# 一种基于模糊评判规则的 P2P 流识别算法 *)

宫 婧[1,2] 孙知信[2] 顾 强[2]

(南京邮电大学数理学院信息与计算科学系 南京 210003)[1] (南京邮电大学计算机学院 南京 210003)[2]

**摘 要** 针对难以将那些协议未公开、特征不明显的 P2P 应用识别出来的问题,本文首次将模糊数学理论应用在 P2P 流的识别中,提出了一种基于模糊评判规则的 P2P 流识别算法 FJRRA。该算法首先对网络数据包进行统一描述,接着定义了长度、出现时机、特征描述以及位置这四个特征的隶属函数,随后建立相应的模糊类,最后用模糊评判规则评定它是归属于某类 P2P 应用。实验结果说明,运用 FJRRA 算法可以识别出 Skype 流,而且准确率较高。

**关键词** P2P 流,识别算法,模糊评判,Skype

## A Kind of P2P Flow Recognition Algorithm Based on Fuzzy Judgment Rules

GONG Jing[1,2] SUN Zhi-xin[2] GU Qiang[2]

(College of Applied Mathematics & Physics of Nanjing University of Posts and Telecommunications, Nanjing 210003,China)[1]

(College of Computer of Nanjing University of Posts and Telecommunications, Nanjing 210003,China)[2]

**Abstract** The paper aims at the question that some P2P applications are difficult to be identified, whose protocols haven't been opened and whose characters are unconspicuous. The paper applies fuzzy mathematics to the identification of P2P flow for the first time, and presents a kind of P2P flow recognition algorithm based on fuzzy judgment rules (FJRRA). The algorithm describes uniformly network data first, then four character membership functions are defined, viz. length, opportunity, character description and position. Subsequently, relevant fuzzy classes are set, and finally fuzzy judgment rules will judge the data is which kind of P2P applications. Our experiment results verify that the algorithm can identify Skype flow and accuracy is high.

**Keywords** P2P flow, Recognition algorithm, Fuzzy judgement, Skype

## 1 引言

P2P(Peer-to-Peer)技术自出现以来,便得到了快速的普及和发展,尤其是应用最为广泛的 P2P 文件共享技术[1,2]。由于 P2P 软件不断地进行更新,新的 P2P 软件也在不断涌现,并且 P2P 用户所共享的文件大多是最新或者最流行的,越来越多的人被吸引到 P2P 的阵营当中,P2P 流量在整个网络流量中的比重与日俱增。据统计,仅在短短的几年时间内,P2P 流量已经占据了固定网络中 60%以上的带宽,对 Web, Email 等其他网络服务构成了严重威胁[3]。于是,如何识别 P2P 流量以及对其进行控制,渐渐成为了人们比较关心的问题。

以往的 P2P 流量识别方法主要有利用端口和利用关键字两类。利用端口进行 P2P 流量识别即对各种 P2P 软件的相应流量进行研究,并归纳出常用的一个或多个固定端口(如 KuGoo 软件通用的商业端口是 7000)。然后在流量检测过程中,一旦发现有流量的端口与已归纳出的端口相同,就可以确定该流量属于 P2P 流量,并属于某一种 P2P 软件[4-6]。利用关键字进行 P2P 流量识别也是在研究各种 P2P 软件相应流量的基础上实现的[7-10]。这时对流量研究的目的不再是归纳特征端口,而是归纳出流量所有数据包中都含有的或者出现频率最高的特征字符串即关键字,一般关键字的出现位置也是有严格要求的。然后在流量检测过程中,对数据包进行深度检测。如果关键字匹配成功,就可以确定该流量属于 P2P 流量及其所属软件类别。

但是随着 P2P 技术的不断发展,上述两种方法已出现了明显的弊端[11]。对于利用端口识别 P2P 流量,现在大多数 P2P 软件都不再使用固定端口,或使用动态端口,或在软件中设有端口设置功能供用户自行设置端口,甚至有的 P2P 软件使用 80 等其它业务的固定端口号,以欺骗流量检测设备[12]。对于利用关键字识别 P2P 流量,关键字符串的部分或全部字节可能随软件的运行环境、版本等的改变而改变[13-15]。随之相应的流量检测软件就要始终监视各种 P2P 软件的流量,跟踪其版本、协议等最新信息,以保证精确地获取该 P2P 应用的最新特征字符串。这样,流量检测设备的开发队伍在寻找各种软件的关键字上所花费的力气远远大于对 P2P 流量识别方法的研究,导致本末倒置。

实际上,无论是利用端口还是利用关键字进行流量检测,都是在对某一 P2P 软件相关流量进行深入研究后,总结出该种 P2P 应用的固有特征,然后针对这些固有特征(净荷长度、关键字、出现时机等)成功识别流量。也就是说,如果能够总结出某种 P2P 应用的流特征,就能识别出该种 P2P 流量。现在 P2P 技术越来越成熟,软件的更新以及新软件的发布速度也越来越快,对 P2P 应用的流特性进行分析,之后对其流特

*)本文得到国家自然基金(NO. 60573141),教育部和南京市归国人员项目资助,华为资金资助。宫 婧 硕士,讲师,主要研究方向为计算机网络、P2P 应用、网络安全等;孙知信 博士,教授,主要研究方向为计算机网络、P2P 应用、网络安全等;顾 强 硕士,讲师,主要研究方向为数据挖掘、数据库、计算机网络等。

性进行统一描述，可以较好地检测出该种 P2P 应用。

因此，为了解决无法根据关键字、端口＋IP 的方式进行识别的 P2P 流，本文首次将模糊数学理论运用到 P2P 流识别中，提出了一种基于模糊评判规则的 P2P 流识别算法 FJRRA (Recognition Algorithm based on Fuzzy Judgment Rules)。本文的组织结构如下：第 2 节介绍模糊评判规则的理论依据和设计思想；第 3 节介绍算法的具体实现；第 4 节是算法在识别 Skype 中的应用，对实验数据进行分析和讨论；最后是总结。

## 2 模糊评判规则

模糊集合[16]用来描述一个模糊概念，它是内涵和外延都不明确的集合。本文利用模糊集合的概念，对于网络中的特征不明确、协议尚未公开的 P2P 流进行规范化描述，接着由四个特征隶属度函数建立相应的模糊类集合描述，随后通过综合评判，有效地识别出 P2P 流。

### 2.1 数据包集合的描述

**定义 1**(数据包的描述)

网络中捕获取的信令数据包，可以从以下几个方面加以描述 $D(L,T,O,C)$。

L(length)：指该信令数据包的包长，例如包长度为 11 的包，在 Skype 中出现，是一个非常有特色的特征信息；

T(Type)：指报文的类型，UDP 包或是 TCP 包；

O(Opportunity)：是指其出现的时机和频率等。例如，某些数据包会在用户端和服务器端交互的时候出现；某些包只在开始建立连接的时候出现；某些包在进行 TCP 三次握手的时候出现；

C(characteristic)：是指其出现时的特性，即某些字节代表的是何种含义，如表示是关键字模式串的。例如：在 QQ 应用中，有“0x02”，“0x0A 0x1D”这样的特征字符串，它们可作为识别时的关键字；

**推论** 根据定义 1，所捕获的数据包被描述为 $D(L,T,O,C)$，此时根据其不同的描述内容加以分组，就可以获取基于“$L$”一类的数据包的集合，称之为“$L$”类集合。同样，有“$T$”类集合、“$O$”类集合，以及“$C$”类集合。

为了对所截获的数据包进行规范化描述，可以进一步对其某些特征进行细致描述。本文以关键字模式串为主，进一步定义了规范化的关键字描述。

**定义 2**(关键字的描述)

对于某个关键字模式串，可以从以下方面加以描述 $D(L,N,P,C)$。

L(length)：指该关键字模式串所在的信令数据包的包长。包长度为 11 的包、长度为 14 的包等等，这都是一些会出现关键字的包长；

N(Number)：关键字模式串的字节数。例如，有的特征字符串是 2 个字节，有的则可能有 6～8 个字节等。

P(Position)：指关键字模式串在报文中的位置，即出现在哪个字节上。例如：在 QQ 应用中，“0x02” 是指包标识，“0x0A 0x1D”出现在第 1～2 个字节。

C(Content)：指该关键字模式串所代表的含义。例如，在 QQ 应用中，“0x02”出现在第 0 个字节，“0x0A 0x1D”是版本 QQ2003(0808)的标识。

根据定义 1 和定义 2 的描述，本文对网络中的数据包的描述主要分为：$L$ 类集合、$O$ 类集合、$C$ 类集合、$P$ 类集合等几类集合。为了实现 P2P 流的识别，根据模糊数学理论，就需要对各类集合建立相应的隶属函数，根据模糊集合间的关系以及所占的权重，才能进行计算，得到评判结论。由于网络中的数据其实是一些随机现象，因此选取隶属函数时，本文是根据数据某项特征的密度分布函数来确定其隶属度的函数的[20]。下面就对各类集合定义其相应的隶属函数。

### 2.2 隶属度函数的定义

对于一个集合 $A$ 可用隶属度函数 $\mu$ 来表示，$\mu$ 的定义域是与该模糊概念相关的对象的集合，称为 $A$ 的论域 $U$，而 $\mu$ 表示了 $U$ 中每个对象隶属于 $A$ 的程度，称为隶属度，$\mu\in[0,1]$。模糊集合是模糊数学的基本概念，通过它可以建立与普通集合的关系，继而用经典数学的方法来研究模糊现象。

**定义 3**(长度特征的隶属度函数)

$$\mu_l(t)=\begin{cases}0, & t<l_1 \text{ 或 } t>l_3\\ \dfrac{t-l_1}{l_2-l_1}, & l_1\leqslant t\leqslant l_2\\ \dfrac{l_3-t}{l_3-l_2}, & l_2\leqslant t\leqslant l_3\end{cases}\tag{1}$$

(1)式中的 $l_1,l_2,l_3$ 表示长度特征在观测过程中发生变化的几个临界点，根据所捕获的数据包的情况来确定。在描述数据包时，$\mu_l(t)$ 表示的是数据包净荷的长度特征的隶属度；在描述关键字串时 $\mu_l(t)$ 表示的是关键字符串的长度特征隶属度。于是，$\mu_l$ 表示将 $L$ 类集合归为 $L$ 类模糊集。

**定义 4**(出现时机的隶属度)

$$\mu_s(t)=\begin{cases}0 & t\leqslant \bar{f}-s\\ 1 & \bar{f}-s<t\leqslant \bar{f}+s\\ 0 & t>\bar{f}+s\end{cases}\tag{2}$$

(2)式中，$f_i$ 表示被研究的数据包 $x_i$ 出现的频率，而

$$\bar{f}=\frac{\sum_{i=1}^{n}f_i}{n},\ s=\sqrt{\frac{\sum_{i=1}^{n}(f_i-\bar{f})^2}{n-1}}$$

定义 1 和定义 2 中对数据包和其关键字串的特征都进行了描述，它们是进行评判时候的重要依据。为了将这种描述性的语义由模糊集合来表达，本文先将其数值化，之后再定义相应的隶属度函数。

**定义 5**(特征描述函数)

$$CX=P\circ K\circ X=(P(x_1,c_1)\quad P(x_2,c_2)\quad\cdots\quad P(x_n,c_n))\circ\begin{pmatrix}K(x_1,c_1)\\K(x_2,c_2)\\\vdots\\K(x_n,c_n)\end{pmatrix}\circ(x_1\quad x_2\quad\cdots\quad x_n)\tag{3}$$

(3)式表示对数据包的特性进行量化规范，其中 $c_i$ 表示数据包 $x_i$ 中的特征字段，$P(x_i,c_i)\in(0,1)$ 表示某个特征字段出现的概率；$K(x_i,c_i)\in(-1,1)$ 是指该特征字段的权值，即该特性对确定 P2P 流类型的可信程度。

例如，Skype 中的“02”字段，$x_1,x_2,x_3$ 分别表示在 11，18 和 23 字节包中出现，

$$CX=P\circ K\circ X=0.35\quad 0.78\quad 0.22\circ\begin{pmatrix}0.1\\0.8\\0.3\end{pmatrix}\circ(x_1\quad x_2\quad x_3)=0.725(x_1\quad x_2\quad x_3)$$

是对“02”字段在 11 字节，18 字节和 23 字节的包中出现的量化描述。

**定义 6**(特性描述隶属度)

$$\mu_c(t)=\begin{cases}1-\dfrac{C(x_i)}{2}t & 0\leqslant C(x_i)\leqslant 1\\ 0 & C(x_i)>1\end{cases}\tag{4}$$

(4)式中 $C(x_i)$ 是由公式(3)计算出来的。$\mu_c$ 表示将 C 类集合归为 C 类模糊集。

**定义 7**(位置隶属度)

$$\mu_p(t)=\begin{cases}0 & t<p_1 \text{ 或 } t>p_3\\ \dfrac{t-p_1}{p_2-p_1} & p_1\leqslant t\leqslant p_2\\ \dfrac{p_3-t}{p_3-p_2} & p_2<t\leqslant p_3\end{cases} \tag{5}$$

(5)式中,在描述数据包时,$\mu_p$ 表示数据包 $x_i$ 的位置权值 $p(x_i)=\dfrac{time(x_i)}{total\text{-}time}$ 的隶属度,$time(x_i)$ 是指获取数据包 $x_i$ 的时刻,total-time 是一次实验中的总的抓包时长。在描述数据包的特征字($key;x_i$)时,$\mu_p$ 表示该特征字的位置权值 $p(key;x_i)=\dfrac{position(key)}{length(x_i)}$ 的隶属度,$position(key)$ 是指 key 在数据包 $x_i$ 中的位置,$length(x_i)$ 表示对应数据包 $x_i$ 的总长度。

### 2.3 FJRRA 算法的评判规则

在本文中,FJRRA 算法借鉴了模糊数学中的概念。为了识别网络中的 P2P 流量,本节制定了评价规则。

为了断定某个数据流是否属于某种 P2P 应用,根据模糊关系的理论,由模糊关系系数定义评价向量。根据该评价向量,进行综合评判。所谓模糊关系系数[18]是关系的推广。在模糊关系中,事物间的关系不是仅用"有"或"无"来描述,而是用隶属度来表述模糊类之间的关联程度。在本文中,用模糊评判规则将比较复杂的事件或对事物的整体评估分成许多比较简单的小部分来分别评估,最终得出对整个事件或事物的结论。下面给出进行模糊评判时评判值的定义。

**定义 8**(评价向量) 是二元组 $E=\langle U,\lambda\rangle$,其中 $U=\{\mu_l,\mu_c,\mu_s,\mu_p\}$ 是隶属度集;$\lambda=\{\lambda_l,\lambda_c,\lambda_s,\lambda_p\}$ 是权重向量,表示所对应的隶属度在评判中所占的比重。则评判值

$$E=U\cdot\lambda^T=\mu_l\cdot\lambda_l+\mu_c\cdot\lambda_c+\mu_s\cdot\lambda_s+\mu_p\cdot\lambda_p \tag{6}$$

且 $\lambda_l+\lambda_c+\lambda_s+\lambda_p=1$。

对于特征不明确、协议尚未公开的 P2P 流,无法通过关键字检测或端口检测的方式判断出来。FJRRA 算法首先对网络中的数据流进行规范化的统一描述,并据此建立各类模糊类和隶属度函数,之后计算评判值。关于如何进行判断识别,将在第 3 节中进行具体描述。

## 3 基于模糊评判规则的 P2P 流识别算法 FJRRA

### 3.1 FJRRA 识别方法的评判结论的确立

在识别 P2P 流的方法中,对于有显著特征关键字的 P2P 流,采用模式匹配的方法识别比较有效;对于有特征源目 IP 或端口的 P2P 流,可采用观察端口的方式识别。而对于特征不明确、协议尚未公开的 P2P 流,上述识别方法都不适用,所以本文提出了 FJRRA 识别方法。

对网络数据流量的分析可看成是做若干次随机试验。皮尔逊[18]曾证明:当试验次数 $n$ 充分大时,虽然某个事件出现的频率 $f_i/n$ 与它的概率 $p_i$ 有差异,但仍然可以用 $\chi^2$ 检验法来检验。同时,经过多次的观测和研究分析,发现网络流量总体的统计属性分布具有自相似性[19]和稳定性的特性[17]。例如,在 Skype 仿真试验中,我们每隔一定时间 $t$ 对网络中数据流量连续采集 100 个数据包,就数据包长度来说,每次观测数据是两两不相容的事件 $A_0,A_1,\cdots,A_n$,可以把一些常见的长度如 11 字节、14 字节、18 字节、23 字节等常出现的字节长度作为各个分组类,并把那些极少出现的长度合并成为一个分组。这样的 $k$ 个分组可以保证在假设检验过程中每个分组都不会出现 $np_i<5$ 的情况。对于网络流量分布的统计量 $\chi^2=\sum_{i=1}^{k}\dfrac{(f_i-np_i)^2}{np_i}$,就满足在置信度 $\alpha$ 下,$\chi^2<\chi_\alpha^2(k-1)$,因此可以使用 $\chi2$ 检验法来进行检验。经过多次的观测和分析,P2P 流的密度函数近似于 $\chi^2$ 分布,当隶属度为 $U=\{\mu_l,\mu_c,\mu_s,\mu_p\}$ 的情况下,令

$$\bar{\mu}=\frac{\mu_l+\mu_c+\mu_s+\mu_p}{k} \tag{7}$$

$$\bar{E}=\sqrt{\frac{(\mu_l-\bar{\mu})^2+(\mu_c-\bar{\mu})^2+(\mu_s-\bar{\mu})^2+(\mu_p-\bar{\mu})^2}{k-1}} \tag{8}$$

$$\because\frac{(k-1)\bar{E}^2}{\bar{\mu}^2}\sim\chi^2(k-1)$$

$$\therefore P\{\chi^2_{1-\alpha/2}(k-1)<\frac{(k-1)\bar{E}^2}{\delta^2}<\chi^2_{\alpha/2}(k-1)\}=1-\alpha$$

成立,

即得到隶属度 $\bar{\mu}^2$ 的一个置信度为 $\alpha$ 的置信区间

$$\left(\frac{(k-1)\bar{E}^2}{\chi^2_{\alpha/2}(k-1)},\frac{(k-1)\bar{E}^2}{\chi^2_{1-\alpha/2}(k-1)}\right) \tag{9}$$

所以,FJRRA 算法中,评判值 $E\in\left(\dfrac{\sqrt{k-1}\bar{E}}{\sqrt{\chi^2_{\alpha/2}(k-1)}},\dfrac{\sqrt{k-1}\bar{E}}{\sqrt{\chi^2_{1-\alpha/2}(k-1)}}\right)$ 则为可识别流。

### 3.2 FJRRA 算法的实现

基于 FJRRA 识别 P2P 流的算法,可以识别网络中"特征不明确、协议尚未公开的"这一类 P2P 流。例如试验中的 Skype 就是该类 P2P 流的典型代表。算法实现主要分信息提取、规范化描述、信息交互和决策判断等步骤来完成。FJRRA 算法实现的流程总结如下:

Step1. 网络中的数据流量经过检测设备,首先经过关键字、端口、IP 等方式识别,对于剩下的无法识别的可能为 P2P 流的数据进行捕获。

Step2. 对数据包进行捕获分析,提取出其中长度、出现的时机、特征描述以及位置特征,按照前述定义和规范化描述,描述成 $D1(L,T,O,C)$ 和 $D2(L,N,P,C)$ 的格式。

Step3. 根据公式(1)、(2)、(4)、(5),计算隶属度函数值 $\mu_l,\mu_c,\mu_s,\mu_p$,得到 $L$ 类、$C$ 类、$S$ 类和 $P$ 类模糊集合类。

Step4. 在置信度 $\alpha$ 情况下,权重向量 $\lambda=\{\lambda_l,\lambda_c,\lambda_s,\lambda_p\}$,计算出评判值的置信区间。

Step5. 根据公式(6),计算出评判值 $E$。

Step6. 由决策器进行评判。若 $E$ 落在置信区间内,则标记为 P2P 流。

决策器是根据系统中的已有数据,可以识别出网络流量是属于何种 P2P 流。它根据捕获到数据的实际情况,对定义 8 中的权重向量 $\lambda=\{\lambda_l,\lambda_c,\lambda_s,\lambda_p\}$ 和置信度 $\alpha$,结合网络流量状态进行动态调整,同时可以将判断的结果传送回系统的数据库中,指导后续参数分析。

## 4 FJRRA 算法在识别 Skepe 流中的应用和分析

Skype 的协议是经过加密的一种 P2P 应用协议,属于私有不公开协议,因此对该类应用的识别与分析比较困难。在实验中,数据由 COMMVIEW V5.0 抓取,保存成 txt 格式,然后通过提取数据程序提取出来。主要信息有序号、在原来抓包过程中的序号、源 IP、端口号、目的 IP、端口号、协议及数据部分,通过对流经网络的数据进行采集和规范化描述,利用 FJRRA 算法,实现了识别 Skype 流的目的。

### 4.1 实验数据

实验数据 1：超级节点(super node，SN)回复给客户端(Skype client，SC)的数据包。

抓取长度为 11 字节(本文中所有提到的包的长度都是指数据部分)的数据包，这是一种非常有代表性的数据包，它在 SN 回复 SC 的时候出现。其结构如图 1 所示。

| No. | Protocol | MAC Addresses |
|---|---|---|
| 8 | IP/UDP | RealtekSem:B2:35:A5 => Cisco:84:F1:48 |

```
0x0000   00 E0 4C B2 35 A5 00 03-E3 84 F1 48 08 00 45 00
0x0010   00 27 AE FD 00 00 70 11-36 36 DB 4D 30 31 0A 0A
0x0020   50 0A FC 57 1A 0A 00 13-F8 76 5C E9 47 DA 02 D8
0x0030   13 5F D0 61 00 00 00 00-00 00 00 00
```

图 1 SN 回复给 SC 的数据包(11 字节)

分析：在图 1 中，(1)UDP 头部分的 0013(十进制是 19)说明所捕获的数据包总长度，因为本文中所有提到的包的长度都是指数据部分，所以减去 UDP 的包头 8 个字节，可以得知数据包的数据部分是 11 字节。(2)从登录过程的全部数据包来看，将长度为 11 的包提取出来，与数据包总体进行比较，长度为 11 的包所占的比例基本上都在 90%。(3)由于这是登录时的数据包，因此在 IP 头部分，可以得到 SN 的 IP 地址是：DB 4D 30 31(即 219.77.48.49)。

这样根据定义 1，可以从四个方面描述该数据包，即 $D1(L,T,O,C)=D1(11,\text{UDP},\text{登录时出现频率 }90\%,\text{登录中参与交互的 SN 的 IP 地址是 }219.77.48.49)$。

实验数据 2：包含“02”特征字的包

从多次试验数据中，发现“02”是一个典型的模式串。典型的数据包有：在登录中第一个包就会发的、之后也不断发的、长度为 18 字节的包，如图 2 所示。

| No. | Protocol | MAC Addresses |
|---|---|---|
| 2 | IP/UDP | RealtekSem:B2:35:A5 => Cisco:84:F1:48 |

```
0x0000   00 03 E3 84 F1 48 00 E0-4C B2 35 A5 08 00 45 00
0x0010   00 2E 6E 5E 00 00 80 11-67 33 0A 0A 50 0A D5 D6
0x0020   35 43 1A 0A 05 2E 00 1A-C1 69 5C E7 02 E3 DC 18
0x0030   C4 DD CF 70 27 1D F6 77-13 2B B8 F8
```

图 2 包含“02”特征字的数据包(18 字节)

分析：在图 2 中，(1)UDP 头部分的 001A(十进制是 26)说明所捕获的数据包总长度，减去 UDP 的包头 8 个字节，得知数据包的数据部分是 18 字节。(2)含有特征字“02”位于数据部分的第 3 个字节上。(3)该特征字的长度是 1 个字节。(4)所捕获的数据包通常是在连接时的一次呼叫中得到的。

因此，根据定义 2，对含“02”特征字的包描述为 $D2(L,N,P,C)=D2(18,1,3,\text{连接时的一次呼叫})$。

实验数据 3：TCP 流量的特征包(图 3)

```
0x0000   00 03 E3 84 F1 48 00 E0-4C B2 35 A5 08 00 45 00
0x0010   00 2D 6E 94 40 00 80 06-2C 4D 0A 0A 50 0A D4 48
0x0020   31 8D 05 21 81 09 70 51-E6 B7 85 CB 9E 61 50 18
0x0030   44 70 F3 89 00 00 16 03-01 00 00
```

```
0x0000   00 E0 4C B2 35 A5 00 03-E3 84 F1 48 08 00 45 00
0x0010   00 2D F5 67 40 00 2A 06-FB 79 D4 48 31 8D 0A 0A
0x0020   50 0A 81 09 05 21 85 CB-9E 61 70 51 E6 BC 50 18
0x0030   16 D0 1F A5 00 00 17 03-01 00 00 00
```

图 3 TCP 流量的特征包

分析：建立连接后，前面两个交互包固定，客户端向对方发送一个 5 个字节的包，内容为 16 03 01 00 00。对方回复一个 7 个字节的包，内容为 17 03 01 00 00 00。这种情况在很多登录情况下都会出现，对方的 IP 也比较固定，212.72.49.141 是较常出现的一个。

因此根据定义，得到 $D3(L,N,P,C)=D3(72,5,1,\text{请求建立 TCP 连接})$，$D4(L,N,P,C)=D4(93,16,1,\text{对方回复建立 TCP 连接})$。

### 4.2 结果分析

经过长时间的多次捕获，对于所捕获的数据包按照长度、出现的时机、特征描述以及位置进行分析，并根据公式(1)、(2)、(4)、(5)计算出相应的隶属函数值，并做出图形，如图 4。

在 $\alpha=0.1,\lambda=\{\lambda_l,\lambda_c,\lambda_t,\lambda_p\}=0.5,0.1,0.3,0.1$ 的情况下，表 1 是根据对图 5 中的不同点进行取值，并根据公式(6)-(8)对其进行了评判值计算。

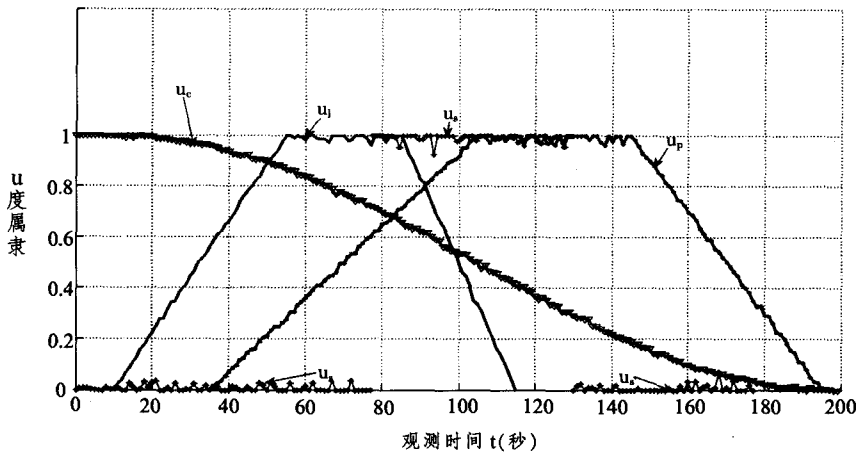


图 4 隶属度函数图

表 1 Skype 试验数据表

| 序号 | $\mu_l$ | $\mu_s$ | $\mu_c$ | $\mu_p$ | E | 置信区间 | 结果 |
|---|---|---|---|---|---|---|---|
| 1 | 0.1 | 0 | 0.99 | 0 | 0.337 | (0.088685,1.968963) | 可识别 |
| 2 | 0.23 | 0.05 | 0.97 | 0 | 0.388 | (0.077502,1.720668) | 可识别 |
| 3 | 0.5 | 0.05 | 0.93 | 0.05 | 0.494 | (0.068417,1.518963) | 可识别 |
| 4 | 0.75 | 0.1 | 0.9 | 0.13 | 0.606 | (0.066001,1.465341) | 可识别 |
| 5 | 1 | 0.1 | 0.85 | 0.25 | 0.715 | (0.074856,1.661932) | 可识别 |
| 6 | 1 | 0.98 | 0.6 | 0.5 | 0.778 | (0.025438,0.564773) | 不可识别 |
| 7 | 0.5 | 0.99 | 0.5 | 0.8 | 0.609 | (0.022274,0.494531) | 不可识别 |
| 8 | 0.1 | 1 | 0.3 | 1 | 0.43 | (0.084453,1.875) | 可识别 |
| 9 | 0 | 1 | 0.2 | 0.98 | 0.356 | (0.103941,2.30767) | 可识别 |
| 10 | 0 | 0.1 | 0.1 | 0.5 | 0.14 | (0.018874,0.419034) | 可识别 |

从实验数据分析中可以看出，采用 FJRRA 算法，有 80% 的数据是可以识别出来的。该算法对 Skype 流量的识别具有较好的性能。

### 4.3 FJRRA 在识别其他 P2P 中的应用

网络游戏也是一种应用广泛但不容易识别出的网络流量。对于某些大型的网络游戏，用 FJRRA 的方法同样可以较好地识别。以网络游戏魔兽世界为例，对其进行一段时间的观测后，得出魔兽世界的特点：

(1)在登录过程中，得到登录服务器的 IP、使用的端口以及一些固定的字符串信息，可以对其进行描述。Server IP：219.133.56.109；Server Port：魔兽的服务器端口是 3724；Login Keyword：是指游戏的登录过程中所出现的关键字模式串及其特征，对所有客户端来说都是固定的。对登录过程进行描述，得到登录类集合。

(2)在游戏进行过程中，角色在走动时，有一些数据部分长度固定为 34bytes，并且有固定字段。另外，局部固定字段表示游戏在某个区域内(例如提瑞斯法林地：0x73，0x00，丧钟墓：0x12，0x01)。

(3)在游戏进行过程中，角色在对抗时，数据包数据部分长度固定为 20bytes，并有固定字段。对游戏交互过程进行描述，得到游戏交互类集合。

(4)魔兽世界的更新是强制性的，玩家每次玩时都要保证所用的版本是最新的，否则玩不了，所以更新过程也非常具有特点，可以作为识别时的因素之一。对更新过程进行描述，得到更新类集合。

随后，分析所捕获的数据，得到对各个特征的描述以及它们的隶属度函数，可以对网络中的流量进行模糊识别。

**结束语** 迄今为止，对于如何识别某些协议未公开、特征不明显的 P2P 流，仍然处在研究探讨阶段。本文提出的 FJRRA 算法，能够较好地识别网络流量中的 Skype 流，对于其他的 P2P 流量识别同样适用，具有较好的准确性和可扩展性。

## 参 考 文 献


[1] Karagiannis T, et al. Is P2P dying or just hiding? // Proc. 2004 IEEE Global Telecommunications Conf. (GLOBECOM 04). IEEE Press, 2004: 1532-1538

[2] Gerber A, Houle J, Nguyen H, et al. P2P the gorilla in the cable//Proc. National Cable & Telecommunications Association (NCTA) 2003 National Show. 2003

[3] Azzouna N B, Guillemin F. Impact of peer-to-peer applications on wide area network traffic: An experimental approach. In Globecom. IEEE, Dallas, 2004

[4] Gerber A, Houle J, Nguyen H, et al. P2P The Gorilla in the Cable//National Cable & Telecommunications Association (NCTA) 2003 National Show. Chicago, IL, June 2003

[5] Sen S, Wang J. Analyzing peer-to-peer traffic across large networks//Proceedings of ACM SIGCOMM Internet Measurement Workshop. Marseilles, France, November 2002

[6] Saroiu S, Gummadi K P, Dunn R J, et al. An Analysis of Internet Content Delivery Systems// Proceedings of the 5th Symposium on Operating Systems Design and Implementation. 2002

[7] Karagiannis T, Broido A, Faloutsos M, et al. Transport layer identification of P2P traffic//Proc. 4th ACM SIGCOMM Conference on Internet Measurement. 2004:121-134

[8] Sen S, Spatscheck O, Wang D. Accurate, scalable in network identification of P2P traffic using application signatures//Proc. 13th International Conference on World Wide Web. 2004:512-521

[9] Karagiannis T, Broido A, Brownlee N, et al. File-sharing in the Internet: A characterization of P2P traffic in the backbone

[10] Spognardi A, Lucarelli A, Pietro R D. A methodology for P2P file-sharing traffic detection//Proc. International Workshop on Hot Topics in Peer-to-Peer Systems. 2005:52-61

[11] Schmidt S E(Guran), Soysal M. An Intrusion Detection Based Approach for the Scalable Detection of P2P Traffic in the National Academic Network Backbone //Proceedings of the Seventh IEEE International Symposium on Computer Networks(ISCN' 06). 2006

[12] Leibowitz N, Ripeanu M, Wierzbicki A. Deconstructing the Kazaa Network//3rd IEEE Workshop on Internet Applications (WIAPP'03). 2003:112-120

[13] Saroiu S, Gummadi P K, Gribble S D. A Measurement Study of Peer-to-Peer File Sharing Systems. In MMCN, 2002

[14] Tutschku K. A Measurement-based Traffic Profile of the eDonkey Filesharing Service// PAM. 2004

[15] Leibowitz N, Bergman A, Ben-Shaul R, et al. Are File Swapping Networks Cacheable? Characterizing P2P Traffic// 7th IWCW. 2002

[16] 何新贵. 模糊知识处理的理论和技术. 第二版. 北京：国防工业出版社，1998

[17] 张剑，龚俭. 一种基于模糊综合评判的入侵异常检测方法. 计算机研究与发展，2003,40 (6):776-783

[18] 盛骤，谢世千，潘承毅. 概率论与数理统计. 第三版. 北京：高等教育出版社，2001

[19] Rincon D, Martinez S, Cano C, et al. Synthesis and analysis of fractal LAN traffic at high speeds//Local and Metropolitan Area Networks, LANMAN 2004//The 13th IEEE Workshop. April 2004:259-264