^{t} \cdot (g)^{广告笔^{b}}_{我} = salt^{b}_{a \cdot t}$$

其中 *盐*$_{-}$ $= 盐_0 + 广告$。这是一个有效的密文，因为一个不具备的知识*摹*^{DT}无与伦比令牌*牛逼*，因此不能决定*gkdt*其中 $\dot{k} \in \check{z}$ 。因此，从进攻的角度来看一个，他与交互他在一个安全的游戏所提出的挑战者协议。

最后，Ç确定其在该实例（*克*，*克*^{一个}，*克*^{b}，$\check{Z}$），$\check{Z} = 克^{AB}$

如果A输出$b' = b$；否则，决定z g_{ab}。

需要注意 的是 一个 具有 的 能力 ，以 打破 该 提议 的协议。因此， 如果 它 是 确实$\check{z} = 克^{AB}$， 然后 所有 密文中{ c_1， $\cdots$， c_n }的格式正确，因此A决定它们是T^0还是T^1的密文。如果$\check{z} ƒ = 克^{AB}$，然后所有密文将成为随机值，并因此阿将是可能以失败与概率与1/2 。因此，我们有 的优势 的A as

$$进阶_{A} = \Pr[A(^{I\lambda} 获胜)] - \frac{1}{2}个$$

作为 一个 结果， Ç也 具有 优点 等于 至 进阶$_{甲}$。 然而，没有已知的算法可以有效地区分DDH问题，这表明C区分DDH问题的概率可忽略不计。这意味着该 进阶$_{甲}$也是可以忽略不计，即，进阶$_{甲} \leq$ negl (λ)。因此，在我们的提议中，MB中没有这样的对手A可以打破不匹配令牌的不可区分性
协议。 □

另一方面，我们协议提供的安全性还取决于所使用的哈希函数的安全性。请注意， 安全散列函数要求从其散列值恢复 消息输入是不可行的。现在我们看到，我们的协议满足安全目标，只要参数 散列函数和离散对数问题被 适当地 设定。

正则表达式评估。回想一下 ， 在 我们 用于正则表达式评估的协议中，每个传入的 加密
到$\dot{k}$烯OV ER的小号P甲连接是的的形式$C \wedge = 克^{s一升吨_{\dot{k}} \cdot 米} \text{MOD} \tilde{N}$。正如我们已经证明引理1以及每个令牌的长度为5个字节一样，m的明文空间约为Θ（256^5）= Θ（2^{40}）。显然，这不能在多项式时间内用蛮力法处理。这确实 阻止我们
协议容易受到暴力攻击，但它会导致执行正则表达式评估的DFA的规模很大，因为从状态到状态的转换选项（即DFA字母的大小）与一个令牌的可能情况（大约 Θ（2^{40}））。显然，这在现实生活中是不可接受的。因此，我们的协议采用了Yao的乱码电路技术[49]（或类似的IV-C部分所示的乱码技术）来保持DFA尺寸小，同时提供所需的 安全性。 然后 1-出-2 不经意 传输 [26]来
在 以 帮助 [R 知道 的 随机 字符串 对应 于 该输入 到 该 电路 没有 让 该 MB 学到 什么 字符串

R得到。结合这两种技术，我们的
协议可以确保MB无法学习良性标记，并且R对正则表达式 的基础结构一无所知 。正式地，我们可以将我们的安全保证总结 为：

引理2.*该协议对恶意软件是完全安全的*

接收者，并针对诚实但好奇的中间人框。

证明。由于我们的协议是基于[37]，[26]中的技术构建的，因此可以参考它们以获取详细的证明。 □

通过机器学习检测恶意软件。回想一下，此过程在很大程度上取决于IV-A节中所述的同构性质。所有的主要计算

机器学习 过程 中进行 的内 MB 和
仅 在 加密 分类 结果 将 被 转发 到

R。因此，此操作所需的安全保证是

1) R不知道训练有素的机器学习 模型的参数。
2) 该 MB 不能 学 的 分类 结果。

与正则表达式运算不同，R不会对主要结果进行运算，除非解密分类结果。它是

不难看出，我们提出的协议可以达到安全性目标（i），因为R仅具有分类知识

结果和输入特征，这不足以使用一个方程式求解所有未知参数变量。至于安全目标（ii），这是源于IV-A部分的两个同态属性以及上述关键字匹配操作的安全性分析的结果。一个可以看到的是，加密也类似于Paillier密码系统， 并且可以参考[40]进行了详细的证明。具体而言，在 一方面，通过提供的两个homomophirc属性我们的加密 方法 确保 其 所述 得到的 密文 从

求幂和乘法是正确的加密的形式，即，*克*$^{s一升吨_{\dot{k}} \cdot 米}$与*米*作为所述得到的状子èXT; 就在其他方面，我们已经证明上面那个没有袭击者在该
MB可以学习任何私人价值不是来自一个关键字一个 结构良好的密文。因此，有了这两个保证，MB中的任何攻击者都无法获得由机器学习过程产生的结果，这就是我们的安全目标（ii）。 此外， 作为*小号*（盐）用于在加密的的n-gram是

与用于加密令牌（等式2）的盐不同，MB无法将n-gram与任何关键字令牌相关联，即使它们的明文相同。

V. 小号YSTEM我MPLEMENTATION

在本节中，我们显示了以下内容的详细实现

用于RG，客户端和MB的SPABox。

在规则生成器端。LIBSVM 3.21库[25]被修改并用于训练SVM模型。

在 客户 端。 我们在OpenSSL-1.0.2d库[7]之上为C中的客户端实现 SPABox 。我们还修改了OpenSSL库中的SSL握手过程，以便我们可以提取AUXInfo。GMP 6.0.0库用于将每个令牌（我们选择为5个字节）转换为大整数（ mpz t格式），然后按照等式1中的说明对每个令牌进行哈希和加密。2使用相应的大整数。我们选择g， $salt_0$，s，d，n和N^2 分别为80、20、20、10、160和4096位。我们使用的哈希函数是SHA-1。根据我们的安全性分析，r设置为20到
40.当s打开一个连接，它创建两个插座，一个 用于SPABox握手，发送正常HTTPS通信和所述另一个用于加密令牌传输。特点为ML的标记化业务后发送。 R的实现与S的实现类似，只是电路出现了乱码[37]和OT [26]。若R成功匹配 由MB发送的正则表达式的流量或得到 一个 积极的 结果 为 恶意软件， 它 停止 了 连接。

在中间箱侧。我们在带有DPDK [1]的Click模块化路由器[33]中实现MB 。我们基于Google密集哈希图[2]构建THT和HHT。我们让THT开始与 65536个 插槽 和 调整大小 时，它更超过50％满。对于*关键字匹配*，所有线程的一半被用于对在加密令牌匹配THT，和一个线程用于以 搜索 在 该 HHT 如果 有 是 关键字 的令牌 相匹配。如果规则匹配且不需要进一步的*正则表达式*求值，则MB将阻止连接并通知R；如果正则表达式需要给被运行在所述接收器侧，
[26]用于不正确地传输对应于乱码DFA的输入密钥字符串。螺纹的其余部分被用于*恶意软件检测*是基于实现上GMP 6.0.0库。如果未阻止连接，则ML分类结果将与其他流量一起发送到R。MB和客户端通过园区中的1Gbps LAN连接。

VI. P ERFORMANCE Ë估价

为了证明我们提出的协议是可行的，在本节中，我们将在客户端和MB端都显示SPABox的性能评估。SPABox需要（i）在客户端进行额外的令牌化和加密；（ii）建立并搜索多个哈希表，乱码和OT以进行正则表达式评估，并在MB端执行分类。本节中显示的统计数据是平均结果。

	盲盒	水疗盒	Paillier
加密（5 个字节）	硬件支持101 *ns* 无硬件支持1022 *ns*	1015 *ns*	20.6*毫秒*

表II：SPABox和Blindbox的加密微基准

S比较

A. *数据 集*

Snort Emerging Threats [11]中的关键字规则集具有大约3K规则，其中包含11,202个关键字和21,035个不同的关键字标记。

对于恶意软件检测，我们使用两个不同的数据集：恶意软件数据集和良性数据集。我们的恶意软件主体包含来自VX Heaven [14]和Microsoft恶意软件分类挑战[5]的17258个恶意程序，其中包括代表不同类型恶意软件的不同恶意软件家族。我们的良性数据集包含1000名合法的可执行文件和动态链接图书馆的（DLL）中，大多数的它们是从运行的机器采集系统文件对我们的校园。我们使用这些文件作为培训的70％设定，并在其余的30％为中 测试 集。

我们运行SPABox客户端和MB在合成流量包含了用于恶意软件检测的数据集，有效载荷从ICTF2010网络跟踪[3]和36中提取长达一小时的未加密的 现实 世界 的痕迹 收集 在 该 访问 链接会在和出的我们的校园。

B. *性能*

现在，我们调查SPABox在客户端，MB和网络上的性能。据我们所知， Blindbox [43]是唯一 能够通过加密流量（即关键字匹配和正则表达式）启用本文中考虑的DPI功能的 一部分 的系统。我们实现了加密方法以及Blindbox中使用的关键字匹配算法，并将其与SPABox进行比较。所有统计数据均为平均结果。

*1) 在客户端：*两个配备Intel Core i7处理器和16GB内存的台式机用于运行我们的客户端原型。这些机器是多核的，但是每个客户端只 使用一个线程（吞吐量测试除外）。启用了超线程。CPU支持AES-NI指令，因此 认为 我们 可以 比较 我们的 解决方案 与 该 Blindbox。

如何 长时间 没有 它 需要 的 小号 来 加密 一个 令牌？ 表II显示了使用SPABox和Blindbox进行加密的微基准。在SPABox中，对一个5字节块的加密平均需要1015 *ns*。此定时结果包括时间用于将令牌插入一个大的整数，散列与SHA-1和加密。与Blindbox相比，我们的解决方案采用9 ×更多的时间。的主要理由是该Blindbox需要
优势 的 硬件 支持 的 AES 加密。 不带
硬件支持方面，Blindbox花费的时间与SPABox相似。与Paillier相比，我们的加密方法节省了近20倍的时间。
如何 长时间 没有 它 需要 为 [R 来 评估 常规 表达式？
多少 开销 不会OT招致？ 在SPABox，ř需要执行，以防MB正则表达式的评估发现该交通可疑（由关键字匹配）。这个过程

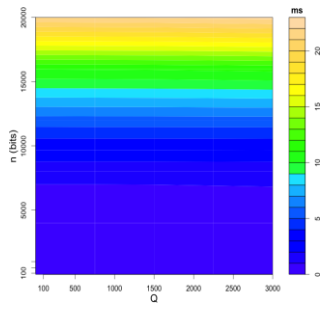

图4：正则表达式评估时间

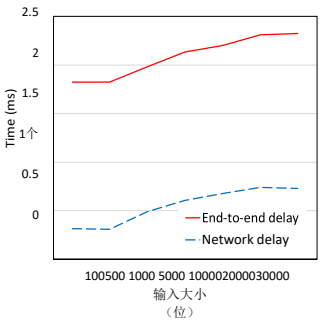

图5：端到端延迟到OT

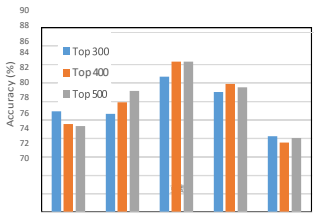

图6：恶意软件检测精度

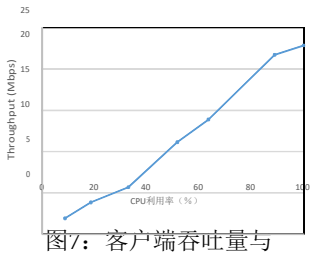

图7：客户端吞吐量与CPU利用率

在接收方增加了更多的计算开销。无花果。

图4显示了在给定不同输入字符串长度（*x*轴）和DFA大小（*y*轴）的情况下R处的正则表达式评估时间。它表明，随着输入大小的增加，评估时间线性增加，但随着状态数的增加，评估时间却缓慢增加。这是因为，对于每个位，仅需要进行一次哈希计算，这在评估时间中占主导地位。用于处理的正则表达式的时间可通过优化来进一步降低该底层 数据 结构 （即，DFA） [38]，[17]，但它是超出这个范围 纸。

评估的正则表达式时，另一个开销来从OT。为了满足性能要求，我们在[26]中使用了 OT 实现。此协议自举离的Diffie- Hellman密钥交换协议[28]，因此是非常有效的。我们评估了总端至端延迟时一个定期需要与真实世界的痕迹表达，如图中图5，包括网络延迟和所造成的延迟OT本身在MB和R.这是明确表示，延迟所造成的OT 增长缓慢，随着输入规模的增大，并且 这将是数量级的相同的顺序内的网络延迟，这意味着由开销造成OT本身是相当 有限。

恶意软件检测的准确性如何？为了回答 这个问题，我们首先应该找出n-gram的最佳*n*值是什么。此外，我们还需要确定多少

可以在平衡性能和开销（即计算和带宽开销）的同时从每个文件中进行选择的功能。图6显示了使用SVM进行恶意软件检测的准确性。准确度被定义为真阳性率，是衡量比例的被正确地识别为恶意软件阳性程序。 虽然，使用前500最常见的6克给出了最准确的分类结果，我们选择$k = 400$，因为它

		盲盒	水疗盒
关键字匹配	1个令牌，1条规则	27*纳秒*	103*纳秒*
	1个令牌，3K规则	162*纳秒*	114 *ns*
	1个关键字，3K规则	708 *ns*	499 *ns*
恶意软件检测	前400 5克	不适用	239个*μ小号*

表III：MB上的关键字匹配和机器学习微型基准，比较了SPABox和Blindbox

导致较小的通信开销，同时仍可实现较高的精度。因此，我们使用5个字节作为

对于滑动窗口标记化（见第IV-B为在两个不同的原因），选择顶部400的n-gram的特征 为在我们的实现中的恶意软件检测。由于我们使用的内核功能的限制，无法轻松提高检测精度。在第VIII-B节中讨论了如何启用其他内核功能和分类模型。什么是对吞吐量的一个客户端？ 根据我们的建议

加密 方法 需要 较长的 比 对 一个 在 Blindbox， 我们期望SPABox产生较低的吞吐量。图7显示了在发送方使用两个核心并禁用超线程进行加密的成本。这些数字包括时间为 标记化，令牌选择（恶意软件检测），散列与SHA-1，加密和传输过程。在5Mbps时，加密成本非常有限，因为CPU可以连续 生成数据。但是当吞吐量增加到23Mbps， CPU 在 该 发送 方 可以 没有 再 继续 了。 这开销

可以通过两种方式缓解：（i）令牌化，哈希和加密可以轻松地与额外的内核进一步并行化；

（ii）我们也可以 使用 基于 定界符的令牌 化代替 基于窗口的令牌化[43]，以减少生成的令牌的数量。然而，实现的吞吐量由我们的解决方案是不够典型的宽带家庭上行链路和应用程序处理小文件。R承担与S相似的成本，因为R必须检查S发送的加密令牌是否与从SSL恢复的明文匹配。

*2) 在*中间盒*方面：*我们的MB原型是在具有两个2.0GHz Xeon E5335内核和16GBRAM的服务器上实现的。该CPU不具备AES指令的支持。

如何 长时间 没有 它 采取 了 对 MB 到 匹配的关键字？怎么样时间拍摄到运行机器学习？ 为了 展示 我们方法的有效性，我们比较了SPABox和Blindbox的关键字 匹配性能。由于在缺乏AES硬件支持下，我们实现了MB上采用Intel酷睿i7和8GB内存为桌面的COM型坯与Blindbox。表 III显示了将Blindbox和SPABox与启用hpyer- threading进行比较的检测微型基准。对于第一种情况，其中MB需要以比较 所述 接收到的 加密的 令牌 在 所述SPA在27与一个关键字令牌连接，Blindbox可以完成该比较*纳秒*而SPABox增加值到103*纳秒*。大部分的开销来源于散列加密令牌，因为 我们首先需要每个令牌的转换 MPZ牛逼格式转换成一个字符串，然后散了。然而，对于3K规则（所有关键字均已标记化），SPABox只需114 *ns*即可匹配令牌，而Blindbox平均将使用162 *ns*。这是因为我们的实现使用哈希表而不是使用搜索树进行关键字搜索。为了匹配一个关键字，其具有

与Blindbox相比，SPABox平均可节省4个令牌。要执行分类，MB只需239 μs。显然，产生的开销很小。

在middlebox中存储THT的内存开销是多少？ 为了跟上搜索速度，我们使用哈希表 来代替 的 搜索 树 在 我们的 算法 为 关键字
匹配。每个关键字标记被散列并存储在该哈希表，使得MB可以迅速找出如果一个加密令牌在SPA 连接存在于该 规则列表。我们方法的一个可能缺陷是更多的 内存
用法。具体来说，如果使用搜索树，则总内存消耗约为0.822MB，而我们使用 的 哈希 表的内存开销为2-3倍。在我们的原型，我们测量的内存使用率 ∽3MB。即使使用哈希表增加 了 内存使用情况， 我们 认为 它 的 接受
权衡。此外，随着网络功能虚拟化（NFV）技术的出现[31]，中间盒可以被“虚拟化”并在具有“无限”内存的标准服务器上实现。因此，THT产生的内存开销可以忽略不计。

什么吞吐量 的 MB 维持？ 随着 我们的客户原型不能生成加密令牌足够快（最高到 23Mbps具备双核心），我们预先对数据进行加密，并使用它作为交通。在我们的实验中，我们测量的平均吞吐量
MB时为69Mbps。此数字包括SPABox支持的所有三个DPI功能。主要开销在该MB来自散列的所有加密标记处收到的来自S，同时寻找在THT和执行分类采取限制时间。如果一个专用的 核心 是 使用 了哈希 和搜索，吞吐量可提高到81Mbps。由于对所有加密令牌进行散列可以很容易地并行进行，因此可以进一步提高吞吐量。

3) 网络 开销:

建立连接会产生哪些开销？一个主要 的优势SPABox是，它不需要在连接设置客户端和MB之间的互动阶段。假设规则集 包含 3K 的关键字， 并且该客户端和MB之间的链路的吞吐量为20Mbps的（典型的家用宽带连接）。SPABox可以 完成40内设置 μ S于客户端（见表IV），而Blindbox要求客户端和之间漫长的相互作用的 MB。具体来说，在Blindbox中，每个关键字令牌的乱码为599KB。要 计算所有的乱码电路和发送它们，也需要超过90*分钟* ，并强加在客户端上巨大的计算量。手机或平板电脑将需要 一个 甚至 更长的 时间。 后 接收 所有 的 乱码 电路中，MB需要7.5*小号* 来 评估 他们 所有。 在 该相反，在SPABox，在MB需要从228字节信息的客户端 和MB可以17个内设置THT *MS*接收后的信息。与Blindbox相比，SPABox有更好的支持， 为 短暂的 连接 和 移动 用户。
总的带宽开销是多少？在整个 的实验中， 我们 看到 一个 带宽 开销 一 点点 少比
20次 原因有两个：*a*）我们使用基于窗口的标记化； *b*）只有非常小的开销因发生于正则表达式和恶意软件检测。虽然是实质性的 带宽 开销， 我们 可以 发现 这 对大

	盲盒		水疗盒	
	客户	中间盒	克林特	中间盒
时间	4670 *秒*	7.5*秒*	< 40 μ s	17*毫秒*
带宽	11.4 *GB*		228*字节*	

表IV：比较SPABox和Blindbox的连接设置开销微基准

开销的来源是令牌化。因此，它可以轻松缓解：我们可以不使用基于窗口的令牌化方法，而可以切换到基于定界符的令牌化方法，并且像Blindbox中那样仅令牌化非二进制数据。然而，基于分隔符-分词 可能 会失败 ，以 检测 攻击 的关键词 是 做 不启动之前或之后的分隔符和结束。人们可以选择任何一个方法，那是最好的一个特定的应用程序。

VII. [R心花怒放w ^ ORK

可搜索的 加密。 为了 在DPI场景中对加密数据执行 关键字匹配，人们自然会考虑可搜索的加密。已经有许多工作就因为它的搜索加密[18]，[20]，[44]的介绍。但是，将这些现有工作应用于DPI进行关键字匹配时，需要生成搜索令牌的实体来加密规则，这很可能会将敏感的规则 集信息公开给最终用户。而且，现有 方案均不能同时满足安全性和网络性能要求。具体来说，一方面，确定性搜索加密方案[18]泄漏模式的信息，即，两个单词是否（符合 一个 规则 或 在加密的流量不）是相同的，即使这些方案使得MB建立索引来处理每个令牌更快。这种薄弱的隐私保证使攻击者可以执行频率分析。在另一方面，随机方案[44]提供，因为更强的安全性的保证的随机盐的存在在其生成的密文。然而， 随机盐的使用防止了MB从构建索引结构进行快速令牌匹配，这导致在一个低 吞吐量 在 该 MB。

正则表达式评估。最近，已经有过关于开发中的协议，以一定的工作[42]，[48]使在加密的数据正则表达式搜索。一般来说，这些交互协议允许两方，以私下评估DFA在加密文件。阿非交互的情况下是所提出的功能性的加密方案的Waters [47]。该方案将密钥与特定的DFA绑定在一起，以便仅在DFA接受时，才能将其用于解密密文。 与密文关联的固定字符串。但是，这些方案只能支持正则表达式，而不能支持本文讨论的其他DPI功能。 而且，它们不能满足网络性能要求，特别是功能加密方案。

机器学习。一些工作[21]，[22]专注于对加密数据执行 ML。Bos等人的 工作[21]表明如何通过加密的医学数据医疗预测函数的计算 可以 被 执行 由一个第三方使用完全同态加密。博斯特等。

[22]构造了三个主要的隐私保护分类器，包括超平面决策，朴素贝叶斯和决策树。然而，他们的计划依赖于全同态加密（FHE），其结果在显著的开销，由于到一个事实，即现有FHE结构仍然是不实际的。结果，它们不能满足网络性能要求。此外，协议本身在[22]是相当复杂的，并且需要多轮的相互作用之间小号和该MB。

VIII. d藏密坛城艺术

A. *使用Paillier 密码系统的协议*

在我们的协议加密的令牌是有点类似于在Paillier密码[40]。在这里，我们指出，为什么它是不是一个好主意，以使用Paillier密码直接。

回想一下，在Paillier加密，密文是的形式$C \wedge = 克^{米} \cdot [R^{\tilde{N}}$ MOD $\tilde{N}^2$与*米*作为消息来加密。然而，直接使用Paillier加密在我们的应用场景中无法提供的MB具有的能力进行关键字匹配。准确地说，与仅*Ç*并没有$[R^{\tilde{n}}$MOD $\tilde{N}^2$，该MB不能执行关键字匹配操作过加密的令牌。不幸的是，如果MB可具有*Ç*和- $[R^{\tilde{N}}$MOD $\tilde{N}^2$在该相同的时间，其使得所述匹配操作成为可能，将得到的协议变为不安全，因为MB可以导出$克^{米}$经由除以*Ç*由$\check{r}^{\tilde{N}}$MOD $\tilde{N}^2$。因此，对有一个安全协议基于上Paillier加密（用于阐述，我们称之为这基于Paillier- 协议Π_P），一个直接的方法是引入一个新的随机元素*的盐*在$\check{z}_{\tilde{N}}$。现在所述加密的令牌中协议Π_P变得$Ç' = 克^{s_{一升吨,米}} \cdot [R^{\tilde{N}}$米直径：d $\tilde{N}^2$，和所述MB将RECE我v ë

Ç '，- $[R^{\tilde{N}}$ MOD $\tilde{N}^2$和$克^{盐}$MOD $\tilde{N}^2$关键字匹配操作过每令牌。同样观察到的是与$[R^{\tilde{n}}$

MOD $\tilde{N}^2$和$\tilde{N}^2$这是部分的r公共密钥，所述MB可以计算- $[R$ - $^{\tilde{n}}$米直径：d $\tilde{N}^2$和然后DER我v Ë$克^{s_{一升吨,米}}$米直径：d $\tilde{N}^2$。因此，$[R^{\tilde{N}}$ MOD $\tilde{N}^2$成为协议不必要Π_P为的MB执行关键字匹配。

Ñ øl瓦特协议Π_P加密一个给ķ烯作为$克^{s_{一个升吨,米}}$米直径：d $\tilde{N}^2$这是恰好等式 1.要执行的关键字匹配，所述MB瓦特乌尔德需要$克^{s_{一个升吨,米}}$米直径：d $\tilde{N}^2$和$克^{小号一升吨}$ m o d n^2。ħ öl瓦特év ë[R，该随机*盐*防止MB从执行高效搜索关键字匹配操作。具体而言，MB需要加密每个关键字，记为*瓦特*，在规则集通过计算（$克^{盐}$）^{瓦特}。MB必须将接收到的加密令牌与所有加密关键字进行比较，这些关键字的数量可能会令人生畏。为了提高MB的搜索性能，可以将*salt*定义为伪随机。此更改提供了我们建议的协议。而且，该方法进一步减少了通信开销，因为现在仅S需要对发射$克^{盐}$一次。

B. *恶意软件 检测扩展*

在我们的协议中，我们使用带有SVM的线性内核功能作为恶意软件检测的方法。通过我们的协议，也可以应用其他非线性内核功能。对于例如，考虑高斯内核k（X，X'）= EXP（- || X - x' ||²／2σ^2）。然后在等式中的决策函数。3可以是- [R é令状10作为$\breve{F}$（X）= Σ $\alpha_{我} Y_{我} k$（X，$X_{我}$）与那些$X_{我}$的该构成支持向量。以执行一个直接的方法本SVM模型涉及一个单轮相互作用之间的MB和R.在一个高的水平，所述MB第一安全单位计算（X - $X_{我}$）对于所有支持向量，分别，然后转发这些生成的结果，以R。然后R可以解密这些结果以获得（x - x_i），然后计算所有k（x- x_i）= exp（- || x - x' ||²／2σ^2），其将被加密并发送回给所述MB用于计算的最后步骤$\breve{F}$（X）= $\sum_{我} \alpha_{我} Y_{我} k$（X，$X_{我}$）。注意的是所有的的计算进行中的MB被执行在一个基于第IV-A节所述的同态性质的隐私保护方式。最后，MB发送加密的分类决定R.注意，此方法可以使在更多的代价更好的分类结果的开销在两个客户和该MB两侧。

另一个可能的扩展是使用其他ML模型代替SVM，例如朴素贝叶斯和决策树[22]。但是，这可能会导致更多的开销并失去MB的透明度。通常，它们将涉及MB和R之间的多轮交互。

C. *与Blindbox的比较*

尽管SPABox具有类似于Blindbox的体系结构，但是它使用不同的方法。首先，Blindbox使用对称加密方案作为基础，而SPABox使用公共密钥加密方案。尽管此选择降低了吞吐量，但可以支持更多操作（例如，机器学习），并且可以在我们的系统中极大地避免连接建立开销。为了弥补对在减少通说，我们建立一个多层次的哈希表，仔细设计我们的加密方案，以加快关键字查找在MB。其次，我们将正则表达式评估推给最终用户，而Blindbox通过解密流量在MB上进行。我们的设计可确保端到端对流量进行加密，并限制接收方的开销。请注意，MB在BlindBox和SPABox 设计中都是必需的，以保护规则集（否则最终用户将有权访问它），因此，我们的方法在R的性能与提供正则表达式评估功能和流量的目的之间进行权衡。/ rule同时设置隐私。

IX. çi面上污染物和°F UTURE W¨ ORK

长期以来，使用HTTPS是否会导致DPI死亡一直是争论的热点。在本文中，我们介绍了SPABox，这是第一个基于中间盒的系统，该系统在加密流量上支持基于关键字和基于数据分析的DPI功能，同时保证了中间盒中用户数据的私密性。最显着的特征SPABox是这样的协议设置并不要求之间的任何相互作用或数据传输一个中间盒和客户端。SPABox还可以定期维护隐私使用SVM进行表情评估和机器学习以检测恶意软件。SPABox的性能评估表明，它以有限的开销提供了保护隐私的DPI 。为了进一步提高性能，我们正在致力于保护隐私的正则表达式的评估方案，为我们今后的工作中，可以部分集成有我们的解决方案在为了以充分支持外包DPI。

R参考

[1] 数据平面开发套件。 http：// dpdk 。org。
[2] Google的密集哈希图。http：// goog-sparsehash 。sourceforge 。net / doc / 密集哈希 图。html。
[3] ICTF数据。 https：// ictf 。cs 。ucsb 。edu /。
[4] McAfee Network Security Platform。 http：// www 。mcafee 。com / us / products / network-security-platform 。aspx。
[5] Microsoft 恶意软件 分类 挑战 （BIG 2015）。https：// www 。kaggle 。com / c / malware-classification。
[6] NSA间谍活动依靠AT&T的“极端帮助意愿”。 https：// www。Propublica。或克/条/ nsa- SP应依靠-ON-atts- èxtreme-愿意到帮助。
[7] OpenSSL。 https：// www。openssl 。组织/。
[8] 克服有针对性的攻击：一种新方法。https： // blogs 。mcafee 。com / mcafee-labs / overcoming-targeted-attacks-new-approach /。
[9] Proofpoint与Intel McAfee。https：// www。检验点。com / sites / def ault /文件/文档/ bnt下载/ intel-mcafee-cg 1 。pdf。
[10] 使用Symantec Insight和SONAR防御高级威胁。 https：// www。赛门铁克。com / content / en / us / enterprise / whitespapers / protection-advanced-threats-insight-sonar-wp-21360326 。pdf。
[11] Snort v2.9。 https：// www。哼了一声。组织/下载。
[12] 赛门铁克在反恶意软件工具中添加了深度学习功能，以检测零时差。 http：// www 。电子周刊。com / security / symantec-deep-learning -anti-malware-tools-to-detect-zero-days 。html。
[13] 赛门铁克企业版。 http：// www 。赛门铁克。com / index 。jsp。
[14] VX天堂。 https：// vxheaven 。组织/。
[15] McAfee Labs 威胁 报告。 技术 报告， McAfee Labs， 2015年。
[16] J. Aycock。*计算机病毒和恶意软件*。施普林格， 2006年。
[17] M. Becchi和P. Crowley。用于实际深度包检查的混合有限自动机。在*ACM CoNEXT中*， 2007年。
[18] 贝拉雷先生，A。Boldyreva和A. O'Neill。确定性和有效的可搜索加密。在*CRYPTO中*。施普林格， 2007年。
[19] G. Berry和R. Sethi。从正则表达式到确定性自动机。 *《理论计算机科学》*，48：117–126，1986 。
[20] D. Boneh，G。Di Crescenzo，R。Ostrovsky和G. Persiano。公共密钥加密 用 的关键字 搜索。 在*Eurocrypt中*。施普林格，2004年。
[21] J. W. Bos，K。Lauter和M. Naehrig。对加密医疗数据的私人预测分析。*生物医学信息学杂志*，2014：50：234-243。
[22] R.博斯特，RA波帕，S.涂，和S.戈德瓦塞尔。机器学习分类 过 加密 的数据。*Crypto ePrint档案*，2014年。
[23] A. Bremler -巴尔，Y. Harchol，D.于草，和Y. KORAL。深度包检测 作为一个 服务。 在*ACM CoNEXT*，2014年
[24] D.现金，P.格鲁布斯，J.佩里和T. Ristenpart。针对可搜索加密的泄漏滥用攻击。在*ACM CCS中*， 2015年。
[25] C.-C. Chang和C.-J. 林 LIBSVM：支持向量机的库。 *ACM Transactions on Intelligent Systems and Technology*，2（3）：27，2011 。
[26] T. Chou和C. Orlandi。 遗忘传输 的最简单协议。在*国际会议上的密码学和信息安全中 的拉丁美洲*， 页 40-58。 施普林格， 2015年。
[27] T. Dierks和E. Rescorla。传输层安全性（TLS）协议版本 1.2。技术报告，IETF， 2008年。
[28] W. Diffie和ME Hellman。密码学的新方向。*Infor公司mation 理论，IEEE 交易 上*， 22（6）：644-654， 1976。
[29] T. ElGamal。公钥 密码体制 和 一个 签名方案基于离散对数。*IEEE Transactions on Information Theory*，31（4）：469-472，1985 。
[30] D. Ficara， S. 奴， G. Procissi，F. Vitucci， G. ANTICHI，和 A. Di Pietro。 一种 改进的DFA，用于快速的正则表达式匹配。*ACM SIGCOMM CCR*，38（5）：29-40，2008年。
[31] A. Gember-Jacobson，R。Viswanathan，C。Prakash，R。Grandl，J。Khalid ，
S. Das和A. Akella。OpenNF：启用网络 功能控制方面的创新。在*ACM SIGCOMM*， 2014年
[32] C.-W. Hsu C.-C. Chang C.-J. Lin等。支持向量分类的实用指南， 2003年。
[33] E. 科勒， R. 莫里斯， B. 陈， J. Jannotti， 和 M. F. Kaashoek。在单击模块化路由器。*ACM Transactions on Computer Systems*，18（3）：263–297，2000 。
[34] JZ Kolter和MA Maloof。学会在野外检测恶意可执行文件。在*ACM SIGKDD中*， 2004年。
[35] C. Lan，J。Sherry，RA Popa，S。Ratnasamy和Z. Liu。踏上：安全地外包 中间盒 到 了 云计算。 在*NSDI中*，2016年。
[36] KS McCurley。离散对数问题。在过程中。*的症状*。 1990年，*应用数学*， 第 42卷， 第 49-74页 。
[37] P. Mohassel，S。Niksefat，S。Sadeghian和B. Sadeghiyan。用于DFA评估和应用的有效协议。在*RSA会议的Cryptographers Track中*，第398-415页。施普林格，2012年。
[38] K. Namjoshi和G. Narlikar。强大而快速的模式匹配，可进行入侵检测。在*IEEE INFOCOM中*， 2010年。
[39] D.勒，K. Schomp，M. Varvello， I. Leontiadis，J.布莱克，D. R.大号d佩斯， K. P apagiannaki，P。罗德里格斯罗德里格斯，和P。Steenkiste。多上下文TLS（mcTLS）：在TLS中启用安全的网络内功能 。在*ACM SIGCOMM*，2015年。
[40] P. Paillier。基于复合度残差类的公钥密码系统。在*EUROCRYPT中*。施普林格， 1999年。
[41] K. Rieck，T。霍尔兹，C. W¯illems，P。d ü SSEL，和P。拉斯维加斯ķ Ø v。学习和分类恶意软件行为。在*“入侵和恶意软件的检测以及漏洞评估”中*，第108–125页。施普林格，2008年。
[42] M. A. 萨利希，T.考德威尔，A.费尔南德斯，E.密茨凯维奇，E. W.罗齐尔，S. Zonouz和D. Redberg。RESEED：对云中的加密数据进行正则表达式搜索。在*IEEE CLOUD* 2014。
[43] J. 雪利酒， C. 兰， R. A. 波帕， 和 S. Ratnasamy。 BlindBox： 深度 包检测 过 加密的 流量。 在*ACM SIGCOMM*，2015年。
[44] DX宋，D.瓦格纳，和A. Perrig。搜索加密数据的实用技术。在*IEEE小号&P*， 2000。
[45] SM Tabish，MZ Shafiq和M. Farooq。使用字节级文件内容的统计分析进行恶意软件检测。在*ACM SIGKDD网络安全和情报信息学研讨会上* ，2009年。
[46] Z. Wang， Z。Qian，Q。Xu，Z。Mao和M. Zhang。蜂窝网络中中间盒的一个不为人知的故事。在*ACM SIGCOMM CCR中*，第41卷，第374–385页， 2011年。
[47] B.沃特斯。常规语言的功能加密。在*CRYPTO中*。2012。
[48] L. Wei 和MK Reiter。迈向实用的加密电子邮件，以支持私有正则表达式搜索。*国际杂志的信息安全*，14（5）：397-416，2015年。
[49] A. Yao。如何以产生和交换秘密。在*IEEE FOCS*，1986年。
[50] Y. Ye，D。Wang ， T。Li和D. Ye。Imds：智能恶意软件检测系统。在*ACM SIGKDD中*， 2007年。

范靖远获得工程学士学位。和MS学位毕业于复旦大学，中国和大学的加利福尼亚州，美国洛杉矶，在2012年和2014年，分别。他目前正在攻读博士学位。纽约州立大学布法罗分校计算机科学与工程系获得博士学位。他的研究兴趣在于计算机 网络领域。

关超文获得学士学位。学位来自暨南大学，广州，2011年他在新加坡管理的研究工程师大学从2012年到2013年。现在，他是博士 学生在 纽约州立大学，布法罗。他的主要研究兴趣在于在密码学和信息安全领域。

厦人是计算机科学教授和工程与UbiSeC实验室主任，在新的州立大学纽约水牛城分校（UB）。他获得了博士学位程度从伍斯特理工学院。Kuibs目前的研究兴趣包括云和外包安全性，无线和可穿戴系统安全性以及移动传感和众筹。他的研究得到了NSF，DoE，AFRL，MSR和Amazon的支持。奎伊在同行评议的期刊和会议上发表了广泛的著作，并获得了若干最佳 论文奖包括
S 2017年和2011年ICNP他目前担任副主编的IEEE交易
，在服务计算IEEE交易，IEEE上交易的移动计算，IEEE 无线 通信， IEEE 互联网 的 物联网 杂志，和 一个 编辑 为 Spinger 内裤上的 网络 安全 系统 和 网络。Kui是IEEE的院士，IEEE的杰出讲师，ACM的成员以及Internet隐私的前任董事会成员。 工作 队， 国家 的 伊利诺伊州。

崔勇获得了工程学博士学位。度都来自清华大学。目前，他是一个专业人SOR 在 的 计算机 科学 系 在 清华。他是IETF WG的共同主席，曾在IEEE TPDS，IEEE TCC，IEEE Network和IEEE Internet Computing的编辑委员会任职或任职。他发表了100篇论文，并获得了多个最佳论文奖，并与人合着了约10篇Internet标准文档（RFC）。他的研究兴趣包括移动计算和网络体系结构。

刘春明乔是区分的SUNY亲 fessor，也是计算机科学与工程系，在现任主席大学 布法罗分校。他被 选举 到 IEEE 院士为他的光学和无线的捐款网络体系结构和协议。他目前的重点是联网和自动驾驶汽车。他已经有过的h指数著述颇丰69（根据谷歌学术）。两个他的论文已经从收到的最佳论文奖IEEE和联合ACM / IEEE 场所。 他 也 有 7美国
究已经十资助专利和作为顾问担任了几家IT和电信公司
，包括思科和谷歌，以及更多的十几个国家科学基金会资助。