doi:10.3969/j.issn.1673-4785.2010.02.007

流特征的Skype流量识别

万月亮[1],朱贺军[2],刘宏志[1]
(1.北京工商大学 计算机与信息工程学院,北京 100048; 2. 公安部第三研究所 北京锐安科技有限公司,北京 100044)

摘 要:Skype流识别的研究大多局限于在静态载荷特征和通信机制,没有考虑网络流特征在Skype流量识别中的作用.提出了一种基于朴素贝叶斯分类的Skype流量识别模型.选择流的连接特征和实时特征作为分类特征集,根据流的连接特征组织网络流,再进一步根据流的包长度、平均发送间隔和突发带宽消耗等实时流特征识别Skype流量.在北京联通骨干网络上的实验表明该模型能有效地识别Skype流,是一种有效的Skype流识别算法.

关键词:流量识别;朴素贝叶斯分类;深度包检测;实时流特征
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-4785(2010)02-0139-05

Skype traffic identification based on flow characteristics

WAN Yue-liang[1], ZHU He-jun[2], LIU Hong-zhi[1]
(1. College of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;
2. The Third Research Institute of Ministry of Public Security Run Technologies Co., Ltd., Beijing 100044, China)

Abstract: Most of the Skype traffic identification models are limited to Skype communication mechanisms and static payload characteristics. No net flow characteristics are considered in identification algorithms. To overcome this limitation, a hierarchical Skype traffic identification model based on naive Bayesian classification was developed. Flows were analyzed according to network connection modes. Results were then obtained according to real-time flow characteristics, such as packet size, average inter-packet gap and burstiness of bandwidth consumption. The validity of the algorithm was proven by testing conducted on the Beijing China Unicom backbone network.

Keywords: traffic identification; naive Bayesian classification; deep packet inspection; real-time flow characteristic

Skype流识别是网络策略计费和差异化服务的重要前提之一.由于Skype采用私有通信协议通信,加密了用户之间以及用户与Skype服务器之间的通信内容,使得基于端口和特征的检测方法难以有效识别Skype流量.Skype流识别研究大多集中于静态载荷特征和通信机制的研究.文献[1]研究了Skype网络拓扑结构和关键节点信息,得到了Skype节点注册和登录通信过程及流量特征.文献[2]采用逆向工程技术分析了Skype协议实现细节,指出Skype所有通信内容都是加密传输.文献[3]分析了Skype使用的通讯协议,指出在呼叫建立阶段采用TCP协议;语音传输阶段则多采用UDP协议,最后给出了部分Skype协议的静态载荷特征码.在此基础上,文献[4]提出了基于UDP端口的Skype流量识别方法,这种方法适用于Skype使用UDP作为传输层协议,需要得到Skype注册信息.文献[5]通过网络端节点入站和出站流的特征,结合P2P特征识别Skype流量,其前提是端点所处网络拓扑已知,能够获取单端节点流量.然而现实网络环境难以满足这些条件,从而限制了该方法的实用性.文献[6]指出Skype流承载有

收稿日期:2009-12-27.
通信作者:万月亮. E-mail:yueliang_wan@126.com.

2种不同模式传输VoIP应用数据:一种是端到端方式(end to end,E2E),2个端节点之间传输VoIP数据;另一种是端到电话方式(end to phone,E2P),端节点到传统PSTN电话之间传输VoIP数据.该文献采用Chi-Square分类器识别网络流量中加密流量,再采用贝叶斯分类器结合实时流的特征识别Skype流,算法没有考虑Skype的P2P特征,实验网络环境流量较小,实验结果难以充分说明模型.文献[7]指出P2P通信存在以下2个连接特点:{TCP/UDP}IP对.{IP,Port}对特征.并且通过对UDP数据包目的地址和目的端口统计特征识别P2P应用流.朴素贝叶斯分类(naive Bayesian classification,NBC)在数据挖掘领域取得了很好的效果[8].

1 基于流特征的朴素贝叶斯分类

Skype将语音、视频、聊天和数据复用在一个帧中,再附加各自的称为开始消息(start of message,SoM)的非加密头部后,最后经压缩和加密后封装成UDP或TCP数据包.作为分类特征,这里选择Skype与其他互联网流有明显差异的特征作为分类特征.一方面,作为一个采用P2P架构的应用软件,Skype产生的流具有P2P数据流的特征,另一方面,Skype承载的业务是VoIP应用,使得Skype流具有实时流的特征.

1.1 Skype分类特征选择

Skype是采用基于P2P架构的VoIP系统,其通讯过程具有P2P通信特点,同时其流具有实时流特征.

1.1.1 Skype连接特征

通过分析可知,Skype端节点对间通信仅使用一种协议(或者采用UDP协议或者采用TCP协议),而不同时采用2种传输层协议.Skype通信过程不符合P2P通信的{TCP/UDP}IP对的特征,但是Skype通讯过程中的目的地址和目的端口连接特征符合{IP,Port}对特征.通过式(1)统计数据流的UDP目的地址和目的端口特征,判断数据流是否符合P2P通信特征.

$$(\mathrm{IP}_d,\mathrm{Port}_d)\in\{(\mathrm{IP}_s,\mathrm{Port}_s,\mathrm{IP}_{d_1},\mathrm{Port}_{d_1}),\cdots,(\mathrm{IP}_s,\mathrm{Port}_s,\mathrm{IP}_{d_n},\mathrm{Port}_{d_n})\}.$$
$$\|\,\|\mathrm{IP}_d\|-\|\mathrm{Port}_d\|\,\|<P_T. \tag{1}$$

式中:$(\mathrm{IP}_d,\mathrm{Port}_d)$为属于同一源地址/源端口的网络流量的目的IP地址和目的端口列表,$\|\mathrm{IP}_d\|$为目的地址数量,$\|\mathrm{Port}_d\|$为目的端口数量,P_T是目的IP地址数$\|\mathrm{IP}_d\|$和目的端口数$\|\mathrm{Port}_d\|$差的阈值.

1.1.2 Skype流特征

Skype承载的业务多是VoIP业务流,为了达到VoIP业务的持续和无延迟实时效果,Skype采用UDP协议传输VoIP业务流,这使得Skype产生的业务流数据较传统C/S非实时数据流小,且数据包发送间隔较小.本文采用数据包大小、平均发送间隔和突发带宽消耗作为实时流特征来度量实时流.

1)数据包大小:Skype语音数据大小分布具有自相似性,主要集中在120 Byte,其中50~150 Byte之间数量占大多数.这个范围外的数据包发生呼叫建立时候,属于信令消息.Skype数据包大小如式(2)所示:

$$x=[s_1,s_2,\cdots,s_w]. \tag{2}$$

式中:s_i为连续w数据包中第i个数据包的大小.

2)平均发送间隔:平均发送间隔定义为连续w帧的发送时间$1/w$,如式(3)所示.这个发送间隔在0.02~0.04 s之间,大于0.04 s数据包推测属于信令消息.

$$y=[T]=[(t_w-t_s)/w]. \tag{3}$$

实验发现,当$w>10$基本可消除偶然因素影响.本文w选择45,大约1.5 s.

3)突发带宽消耗:突发带宽消耗有字节率和数据包率2种,而无论字节速率还是数据包速率Skype流在经过开始几秒钟的上升阶段后都保持相当的稳定,这里突发带宽消耗为规定时间间隔T内字节突发带宽消耗,如式(4)所示.

$$z=[\mathrm{BW}]=\sum_{i=1}^{n}\mathrm{bw}_i\mid_T. \tag{4}$$

式中:T取值为1 s,大约Skype的30个数据包带宽消耗.

1.2 朴素贝叶斯分类

朴素贝叶斯方法以概率密度函数为基础,描述分类系统中条件属性和分类属性之间的映射关系,相比于其他算法,具有出错率最小的特点.算法假定可以通过观察的特征量集来描述目标,这些不同特征可形成各自聚类.$\boldsymbol{X}=[X_i]$表示不同样本向量,给定特征序列$\boldsymbol{X}$,$P\{C\mid x\}$表示样本x属于C类的概

率,由先验概率 $P\{x \mid C\}$ 得出后验概率 $P\{C \mid x\}$ 如式(5) 所示:

$$P\{C \mid x\} = \frac{P\{C,x\}}{P\{x\}} = \frac{P\{C,x\}}{P\{x\}} \cdot \frac{P\{C\}}{P\{C\}} = \frac{P\{x \mid C\}}{P\{x\}} P\{C\}. \tag{5}$$

各分量 x_i 相互独立,满足式(6):

$$P\{x \mid C\} = \prod_i P\{x_i \mid C\}. \tag{6}$$

通常情况下,采用最大可能性判定准则评估样本属于 C 类的可能性,$P\{x \mid C\}$ 称为置信度,置信度越大,属于该类的可能性也就越大. 这里采用前面小节定义的实时流量特征作为分类特征集. 其中数据包大小、平均发送间隔和平均带宽消耗置信度分别定义为 B_x、B_y 和 B_z,如式(7):

$$\begin{aligned} B_x(C) &= \frac{1}{w}\sum_{i=1}^{w} \lg P\{x_i \mid C\}, \\ B_y(C) &= \lg P\{y \mid C\}, \\ B_z(C) &= \lg P\{z \mid C\}. \end{aligned} \tag{7}$$

其中,数据包大小由观测值采用高斯分布进行拟合. 其与发送速率 RT,头部长度 H,冗余参数 RF 和编帧时间 $N(u,\sigma)$ 相关. 对于每个 $\{RT,H,RF,\Delta T\}$,数据包大小分布用高斯分布 $N(u,a)$ 表示,其中 u 如式(8) 表示:

$$u = (\text{Rate}\,\Delta + \text{len}(H))RF + \text{len}(SoM). \tag{8}$$

式中:RT 由表 1 所示,$RF \in \{1,2,3,4\}$,$\text{len}(SoM) = 4$ Bytes,$\text{len}(H) = 8$ Bytes,$\Delta T \in \{10,20,\cdots,60\}$,单位为 ms.

表 1　Skype 解码器的特性

Table 1　Skype decoder characteristics

编码器	帧尺寸/ms	比特率/Kbps
ISAC	30,60	10/32
ILBC	20,30	13.3,15.2
G.729	10	8
iPCM-wb	10,20,30,40	80
EG.711A/U	10,20,30,40	48,56,64
PCM A/U	10,20,30,40	64

σ 的取值,对于恒定速率编码,如 G729,σ 取值为 1;对于可变编码如 ISAC,σ 取值为 0.75. 每个实验窗口 k 选择所有置信度中最大的 $B_x^{(k,j)}(c)$ 作为编码器 j 新的置信度,如式(9):

$$E[B_x^{(j)}] = E_k[B_x^{(k,j)}]. \tag{9}$$

式中:$B_x^{(k,j)}$ 为编码器 j 时间序列 k 的置信度,数据包大小特征的置信度如式(10) 所示:

$$\max B_x = \max(E[B_x^{(j)}]). \tag{10}$$

平均发送间隔和带宽消耗不需要考虑数据包大小中的不同编码器因素,窗口 k 置信度分别为 $B_y^{(k)}$ 和 $B_z^{(k)}$,如式(11)、(12) 所示:

$$E[B_y] = E_k[B_y^{(k)}], \tag{11}$$

$$E[B_z] = E_k[B_z^{(k)}]. \tag{12}$$

Skype 实时流量特征置信度由数据包大小、平均发送间隔和带宽消耗特征的置信度中最小值决定,如式(13)所示.

$$B = \min(\max B_x, E[B_y], E[B_z]), \quad B > B_T. \tag{13}$$

式中:B_T 为 Skype 实时流特征置信度阈值.

Skype 流的判定由流的连接特征和实时流置信度决定,满足式(1)和式(13)的网络流量识别为 Skype 流. 也就是说,如果网络数据流满足 P2P 的连接模式,且具有实时流特征,则该数据流认为是 Skype 流.

2　试验结果及分析

实验数据来自北京联通 2 路 OC192 10G 数据,实验时间为 2009 年 4 月 10 号 –2009 年 4 月 11 号,总共 20 h. 实验的数据中不仅包含 Skype 流,还有 TCP 协议数据、TRP/UDP 协议传输的视频数据、P2P 数据、VPN 数据、VoIP 数据.

测试数据集中的 Skype 数据流采用简单载荷识别和组合特征识别方法采捕获,包括 E2P 语音流和 E2E 语音、视频、数据和聊天数据流.

实验的硬件环境为:机架式服务器,4 GB 内存,Xeon E5530 4 核 2.4 GHz 处理器,操作系统为 Linux. 按照错检率(false-Positive,F_P)和漏检率(false-negative,F_N)来衡量识别模型性能,如式(14).

$$F_P = \frac{N_{SN}}{N_{SS}}, \quad F_N = \frac{N_{NS}}{N_S}. \tag{14}$$

式中:N_{SS} 为测试集中属于 Skype 的流识别为 Skype 的数量,N_{SN} 为测试集中不属于 Skype 的流识别为 Skype 的数量,N_S 为测试数据集中的 Skype 流数量,

N_{NS} 属于测试数据集中 Skype 流,但未被识别的 Skype 流数量.

由于 Skype 有效载荷的 SoM 消息的数据包识别需要持续跟踪 Skype 有效载荷 SoM 消息及源节点产生的所有流量. 在真实网络环境下很难满足这些条件. 但是基于有效载荷特征识别结果可靠,本文采用其作为离线交叉验证方法,选用文献[6]基于有效载荷特征识别的流作为基准测试集,来验证 NBC 分类模型.

图1 比较了 NBC 模型和文献[5]提出的 NPA 算法的性能,表2 是 NBC 和 NPA 分类性能详细数据.

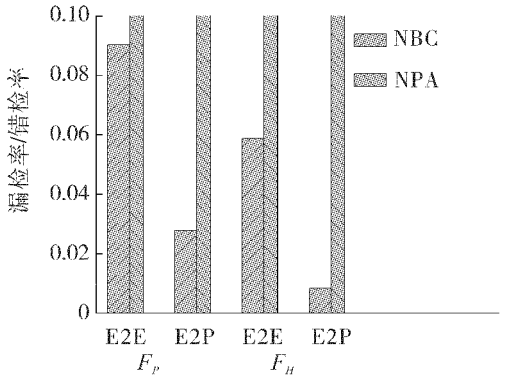


图1 NBC 与 NPA 识别比较

Fig. 1 Results of NBC and NPA

测试数据集通过有效载荷特征识别方法获得,作为 Skype 数据流基准测试数据,包括63 157 条 Skype 语音流,其中54 411条 E2E 流和8 746条 E2P 流. NBC 分类结果为 E2E 错检率为 9.0%,漏检率为 5.9%;E2P 的错检率为 2.8%,漏检率为 0.8%. NPA 分类结果为 E2E 错检率为 10.7%,漏检率为 24.2%;E2P 的错检率为 16.1%,漏检率为 19.3%. NBC 的模型分类性能较 NPA 分类模型效果要好.

表2 NBC 与 NPA 分类算法的分类性能

Table 2 Comparison between NBC and NPA

分类算法	承载类型	N_{SS}	N_{SN}	F_P	N_{NS}	F_N
测试	E2E	54 411	–	–	–	–
数据库	E2P	8 746	–	–	–	–
NBC	E2E	56 291	5 082	9.0	3 202	5.9
	E2P	8 921	248	2.8	73	0.8
NPA	E2E	75 712	8 120	10.7	13 181	24.2
	E2P	12 541	2 101	16.1	1 694	19.3

对于 NBC 错检样本,通过手工识别发现大部分情况为 RTP 承载的 VoIP 流,这些数据与 Skype 承载的 VoIP 实时流特征存在相似性. 对于漏检样本,分析发现基准测试库中包含 Skype 的视频/数据/聊天消息,这部分消息不符合实时流特征,NBC 无法从 E2E 流中分离聊天/数据流. 视频/数据/聊天数据不经过 PSTN 网关,E2P 测试库没有该类数据,因此 NBC 模型的 E2P 漏检索率较 E2E 漏检率要小.

3 结束语

Skype 流特征和其他网络应用数据流的不同特征,传统 C/S 应用流通常使用众所周知的端口作为服务端口,传输中数据包比较大. 其他 P2P 应用多使用明文传输,可采用载荷检测技术进行识别. 互联网流中实时音频与 Skype 流在实时流特征存在一定程度上类似,都是同时使用 TCP 和 UDP 传输数据,但其不具有 P2P 的连接特征. 其他 VoIP 流,如 MSN 或 QQ 也存在实时特征,大多建立在标准协议上,动态协商端口,存在可识别静态载荷特征. 网络游戏通常采用 UDP 协议传输数据,具有网络带宽消耗波动和突发性特征[9]. 本文结合了 Skype 的连接特征和实时流特征,采用朴素贝叶斯分类器识别 Skype 流量. 首先根据网络流的连接模式,识别出具有 P2P 连接模式网络数据流,进一步根据流的实时性特征识别 Skype 流,最后采用有效载荷交叉方法验证了算法的性能. 由于 Skype 属于目前私有协议,没有建立一个标准的测试集,对于 Skype 流的检测算法大多基于互联网流检测基础上,难以客观评价体系.

参考文献:

[1] BASET S A, SCHULZRINNE H. An analysis of the Skype peer-to-peer internet telephony protocol[C]// IEEE Infocom'06. Barcelona, Spain, 2006:1-11.

[2] BIONDI P, DESCLAUX F. Silver needle in the Skype [C]//Black Hat Europe'06. Amsterdam, The Netherlands, 2006, 1:25-47.

[3] YU Y F, LIU D D, LI J, et al. Traffic identification and overlay measurement of Skype[C]//Proc of IEEE International Conference on Computational Intelligence and Security. Guangzhou, China, 2006: 1043-1048.

[4] CHEN K T, HUANG C Y, HUANG P, et al. Quantifying fying Skype user sattsfaction[C]//ACM SIGCOMM'06. Pi-

sa, Italy, 2006:399-410.

[5]LU L, JEFFREY H, SAFAVI-NAINII R, et al. Transport layer identification of Skype traffic[C]//ICOIN 2007. Estoril, Portugal, 2007:465-481.

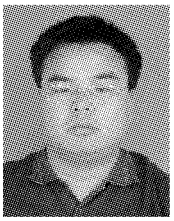
[6] DARIO B, MARCO M, MICHELA M. Revealing Skype traffic when randomness plays with you[C]//ACM Sigcomm'07. Kyoto, Japan, 2006:37-48.

[7] FALOUTSOS M, KARAGIANNIS C K, BROIDO A T. Transport layer identification of P2P traffic[C]// Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement. Taormina, Sicily, Italy, 2004: 121-134.

[8] MOORE A, ZUEV D. Internet raffic classification using Bayesian analysis [C]//ACM Sigmetrics BANFF, CA, 2005:50-60.

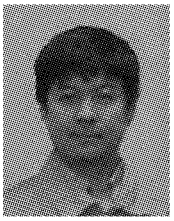
[9]FENG W, CHANG F, FENG W, et al. A traffic characterization of popular on-line games[J]. IEEE/ACM Transactions on Networking,2005,13(3): 488-500.

作者简介:



万月亮,男,1973 年生,讲师,博士,主要研究方向为网络信息安全、海量数据挖掘、网络行为分析、网络视频挖掘.发表学术论文近 10 篇,出版编著 1 部.

朱贺军,男,1973 年生,硕士,主要研究方向为网络信息安全、互联网挖掘与数据挖掘.



刘宏志,男,1964 年生,教授,博士,主要研究方向为信息工程监理与电子政务、软件工程、计算机网络,发表学术论文 60 余篇,主编、参编论著及教材 10 余部.

2010 年全国模式识别学术会议
Chinese Conference on Pattern Recognition 2010

随着人工智能、机器学习和计算机网络等相关技术的快速发展,模式识别研究在近几年来取得了令人瞩目的成就,一批研究成果得到了广泛应用和推广.继 20 世纪 80 年代以来中国自动化学会模式识别与机器智能专业委员会成功主办了多次国内学术会议后,全国模式识别学术会议 2007 年和 2008 年在北京、2009 年在南京举行,得到了国内同行的积极响应,会议取得圆满成功.

为了进一步促进模式识别研究的快速发展,加强国内外同行间的学术交流与合作,2010 年全国模式识别学术会议(Chinese Conference on Pattern Recognition 2010)将于 2010 年 10 月 21 - 23 日金秋时节在美丽的山城重庆召开.会议将邀请国内外著名学者做特邀学术报告,并向国内外同行征集有关模式识别理论方法研究和应用技术的学术论文.会议论文集将由 IEEE 出版,电子版将在 IEEE Xplore 发布并被 EI 数据库检索,并从会议论文中选出 20 篇左右与主题相关的优秀论文以英文形式在国际期刊《小波,多分辨分析与信息处理》(SCI 源刊)、《Frontiers of computer science in China》(EI 源刊)上公开发表.同前几届一样,会议将选择最佳论文和最佳学生论文予以奖励.

重要日期:

投稿截止日期: 2010 年 6 月 10 日;

论文录用通知: 2010 年 7 月 20 日;

最终论文提交: 2010 年 8 月 20 日;

会议时间: 2010 年 10 月 21 - 23 日.

联系方式:

联 系 人: 文　静,陈恒鑫;

联系电话:023-65106125 ;

联系传真:023-65102502;

电子邮件:CCPR2010@ cqu. edu. cn.