doi:10.11959/j.issn.1000-436x.2015070

基于流感知的复杂网络应用识别模型

张洛什[1]，王大伟[2]，薛一波[3,4]

（1. 哈尔滨理工大学 计算机科学与技术学院，黑龙江 哈尔滨 150080；
2. 国家计算机网络应急技术处理协调中心，北京 100029；3. 清华大学 信息技术研究院，北京 100084；
4. 清华大学 信息科学与技术国家实验室，北京 100084）

摘 要：传统协议识别技术多以单网络流为识别手段，不能应对复杂网络应用多服务、多协议等特性，因此在面对复杂网络应用识别时严重失效。针对复杂网络应用的识别难题，提出了一种流感知模型，从空间、时间和流量 3 个维度来刻画复杂网络应用的通信特性，深度分析并挖掘了复杂网络应用的行为和状态特征；基于此模型，提出了一套快速识别复杂网络应用的方法和架构。实验结果表明，流感知模型能有效识别复杂网络应用，具有良好的识别效果。

关键词：协议识别；行为分析；流感知；复杂网络应用
中图分类号：TP393 **文献标识码**：A

Flow-awared identification model of sophisticated network application

ZHANG Luo-shi[1], WANG Da-wei[2], XUE Yi-bo[3, 4]

(1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China;
2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China;
3. Research Institute of Information and Technology, Tsinghua University, Beijing 100084, China;
4. National Lab for Information Science and Technology, Tsinghua University Beijing 100084, China)

Abstract: Traditional methods of protocol identification, which is mainly based on individual flow, lose their effectiveness as dealing with sophisticated network applications. A novel model of identifying sophisticated network applications, called flow-aware model, is addressed. This proposed model abstracts the characteristics of sophisticated network applications from spatial dimension, time dimension and flow dimension, and provides the detailed analysis and deeply mining in characteristics of behaviors and states. Based on this model, a framework and method of sophisticated network applications identification is proposed. The experimental results demonstrate that the proposed method can achieve the purpose of identifying sophisticated network applications effectively.

Key words: protocol identification; behavior analysis; flow aware; sophisticated network application

1 引言

随着网络技术的不断发展以及用户需求的不断变化，网络应用的服务模式也逐渐发生了改变：从提供单一功能的简单网络应用转变为同时提供多种功能的复杂网络应用。

复杂网络应用相比于简单网络应用而言，在一个网络应用中整合了多种不同的功能和业务，并广泛采用分布式技术和负载均衡技术实现服务的效率优化，采用多种网络协议保证数据传输的高效率和安全性。然而，复杂网络应用在提高用户便捷性的同时，也给网络运营商带来了严重的网络管理问题，尤其是针对复杂网络应用的协议识别问题。

传统协议识别方法的主要思路是通过对特定网络协议的分析，挖掘一个固定且区分能力较强的协议特征，并以此为基础对网络流量中的此类协议进行识别。然而，对于复杂网络应用而言，要寻找具有区分能力的协议特征尤为困难。一方面，复杂

收稿日期：2013-11-18；**修回日期**：2014-03-17
基金项目：国家科技支撑计划基金资助项目(2012BAH46B04)
Foundation Item: The National Science & Technology Pillar Program of China (2012BAH46B04)

网络应用的多业务模式带来了协议多样性、流量复杂性等特点，导致很难找到能覆盖特定复杂网络应用所使用的全部网络协议和全部流量的固定特征；另一方面，加密协议的广泛使用更进一步隐藏了可用在网络流量中识别特定应用的特征。所以，传统的协议识别方法在面对复杂网络应用识别时往往显得力不从心，甚至罕有效果。

为了有效识别网络流量中的特定复杂网络应用，本文提出了一种基于流感知的复杂网络应用识别模型。该模型从空间、时间和流量 3 个维度对复杂网络应用的通信流量进行感知，建立特定复杂网络应用的通信模型，并以此模型对网络流量中的特定复杂网络应用进行识别。在空间维度，描述了复杂网络应用的服务器节点部署的空间关联性；在时间维度，描述了复杂网络应用不同功能之间依据时间所进行的转换特性；在流量维度，描述了复杂网络应用不同功能所产生的多个网络流的行为特征。基于此模型，最终提出了一套快速分析和识别复杂网络应用的方法和架构。

本文的主要贡献如下。

1) 首次对复杂网络应用进行了定义，研究了复杂网络应用的通信特性，并分析了传统网络协议识别方法面对复杂网络应用识别时遇到的困难和原因。

2) 根据复杂网络应用的通信特性，在传统流量分析的基础上，提出了“流感知”模型，从 3 个维度对复杂网络应用的通信模式进行了建模。

3) 以“流感知”模型为基础，提出了一套识别复杂网络应用的框架和方法。

2　研究现状

在网络发展的早期，大多数协议遵循 IANA[1] 所颁布的协议端口号标准，并以此作为网络流量识别特征[2]。但是，随着协议数量增多以及随机端口的频繁使用，该方法在面对单流协议时就已经显得力不从心[3]，更无法应对复杂网络应用的识别。

因此，基于载荷的协议识别方法开始利用数据分组载荷中所包含的精确特征或正则表达式特征对协议进行识别[4,5]。但是，由于复杂网络应用所具有的多业务特性，无法利用一个特征对全部业务进行识别，因此，该方法通常仅能够识别复杂网络应用的部分流量，漏报比较严重。

近年来，出于数据安全的考虑，加密协议和私有协议被广泛使用[6]，这就使基于载荷的协议识别方法严重失效。Early 等[7]首次利用网络流的统计特征有效地区分了多种网络应用，并推动了基于流统计特征的协议识别方法的发展。该类方法把协议流量的统计特征作为识别特征，利用机器学习方法（如朴素贝叶斯、SVM、C4.5 等[8~10]）来对加密协议进行有效识别。但是，该类方法在面对复杂网络应用识别时也存在很大的局限性：一方面，机器学习方法很容易受到网络环境变化的影响，其准确性和可用性波动较大；另一方面，复杂网络应用的多业务特点会导致其统计特征的复杂化，很难保证机器学习算法的有效性。

为了有效应对复杂网络应用所带来的协议识别问题，Karagiannis 等[11]提出的 BLINC 方法，为后来基于行为的协议识别方法开启了先河。之后，Moore 等[12]开始利用复杂的分析方法对数据分组之间的顺序进行分析，以了解其所属协议的通信行为。Li 等[13]提出了 HMC 的方法对 P2P 协议进行有效识别。Xu[14]则使用信息论对协议的主机层流量行为进行了刻画。Jin[15]提出了 TAGs 方法和 mixed TAGs 方法[16]，利用图论对大规模网络环境下的协议通信模式进行了分析。Shi 等[17]引入了数据挖掘的概念，通过计算流与流之间的关联性对流进行分类。

近年来，针对复杂网络应用的识别主要集中在 2 个方面：一方面，尝试寻找固定不变的特征来识别特定复杂网络应用的全部网络流量；另一方面，则是对复杂网络应用的特定服务进行精细化识别。虽然在一定程度上解决了复杂网络应用的识别问题，但是其研究思路依然是针对单网络流进行分析，缺乏对全部网络流量的宏观掌控，识别结果有严重的误报及漏报，这正是本文要解决的问题。

3　“流感知”模型

3.1　复杂网络应用

用户需求的不断变化以及 IT 厂商的持续整合引发了采用多协议进行通信，并提供多服务和多业务的复杂网络应用的快速发展。

定义 1　复杂网络应用是指能够整合多种协议，并向用户提供多种业务的网络应用，具有多功能共存、多协议灵活切换、海量服务资源以及通信过程复杂等特点。

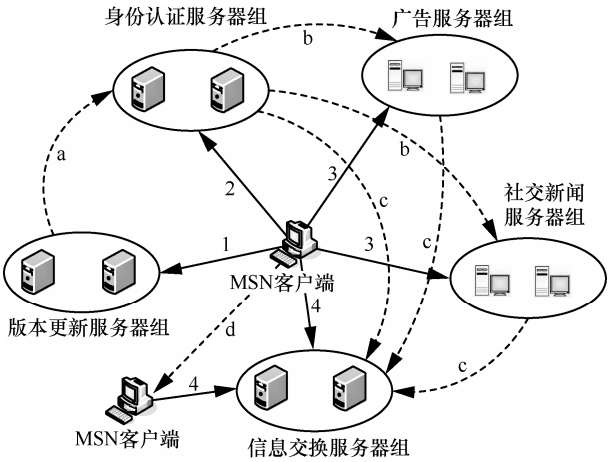


图 1 复杂网络应用 MSN 的通信过程

MSN 是一种典型的复杂网络应用，图 1 所示为其复杂通信过程，其通信过程具备以下 3 个方面的特性。

1) 空间分布特性。相同功能的服务资源所形成的服务器组部署在同一地域或同一个 IP 段，如图 1 中的版本更新服务器组等。

2) 时间关联特性。复杂网络应用运行时会按照固定顺序执行特定的功能（如图 1 中的 a、b、c、d），其执行顺序并不因用户和环境的改变而发生变化。

3) 流量聚类特性。复杂网络应用运行时会同时启用多个网络流，这些网络流使用不同协议、承载不同内容，按功能而划分，同一个功能的多个连续网络流的通信特性相对较为独立。

基于对多个复杂网络应用的详细分析，总结出复杂网络应用具有以下几个特点。

首先，功能融合所带来的协议复杂性。复杂网络应用中多个功能的整合导致了多协议并存的情况，无法利用同一个流量特征对复杂网络应用的多种协议进行识别。

其次，负载均衡和分布式处理带来的交互复杂性。负载均衡技术和分布式技术的使用在提高复杂网络应用服务效率的同时，增加了交互模式的复杂性，提高了识别的难度。

再次，加密或私有协议以及协议混淆技术所带来的流量复杂性。数据安全性的需求导致加密协议在复杂网络应用中被广泛使用。同时，为了逃避检测，一些特殊的复杂网络应用还会采用混淆、模糊、随机填充等技术对网络流量进行“整容”，防止被识别。这些技术消除了可供识别的固定特征，使传统协议识别技术失去了作用。

最后，版本频繁更新所带来的特征复杂性。复杂网络应用的频繁更新满足了用户需求的改变，但也导致协议特征的频繁变化，给特征提取带来了巨大的困难。

综上所述，复杂网络应用的出现虽然大大提高了使用便利性和用户体验，但却对协议识别方法提出了严重挑战。

3.2 “流感知”模型

流感知模型是指对复杂网络应用通信过程的抽象，将复杂网络应用所产生的全部网络流看作一个整体，通过时间、空间以及流量 3 个维度对聚合后的流量进行感知，最终描述复杂网络应用的通信本质。

空间维度描述了复杂网络应用通信过程中所使用的服务资源的空间分布特性，表示了复杂网络应用的空间行为特性；时间维度描述了复杂网络应用通信过程中不同功能之间的时间关联特性及行为变化规律，显示了复杂网络应用的时间行为特性；流量维度描述了复杂网络应用通信过程中的流量聚类特性，将同一功能中的多个网络流进行聚类，并按类提取统计特征，表示了复杂网络应用的流量行为特性。

如图 2 所示，3 个维度表征了复杂网络应用不同层次的通信行为，并存在互相支撑的关系。空间维度将网络流量规划在限定的范围内，减少了其他维度的处理范围；流量维度所表征的不同阶段的流量特征，是构成时间维度分析的基础；时间维度则是将流量维度的不同行为阶段在时间上进行关联，以此来识别不同的复杂网络应用。

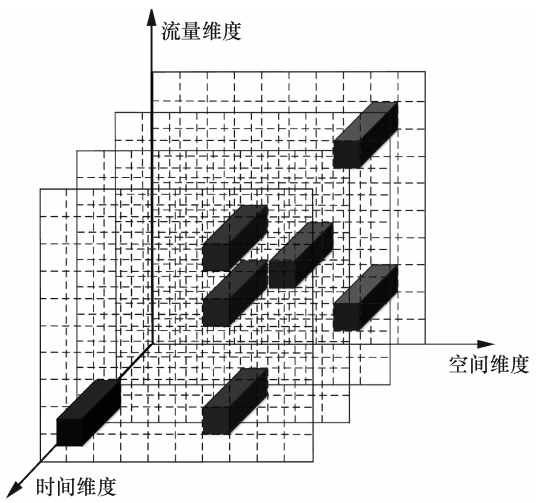


图 2 “流感知”模型

3.2.1　空间维度分析

复杂网络应用所具有的空间分布特性是空间维度分析的基础，采用关联思想可以对复杂网络应用所使用的服务资源进行分析和定位。

复杂网络应用通常包括 3 种基本关联策略。

1) 主机关联策略：同一个客户端在较短的时间间隔内所访问的多个服务主机可能隶属于同一个复杂网络应用。

2) IP 关联策略：相同功能的多台服务器的 IP 地址通常分配在同一个 C 段范围内。

3) 地域关联策略：由于设备部署原因，同一复杂网络应用的服务器主机节点的所在地理位置通常比较接近。

这 3 种关联策略从复杂网络应用的通信机制出发，描述了复杂网络应用服务资源的部署策略和原理，在面对具有大量服务资源的复杂网络应用时，有效地筛选了待识别的网络流量的范围。

图 3 描述了复杂网络应用空间维度分析的关联策略，其中，C_i（i=1,2,…,5）和 S_j（j=1,2,3,4）分别表示同一复杂网络应用的多个客户端和服务端。图 3（a）是在网关处获得的一个真实复杂网络应用的流量截图，显示了其通信过程和使用的服务资源；图 3（b）展示了主机关联策略，利用主机关联策略对真实流量中的客户端和其所访问的服务器进行关联分析后，可以得出如图 3（c）所示的服务资源隐含关系，其中，S_1、S_3 以及 S_2、S_4 分属于 2 个不同的服务器组，2 个服务器组之间存在时间上的先后关系，只有客户端访问 S_1 或 S_3 服务器之后才访问 S_2 或 S_4 服务器。

3.2.2　时间维度分析

复杂网络应用通信时的功能执行顺序可以由如图 4 所示的有限状态机进行表示（以 MSN 为例），描述了在一定时间段内复杂网络应用的通信行为转移模式。

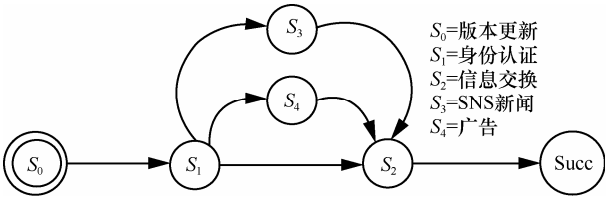
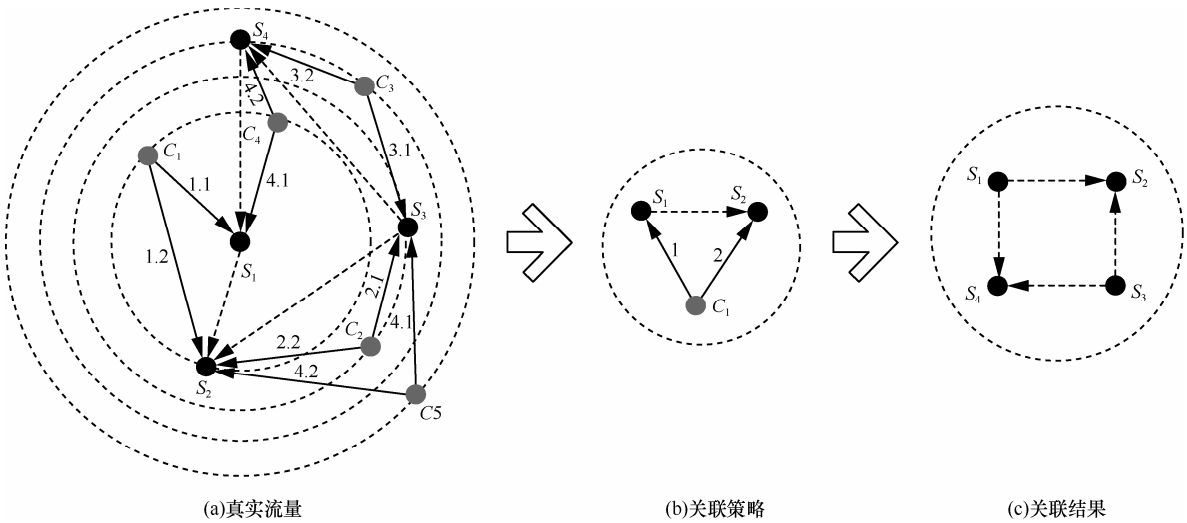


图 4　MSN 状态转移

然而，对于网络流量的观察者而言，其通信过程所代表的有限状态机位于程序内部，其状态及状态间的转移特征均被屏蔽，只能观察到状态执行过程中所产生的网络流量。因此，复杂网络应用的通信过程类似于隐马尔可夫模型（HMM）。其包含 2 个部分：一个是具有一定状态数目的马尔可夫链，使用状态转移概率描述状态之间的转移特性；另一个是描述状态和观测值之间对应关系的一般随机过程。其中，可见的观测值序列是指可观察到的多个网络流量的统计特征集合，而不可见的状态序列是指复杂网络应用通信时所执行的多个连续的操作序列，即运行在程序内部的一个有限状态机。

假设复杂网络应用在正常启动时会依次完成 N 个行为，即 N 个状态，表示为 S={$s_1, s_2,\cdots, s_N$}，状态之间并不独立，当前状态通常与上一个状态有关。因此，可将 S 看作 HMM 的隐状态集合，状态间的转移概率可由复杂网络应用通信过程所形成



(a)真实流量　　(b)关联策略　　(c)关联结果

图 3　复杂网络应用空间维度分析

的有限状态机 M 进行描述，处于状态 s_i 时的观测值为此状态下所产生的网络流簇的统计特征向量。当网络通信正常时，同一个复杂网络应用的 HMM 中的状态及对应的网络流簇统计特征是相同的，由通信过程所对应的有限状态机和网络流簇到达顺序唯一确定。

同一复杂网络应用其所要完成的功能和传递的信息具有一定的相似性。但是，不同的复杂网络应用因为其设计理念和功能的差异性导致其行为模型或传递内容之间存在较大不同，利用这种行为模式的相似性及差异性，可以对复杂网络应用进行建模，并利用它对应用进行识别。

3.2.3 流量维度分析

复杂网络应用的同一通信行为所产生的多个网络流之间存在互相协作的关系，形成具有固定行为统计特征的网络流簇。

对于网络流簇中的 TCP 流量而言，根据其使用情况可分为 3 个情况。

1) 连接失败的 TCP 流（T_F）。通常表示所连接的服务器不可达，即客户端所发送的 SYN 数据分组无应答。

2) TCP 控制流（T_C）。基于实验观察，本文将少于 15 个数据分组的 TCP 流称为 TCP 控制流。通常情况下，复杂网络应用使用 TCP 控制流进行控制命令和关键信息的传输。

3) TCP 数据流（T_D）。同样基于实验观察，本文将多于15个数据分组的TCP流称为TCP数据流。通常情况下，复杂网络应用使用 TCP 数据流传输应用数据。

对于网络流簇中的 UDP 流量而言，可以按照功能将其分为 2 个情况。

1) DNS 流（U_{DNS}）。复杂网络应用经常使用 DNS 协议获得服务器 IP 地址，并以此来实现负载均衡。因此，DNS 流通常代表了复杂网络应用一个行为阶段的开始。

2) UDP 数据流（U_{Data}）。复杂网络应用通常在传输大量数据时使用 UDP 协议以便提高传输效率。

综合上述 2 种协议特征，本文定义了复杂网络应用不同行为阶段的网络流簇特征，如式(1)所示。

$$<T_F, T_C, T_D, U_{DNS}, U_{Data}> \quad (1)$$

复杂网络应用的流量维度分析从流量之间的关联性出发，关注流量的交互行为，从整体的角度考虑了复杂网络应用的行为特性。

3.3 基于流感知模型的复杂网络应用识别

利用流感知模型可以为特定的复杂网络应用构建一个特有的行为模型，并以此为基础对网络流量中的特定复杂网络应用进行识别。

图 5 展示了基于流感知模型的复杂网络应用识别过程，主要分为 4 个步骤：基于空间维度的流量过滤，基于流量维度的网络流簇统计特征提取，基于时间维度的行为状态序列识别，以及基于决策树的复杂网络应用识别。

3.3.1 基于空间维度的流量过滤

基于空间维度的流量过滤分为先验知识构造和流量过滤 2 个阶段。

先验知识构造阶段主要是在可控网络环境下获取纯净的特定复杂网络应用流量，并提取全部服务器 IP 地址，以此来构建特定复杂网络应用的服务资源列表。

流量过滤阶段主要采用空间维度分析的关联策略，对同一个客户端 IP 所发送或接收的全部网络流量进行过滤，剔除不需要识别的网络流量。

为了更有效地对多个网络流进行分析，本文在网络流表的基础上引入了“DTable”数据结构，增加了 2 个散列表，分别记录网关内部 IP 节点流量信息（Client 表）和网关外部 IP 节点流量信息（Server 表），以 IP 为聚合点存储网络内外的多条网络流，并与流表之间存在关联。图 6 展示了“DTable”的结构。

如图 6 所示，客户端 C_1 连续访问了多个服务器 $S_i(i = 1,2,\cdots,7)$，产生了多条网络流 $f_i(i = 1,2,\cdots,7)$。其中，S_1 服务器是事先获得的已知服务资源，以 S_1 为基础，通过时间关联策略，将 C_1 访问 S_2 服务器的网络流量 f_2 与 f_1 进行关联；通过地域关联策略，将 C_1 访问 S_3 与 S_6 服务器的网络流量 f_3、f_6 与 f_1 进行关联；通过 IP 关联策略，将 C_1 访问 S_4 与 S_7 服务器的网络流量 f_4、f_7 与 f_1 进行关联。其中，S_5 服务器由于不符合 S_1 服务器的 IP 关联策略以及地域关联策略，并且网络流量 f_5 与 f_1 之间的时间间隔较长，因此，网络流量 f_5 和其他网络流不属于同一个网络流簇。

通过这种方式，可有效地过滤和整合同一个复杂网络应用所产生的网络流量，为后续的识别奠定了基础。

3.3.2 基于流量维度的网络流簇统计特征提取

本文收集在非连续的 2 次 U_{DNS} 类型的网络流之间的多个网络流，将同种类型的网络流进行聚

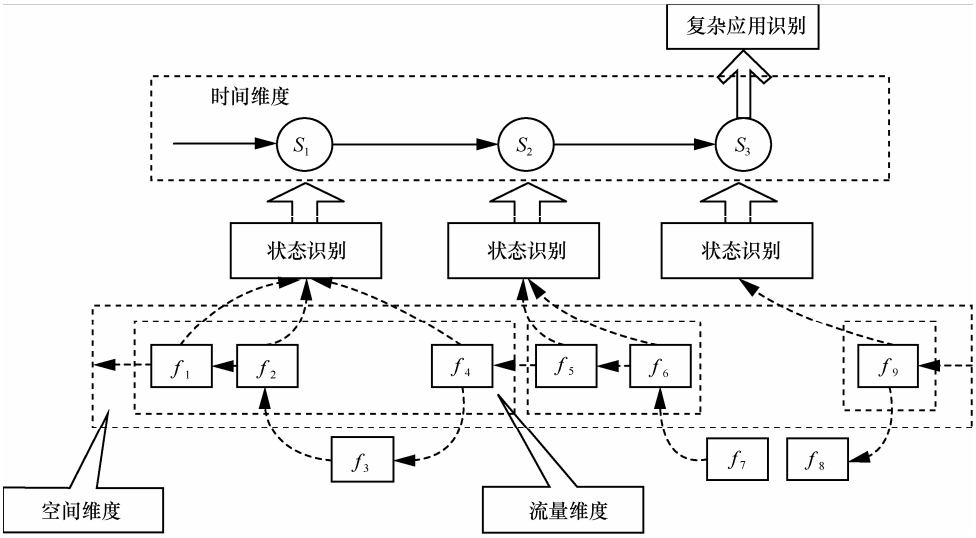


图 5　基于流感知模型的复杂网络应用识别示意

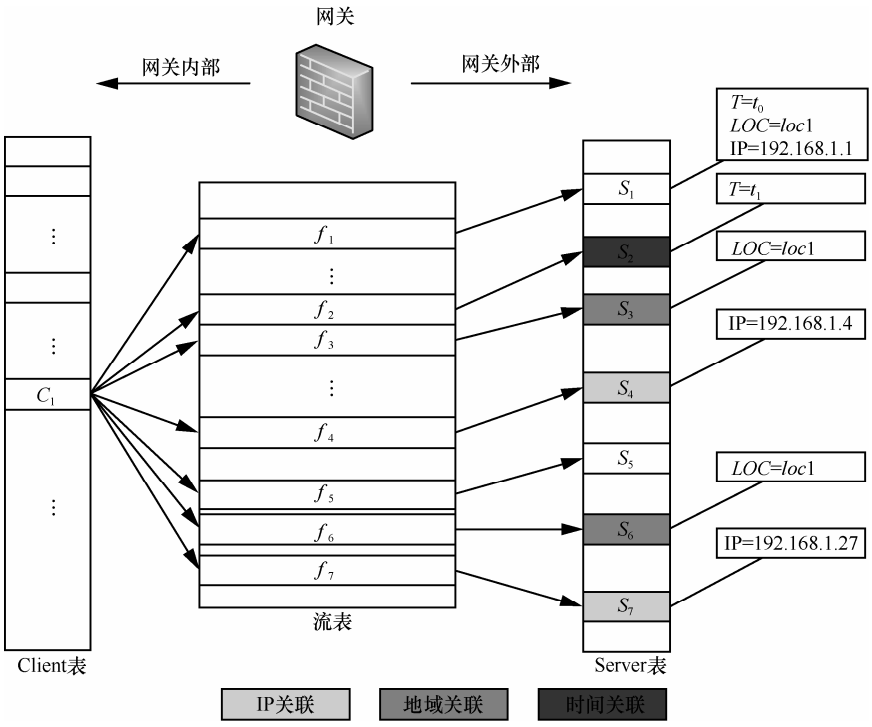


图 6 “DTable”结构图

合，并提取这种类型网络流簇的整体统计特征。

特征如表 1 所示。

表 1　每种协议类型网络流量的统计特征

编号	特征名称	特征描述
1	Ip_num	同一个内网 IP 所连接的外网 IP 数
2	Flow_num	同一个内网 IP 收发的网络流数
3	Packets_num	同一个内网 IP 收发的数据分组数
4	Bytes_num	同一个内网 IP 收发的字节数

这样，每种类型网络流的统计特征为一个四维向量

<Ip_num, Flow_num, Packets_num, Bytes_num>　(2)

由于 T_F 和 U_{DNS} 类型的网络流包含的内容较少，因此，不考虑这 2 种流量所带来的统计特征变化，由此，复杂网络应用每一个行为阶段可得到一个 12 维的统计特征向量作为隐马尔可夫模型中的观察值序列。

3.3.3 基于时间维度的行为状态序列识别

在得到多个连续网络流簇的统计特征向量作为观察值序列 O 后，通过隐马尔可夫模型可以得到对应的复杂网络应用通信过程的行为状态序列 Q。

利用隐马尔可夫模型（HMM）对复杂网络应用通信过程进行识别时，其参数被表示为

$$\lambda=(A,B,\Pi) \tag{3}$$

其中，

1) 通信行为状态为

$$\boldsymbol{S}=\{S_1,S_2,\cdots,S_N\} \tag{4}$$

2) 不同的观察值向量为

$$\boldsymbol{V}=\{v_1,v_2,\cdots,v_M\} \tag{5}$$

3) A：状态转移概率

$$A=\left[a_{ij}\right],\ a_{ij}\equiv P\left(q_{t+1}=S_j\mid q_t=S_i\right) \tag{6}$$

其中，a_{ij} 表示复杂网络应用在一个完整通信过程中，从行为 S_i 跳转到 S_j 的概率。

4) B：观测概率

$$B=\left[b_j(m)\right],\ \text{其中}\ b_j(m)\equiv P\left(O=v_m\mid q_t=S_j\right) \tag{7}$$

5) 初始状态概率

$$\Pi=[\pi_i],\ \text{其中}\ \pi_i\equiv P\left(q_1=S_i\right) \tag{8}$$

其流程共包括 2 个阶段。

1) 学习模型参数

该阶段主要目标是给定观测序列组成的训练集以便得到最优的隐马尔可夫模型。

假设在可控网络下获得的训练集合为 T，包括 K 条不同训练集数据，则 $T=\{O^{(k)},k=1,2,\cdots,K\}$，同时假设全部训练集均符合同一个隐马尔可夫模型且相互独立。

本文选择了隐马尔可夫模型中的前向后向算法（Baum-Welch 算法）来训练模型参数[18,19]，并最终得到特定复杂网络应用的隐马尔可夫模型。

2) 寻找状态序列

寻找状态序列是指利用确定的模型参数和获得的网络流簇的统计特征向量作为观察序列，递归计算其对应并隐藏的复杂网络应用的通信行为状态序列。

$$\delta_t(i)\equiv\max_{q_1q_2\cdots q_{t-1}}P\left(q_1q_2\cdots q_{t-1},q_t=S_i,O_1\cdots O_t\mid\lambda\right) \tag{9}$$

本文选择维特比（Viterbi）算法[18]计算得到最终的行为状态序列。

3.3.4 基于决策树的复杂网络应用识别

以所得到的通信状态行为序列作为统计特征，本文使用 C4.5 决策树算法[20]训练识别模型，并采用此算法进一步从背景流量中识别复杂网络应用所产生的网络流量。

为了提高复杂网络应用识别的效率以及适应性，通过对多个复杂网络应用的分析，本文设定行为状态最大长度为 5，即决策树的统计特征为 5 维，并认定任何超过此阈值的状态数量，均表示此复杂网络应用已经被关闭。

利用流感知模型对复杂网络应用进行识别时，并不是以单一维度特征作为识别条件，而是将 3 个维度的信息进行有效融合，逐个维度地对复杂网络应用的通信过程进行判断，并将其结果进行整合和抽象后进一步识别特定复杂网络应用所产生的网络流量。

4 实验分析

4.1 空间维度行为分析实验

图 7 展示了在某校园网络出口所获得的网络流量中部分 Skype 服务资源的地域分布情况。实验结果表明，Skype 应用的服务资源在全球 96 个国家均有分布，但部署在美国、英国、法国以及俄罗斯的 Skype 服务资源已经超过其全部服务资源数量的 50%，并且这些服务资源被访问的次数和频率远高于其他地域。由此证明了流感知模型中空间维度分析所提出的地域关联策略的有效性。

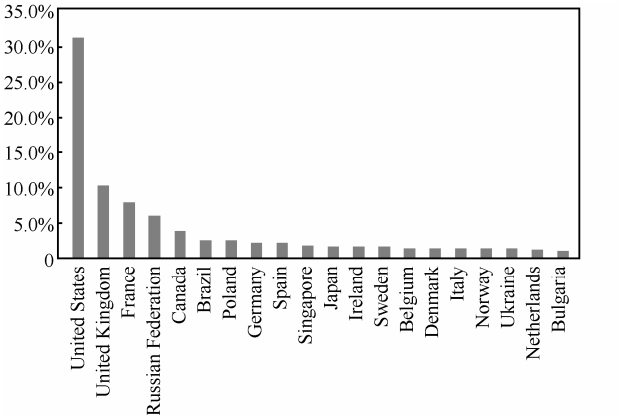


图 7 Skype 应用服务节点国家分布

图 8 所出为 Skype 使用最频繁的前 20 个 IP 地址 C 段。实验结果显示，其段内平均被使用 IP 数量为 23 个，IP 总量占全部服务资源总量的 15.14%，前 20 个 C 段内 IP 被访问次数占总访问量的 66.56%。因此，Skype 服务资源主要集中在特定的

C 段范围内，具有较大的集中性。

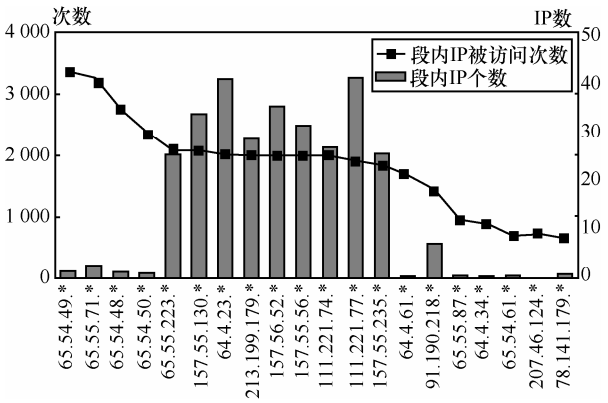


图 8　Skype 应用服务节点 C 段 Top20 分布

由于 Skype 使用了 P2P 技术，因此具备大量的用户自建 Peer 节点，这部分节点无论是地域还是 IP 地址 C 段均分布广泛，其平均被访问次数为 4.6 次，C 段所含 IP 数量平均为 1.01 个，不具备参考价值。

通过对 Skype 服务资源的分析后，发现复杂网络应用所使用的服务资源具有明显的集中性和关联性，验证了空间维度分析方法的有效性和可行性。

4.2　时间维度行为分析实验

图 4 和图 9 分别展示了 3 种复杂网络应用（MSN、Skype、eMule）在启动阶段的行为模式。在本文中，为了达到尽早识别的目的，仅对复杂网络应用的初始启动阶段的行为模式进行分析。

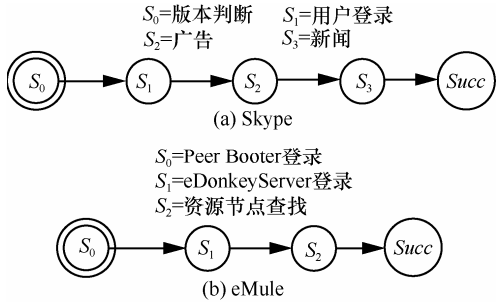


图 9　复杂网络应用启动行为转换

实验结果证明，不同复杂网络应用由于功能差异导致其行为模式大不相同，但同一复杂网络应用其行为模式较为固定。因此，有效地证明了利用时间维度可以区分不同的复杂网络应用的行为模式。

4.3　复杂网络应用识别性能

为了全面评估流感知模型对特定复杂网络应用的识别效果，本文选择了 3 种复杂网络应用（Skype、Thunder、PPS）进行验证。

在某校园网下的可控实验室环境中分别收集了 3 组数据集作为实验数据对象，如表 2 所示。同时，为了保证数据集的纯净，采用了有效的手段避免了操作系统所带来的其他网络流量。

表 2　　3 个数据集的统计信息

数据集		网络流数量	网络分组数量
训练集	Skype	207 532	1 287 402
	Thunder	721 232	16 933 250
	PPS	530 695	48 341 227
测试集	Skype	875 502	5 544 857
	Thunder	1 624 724	38 145 643
	PPS	1 890 933	1 392 988 874

每一组数据集被均分为训练集和测试集 2 个部分。训练集用来训练不同维度的检测模型，而测试集则用来对模型的准确性进行评估。

为了有效地评估流感知模型的有效性，本节选择了 2 个度量来描述其识别结果。

1) 召回率：表示检测出的正确网络流数占测试集中全部正确网络流的比率，衡量模型的查全率。

2) 准确率：表示检测出的正确网络流数占检测出的全部网络流数的比率，衡量模型的查准率。

同时，本文选择目前流行的检测工具 nDPI[21] 作为对比对象，其识别结果如表 3 所示。

表 3　　nDPI 识别实验结果

复杂网络应用	准确率	识别协议	分组百分比	流百分比
Skype	79.66%	Skype	64.75%	79.66%
		SSL	32.23%	16.93%
		DNS	2.18%	1.69%
Thunder	16.47%	BitTorrent	1.91%	16.47%
		DNS	0.62%	4.13%
		HTTP	4.33%	8.78%
		Unknown	92.35%	69.08%
		Other	0.79%	1.54%
PPS	23.32%	H232	0.11%	0.99%
		HTTP	7.53%	22.22%
		DNS	0.50%	20.15%
		Unknown	91.86%	56.64%

从表 3 中可以看出，nDPI 的识别结果包含了多种协议，印证了复杂网络应用的协议复杂性特点。但其识别准确率较低，除 Skype 外，其他 2 种应用的识别结果基本不可用，表明目前传统协议识别方法在面对复杂网络应用时已经力不从心。

流感知模型的识别结果如表 4 所示。

表 4 复杂网络应用识别实验结果

复杂网络应用	召回率	准确率
Skype	93.69%	96.01%
Thunder	95.73%	90.96%
PPS	92.82%	93.66%

实验结果表明，利用流感知模型对复杂网络应用进行识别，其准确率和召回率均可以达到 90%以上，说明其具有很好的检测效果，证明了流感知模型的有效性，可以为复杂网络应用的识别提供很好的基础和参考价值。

5 结束语

针对复杂网络应用的识别难题，本文提出了流感知模型，从空间、时间和流量 3 个维度对其行为特征进行感知，形成一个有效的行为模型。利用此模型对复杂网络应用进行识别，取得了很好的识别效果。该方法从根本上改进了协议识别的思路，扩充了协议识别的研究方向，能够有效地解决复杂网络应用的协议识别难题。

在下一步工作中，会进一步完善和优化流感知模型，并对更多的复杂网络应用进行测试，以攻克复杂网络应用的识别难题。

参考文献：

[1] IANA[EB/OL]. http://www.iana.org/.

[2] SEN S, SPATSCHECK O, WANG D. Accurate, scalable in network identification of P2P traffic using application signatures[A]. Proceedings of the 13th international conference on World Wide Web[C]. New York, USA, 2004.512-521.

[3] KARAGIANNIS T, BROIDO A, BROWNLEE N, *et al*. Is P2P dying or just hiding?[A]. Proceedings of the 47th annual IEEE Global Telecommunications Conference[C]. Dallas, USA, 2004 .1532-1538.

[4] HU C C, YI T, CHEN X F. *et al*. Per-flow queueing by dynamic queue sharing[A]. Proceedings of the 26th IEEE International Conference on Computer Communications[C]. Anchorage, Alaska, 2007. 1613-1621.

[5] SOMMER R, PAXSON V. Enhancing byte-levelnetwork intrusion detection signatures with context[A]. Proceedings of the 10th ACM Conference on Computer and Communications Decurity (CCS 2003)[C]. Chicago, USA, 2003. 262-271.

[6] SMTICH R, ESTAN C, JHA S. XFA: faster signaturematching with extended automata[A]. Proceedings of the 2008 IEEE Symposium on Security and Privacy (sp 2008)[C]. Oakland, USA, 2008. 187-201.

[7] JAMES E, CARLA B, CATHERINE Behavioral authentication of server flows[A]. Proceedings of the 19th Annual Computer Security Applications Conference[C]. 2003. 46–55.

[8] 王一鹏,云晓春,张永铮,李书豪. 基于主动学习和 SVM 方法的网络协议识别技术[J].通信学报,2013,34(10):135-142.
WANG Y P, YUN X C, ZHANG Y Z, LI S H. Network protocol identification based on activelearning and SVM algorithm[J]. Journal of Communications, 2013, 34(10): 135-142.

[9] AULD T, MOORE ANDREW W, STEPHEN F. Bayesian neural networks for Internet traffic classification[J]. IEEE Trans Neural Networks, 2007,18(1): 223–239.

[10] YANG B H, HOU G D, RUAN L Y, *et al*. SMILER: towards proactical online traffic classification[A]. Proceedings of the 2011 ACM/IEEE Seventh Symposium on Architectures for Networking and Communications Systems[C]. DC, USA, 2011. 178-188.

[11] THOMAS K, KONSTANTINA P, MICHALIS F. BLINC: Multilevel traffic classification in the dark[A]. Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications[C]. New York, USA, 2005.229-240.

[12] ANDREW M, KONSTANTINA P. Towards the accurate identification of network applications[A]. Proceedings of 6th International Workshop, PAM 2005[C]. New York, USA, 2005. 50-60.

[13] LI C L, XUE Y B, *et al*. HMC: A novel mechanism for identifying encrypted P2P thunder traffic[A]. Proceedings of Global Telecommunications Conference (GLOBECOM 2010)[C]. Miami, USA, 2010. 1-5.

[14] XU K, ZHANG Z L, BHATTACHARYYA S. Profiling Internet backbone traffic: behavior models and applications[A]. Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'05)[C]. New York, USA, 2005. 169-180.

[15] JIN Y, SHARAFUDDIN E, ZHANG Z L. Unveiling core network-wide communication patterns through application traffic activity graph decomposition[A]. Proceedings of the Eleventh International Joint Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '09)[C]. New York, NY, USA, 2009. 49-60.

[16] JIN Y, DUFFIELD N, *et al*. Can't see forest through the trees? understanding mixed network traffic graphs from application class distribution[A]. Proceedings of the 9th Workshop on Mining and Learning with Graphs (MLG 2011)[C]. San Diego, California, USA, 2011. 20-21.

[17] SHI X G, CHAU C K, CHIU D M. Space-efficient tracking of network-wide flow correlations[A]. Proceedings of INFOCOM'2011[C]. Shanghai, China, 2011.11-15.

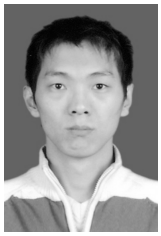
[18] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1889, 77(2) : 257-286.

[19] 谢柏林, 余顺争. 基于应用层协议分析的应用层实时主动防御系统[J]. 计算机学报. 2011, 34(3): 452-463.
XIE B L, YU S Z. Application layer real-time proactive defense system based on application layer protocol analysis[J]. Chinese Journal of Computers, 2011, 34(3): 452-463.

[20] KOTISANTISS.B. Supervised machine learning: a review of classification techniques[J]. Informatica, 2007,31(3): 249-268.

[21] nDPI[EB/OL]. http://www.ntop.org/products/ndpi/.

作者简介：



张洛什（1983-），男，陕西西安人，哈尔滨理工大学博士生，主要研究方向为网络安全、协议识别以及流量管理等。

王大伟（1982-），男，山东烟台人，国家计算机网络应急技术协调处理中心高级工程师，主要研究方向为计算机网络、信息安全、人工免疫。

薛一波（1967-），男，山东莱阳人，清华大学研究员，主要研究方向为计算机网络和信息安全、并行处理、分布式系统。